FARKLI DİLLERDEKİ BELGELERİN BENZERLİĞİNİN TESPİTİ

HAKAN YILMAZER

MERSİN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

MERSİN
TEMMUZ- 2013

FARKLI DİLLERDEKİ BELGELERİN BENZERLİĞİNİN TESPİTİ

HAKAN YILMAZER

MERSİN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

Tez Danışmanı
Yrd. Doç. Dr. Zeki Yetgin

MERSİN
TEMMUZ - 2013

Hakan YILMAZER tarafından Yrd. Doç. Dr. Zeki YETGİN danışmanlığında hazırlanan "Farklı Dillerdeki Belgelerin Benzerliğinin Tespiti" başlıklı bu çalışma aşağıda imzaları bulunan jüri üyeleri tarafından oy birliği ile Yüksek Lisans Tezi olarak kabul edilmiştir.

İmza

Doç. Dr. Ali AKDAĞLI .....

Yrd. Doç. Dr. Zeki YETGİN .....

Yrd. Doç. Dr. Ali YILDIZ .....

Yukarıdaki Jüri kararı Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 20../08../2013..tarih ve 2013.15...../..439..... sayılı kararıyla onaylanmıştır.

Doç.Dr.Mehmet KÜÇÜKASLAN
Enstitü Müdürü

ÖZ

Teknolojinin gelişmesi ile birlikte internet kullanımı ve web dokümanlarının boyutu paralel olarak artmaktadır. Dünyanın farklı coğrafyalarında, farklı dillerde internet ortamında dijital olarak paylaşılan dokümanlar çoğalmaktadır. İnternetin bu devasa bilgi kapasitesinde yer alan dokümanlar elbette tekil değildir.

Farklı dillerdeki bu dijital dokümanların bazıları benzer içeriğe sahip olabilirken, bir kısmı diğerinin alıntısı veya özeti, bir kısmı ise orijinal dökümanın birebir çevirisi olarak yer alabilmektedir. Birçok dokümanın orijinal dilinde kopyaları, alıntıları ve benzerleri, gerek başka dillerde tercümeleri mevcuttur. Bilginin bu kadar önemli olduğu çağda, aranılan metin veya belgelerin farklı dillerdeki mevcudiyetleri bilgiye erişimi kolaylaştıracaktır. Bir dilde yazılmış belgenin başka dillerde doğru çevrimlerini ve alıntılarını hızlı bir şekilde bulmak, araştırmacılar açısından da faydalı olacaktır. Bunun yanında, aynı metnin farklı dillerde bulunabiliyor olması bir akademik çalışmanın orijinal dilinin dışındaki başka dillerdeki intihallerinin bulunmasında da yardımcı olacaktır.

Bu tez çalışmasında bir belgenin farklı dillerdeki anlamsal ve içerik olarak benzerlerinin bulunması için yeni algoritmalar geliştirilmesi amaçlanmıştır. Bu algoritmalarda, dökümanlar metin, kelime, harf benzerliğinin yerine sayısal vektörler olarak düzenlenmiştir. Ayrıca uzaklık ve benzerlik ölçümleri literatürde kullanılan farklı yöntemler ile test edilmiştir.

Anahtar Kelimeler: Farklı dillerde doküman benzerliği, kümeleme, intihal, benzerlik ölçütleri, öznitelik vektörü çıkarım metotları, eşdizimlilik, doküman hizalama.

Danışman: Yrd. Doç. Dr. Zeki YETGİN, Mersin Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalı

ABSTRACT

The use of internet and size of the documents have been increasing in parallel with the development of technology. The documents in different languages, which are digitally shared on Internet, have risen on different geographies of world. The documents placed in the paramount information capacity of Internet are not single for certain.

While some of these digital documents in different languages have similar contents, some of them may be citation or summary of the original or some parts of the document may be literal translation of the original. The copies, citations, and duplications of the documents in original languages as well as their translations in different languages are available. The existence of the searched text or documents in various languages make easy to access the information in this era when the information is very important. It will useful for the monolingual individuals to find the translations and citations of the document which is written in mono language. Further, it will be beneficial for researchers to find a mono-lingual document's right citiations and translations in different languages easily. Beside, the presence of a text in difference languages will be helpful for the detection of plagiarism of an academic study in different languages other from the original language of the study.

In this thesis study, novel algorithms are aimed to be developed in order to find out a documents's similarities in different languages in terms of their semantics and contents. In these algorithms, documents are organized in feature vectors rather than text, word, and letter similarities. Moreover, distance metrics and similarity measures are tested with different state of the art methods that have been used in literature.

Keywords: cross language information retrieval, clustering, plagiarism, similarity metrics, feature vector extraction methods, co-occurence, document alignment.

Advisor: Assist. Prof. Dr. Zeki YETGİN, Department of Computer Engineering, Mersin University

TEŞEKKÜR

Çalışmamın her aşamasında bilgi ve tecrübesini benimle paylaşan, her türlü desteğini benden esirgemeyen danışmanım Sayın Yrd. Doç. Dr. Zeki YETGİN'e, Sayın Doç. Dr. Ahmet UYAR'a, Sayın Doç. Dr. Ali AKDAĞLI'ya, Sayın Yrd. Doç.Dr. Ali YILDIZ'a, arkadaşlarıma, çok sevdiğim aileme ve son olarak da varlığı ile bana güç veren sevgili eşime, teşekkür ederim.

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

Sayfa

SİMGE VE KISALTMALAR DİZİNİ

ÖVÇY-1 : Öznitelik Vektörleri Çıkarım Yöntemi - 1
ÖVÇY-2 : Öznitelik Vektörleri Çıkarım Yöntemi - 2
ÖVÇY-3 : Öznitelik Vektörleri Çıkarım Yöntemi - 3
ÖVÇY-4 : Öznitelik Vektörleri Çıkarım Yöntemi - 4
ÖVÇY-5 : Öznitelik Vektörleri Çıkarım Yöntemi - 5
ÖVÇY-6 : Öznitelik Vektörleri Çıkarım Yöntemi - 6
PC : Personal Computer
ÖV : Öznitelik Vektörü
VSM : Vector Space Model
IDF : Inverse Document Freqeuncy
EUROVOC: European Vocabulary Dictionary
UTF : Unicode Transformation Format
ASCI : American standard Code for Information Interchange
en : İngilizce
de : Almanca
fr : Fransızca
it : İtalyanca
es : İspanyolca
pt : Portekizce
tr : Türkçe
nl : Flemenkçe
t : terim
tf : terim frekansı
w : terim ağırlığı
idf : inverse document frequncy
MI : Karşılıklı bilgi miktarı skoru (Mutual Information)
LLR : Log Likelihood Oranı
H : Entropi değeri
MSE : Ortalama Hata
BY : Benzerlik Yöntemi

1. GİRİŞ

Son yıllarda teknolojinin hızla gelişmesi ile birlikte internet artık hayatımızın değişmez bir parçası haline gelmiştir. Özellikle sosyal medya siteleri, gücü yadsınamaz bir iletişim aracı haline dönüşmüştür. Birçok platformdan insanlar internet hizmeti alabilmektedir. Bilgi çağında insanların gereksinimi gerek yazılı, gerek görsel bilgilerle karşılanmaktadır. Yazılı metinler; sosyal platformlarda, haber sitelerinde, ansiklopedilerde, bloglarda, araştırma sitelerinde paylaşılmaktadırlar. PC veya Mac ortamlarının yanı sıra artık dizüstü, avuç içi veya akıllı telefonlar gibi ortamlar, sunulan bu dijital doküman denizini daha da ulaşılabilir ve zengin kılmaktadır. Fakat dünyanın farklı coğrafyalarından, farklı birçok platformdan, ulusal veya yerel dillerde paylaşıma sunulan bu bilgiler, insanların gereksinimlerine etkili şekilde ancak yazıldıkları diller ölçüsünde cevap vermeye çalışabilmektedirler.

Dünya nüfusu ve internet kullanımı üzerine yapılan araştırmalara bakıldığında, dünya nüfusu yaklaşık 7.1 milyar olarak hesaplanmaktadır [1] ve bu nüfusun %39'u internet kullandığı belirlenmiştir. Buradan 2.7 milyar insanın internet kullandığı anlaşılmaktadır ve bu kullanıcıların %73'den fazlası İngilizce dışında bir dil kullanmaktadır. Yine bu araştırmalar gösteriyor ki İngilizce bilmeyen internet kullanıcılarının nüfusu, İngilizce bilenlere göre yıllara göre daha hızlı artmaktadır.

Çizelge 1.1. 2005–2013 yılları arası dünya nüfusu ve internet kullanım oranları [1].

	2005	2010	2013 (ilk 6 ay)
Dünya nüfusu (milyar)	6.5	6.9	7.1
İnternet kullanıcısı (%)	16	30	39

2000-2011 yılları arasında, arama motorlarında yapılan İngilizce olmayan sorguların oranı, İngilizce sorgulara göre daha fazla artmaktadır. Örneğin Çin dilini konuşan internet kullanıcıları 1478 kat artarken, İspanyolca konuşan internet kullanıcıları ise 800 kat artmıştır [2]. Bu durum da internetin gün geçtikçe dünyanın

her coğrafyasına yayıldığının bir işaretidir. Bir istatistiğe göre internette yer alan doküman sayısının 3.56 milyar sayfa olduğu ifade edilmektedir [3].

Çizelge 1.2. 2000–2011 yılları arası anadillerine göre ayrıştırılmış internet kullanıcıları sayısı ve nufüs artış miktarı [2]

No	Dil	İnternet Kullanıcısı	Yüzde	2000–2011 yılları arası internet kullanım oranındaki yüzdelik artış
1	İngilizce	565,004,126	26.8	301.4
2	Çince	509,965,013	24.2	1,478.7
3	İspanyolca	164,968,742	7.8	807.4
4	Japonca	99,182,000	4.7	110.7
5	Portekizce	82,586,600	3.9	990.1
6	Almanca	75,422,674	3.6	174.1
7	Arapça	65,365,400	3.3	2,501.2
8	Fransızca	59,779,525	3.0	398.2
9	Rusça	59,700,000	3.0	1,825.8
10	Korece	39,440,000	2.0	107.1

İnternette yer alan dokümanların dillerine bakıldığında, 2010 yılında yapılan bir çalışma %26.8'inin İngilizce, %24.2'sinin Çince dilinde ve %0.8'inin Türkçe dilinde olduğunu göstermiştir. Bu bulgular ışığında İnternet boyutunda bulunan dokümanlardan %73.2'lik bir bölümünün İngilizce olmayan dokümanlar olduğunu anlıyoruz. Yine bu veriler Türkçe dili açısından değerlendirildiğinde İnternet boyutunda bulunan dokümanlardan %99.2'lik bir bölümünün Türkçe olmadığı görülmektedir.

İnternetin bu muazzam kapasitesinde dokümanlar elbette tekil değildir. Birçok bilginin gerek orijinal dilinde gerek başka dillerde, kopyaları, alıntıları, benzerleri, tercümeleri ve intihalleri mevcuttur. Araştırmacılara, aranılan bilgilerin orijinal dilinden farklı dillerde sunulması da arama motorlarınca sunulan bir hizmettir. Bu sayede bilgiye erişim zenginliği artmaktadır. Bir dilde yazılmış belgenin başka dillerde çevirilerin ve alıntılarının bulunması, bu bilgilerin web

ortamında bulunması ve sonuçların görüntülenme süreleri tek dilli kişiler açısından da faydalı olacaktır. Ayrıca dokümanların telif hakkı veya intihal gibi durumlar açısından da farklı dillerde benzerlerinin bulunması bir kolaylıktır.

Bu tez çalışması kapsamında, farklı dillerdeki dokümanların benzerlerinin bulunması için yeni yöntemler geliştirilmiştir. Bu yöntemler ile dokümanlar, metin formatından vektörel hale çevrilmiş, dönüşen bu öznitelik vektörleri ile farklı dillerdeki dokümanların içerik olarak benzerlerinin saptanmasına dair literatürde yer alan algoritmalar ile benzerlik ölçümleri yapılmıştır. Ayrıca yeni bir benzerlik yöntemi geliştirilmiş ve geliştirilen öznitelik vektörü çıkartma yöntemlerine, testler sonucunda en pozitif ölçümü yaptığı tespit edilmiştir. Farklı dillerde ama benzer içeriğe sahip dokümanların diğer dokümanlar arasından belirgin bir şekilde üst sıralarda saptandığı tespit edilmiştir. Ayrıca farklı dilde olsa da benzer içeriğe sahip ve intihale neden olan belgelerin saptanması çalışmalarında da yardımcı olacak yöntem olarak akademisyenlere yön vereceği düşünülmüştür.

2. KAYNAK ARAŞTIRMASI

1990'ların ortasından bu yana çok-dilli bilgi erişimi önemli bir araştırma alanı olmaya başlamış ve açık bir çalışma alanıdır [5].

Bu konuda çalışma yapanlar Makine Öğrenme (Machine Learning), Çarpraz Dil (Cross-Lingual) veya Çok Dilli Bilgi Erişimi (Multilingual Information Retrieval) (MLIR) alanlarında araştırmaya yönelmişlerdir. Ama web ortamındaki belgelerin belirgin bir şeması olmamasından dolayı duyarlık (precision) ve geri çağırım (recall) açısından tatmin edici sonuçlar elde edilememiştir.

Doküman benzerliğinin tespitinde, tek dilli (monolingual) veya farklı dillerdeki (interlingual) dokümanların benzerliğinin saptanması genelde 2 aşamalıdır. Öncelikle belgelerin bir Öznitelik Vektörünün (Feature Vectors) oluşturulması sağlanır ve daha sonra bu örnek uzayının sınıflandırılması (Classification) veya kümelendirilmesi (Clustering) yapılır [5].

Benzerlik tespitinin yapılabilmesi için öznitelik uzaylarının oluşturulması gerekir ve dokümanlar; harf, kelime veya cümlelerden ziyade sayısal değerler halinde vektör uzayları olarak temsil edilirler. Sayısal vektörler için doküman veya metin karakterini yansıtacak veriler kullanılır. Bu sayısal değerler Salton'ın VSM (*Vector Space Model)* 'e dayanır [6].

Sınıflandırma veya Kümeleme için kullanılacak belgeler ve normalizasyon işlemlerinden sonra vektörlere çevrilir. Doküman boyutunun ve sayısının çokluğundan dolayı, bu ön işlemler vektör boyutunu azaltan ve işlem süresini hızlı hale getiren yöntemler olarak tercih edilir.

Literatürde yapılan araştırmalarda farklı dillerdeki dokümanların benzerlerinin bulunması için birçok farklı yöntem uygulanmıştır.

Savoy yaptığı çalışmada sorguları indeks'teki farklı dillere çevirerek, farklı dilde ama aynı içerikte doküman getirmeyi denemiştir [7].

Benzer konuda Steinberger tarafından 2002 yılında yapılan diğer bir çalışmada ise EUROVOC® eş-anlamlılar (thesaurus) sözlüğünü kullanarak farklı dillerdeki dokümanların benzerliğini bulmaya dayalı bir yöntem kullanılır [8]. Ama

çalışmalarında kendilerinin de belirttiği gibi kullanılan sözlüğün kapsama alanının darlığı çalışmayı kısıtlamıştır.

Mathieu ve arkadaşları [9] tarafından yapılan çalışmada ise belgelere dilbilimsel analiz yapılıp ekler ve özel isimler (yer, firma, tarih vs.) kaldırılır. Daha sonra terimlerin ağırlıkları hesaplanır. Daha sonra çift-dilli sözlükler kullanarak belge benzerliği saptanır.

Pouliquen [10] coğrafik özel isimlerin farklı dilde belgelerin benzerliğinde kullanılabileceğini söylemiş, Buradan yola çıkan Steinberger ve arkadaşları belgelerde geçen coğrafi isimlerin, kuruluş isim ve kısaltmalarının, tarih ve sayısal yüzdelerini özel isimlerin farklı dilde belgelerde de benzer olabileceğini belirtmişlerdir. Başka araştırmacılarda özel isimlerden doküman benzerliği bulma konusunda çalışmalar yapmışlardır [11-12]. Ama ortak özel isimlere sahip olmayan benzer içerikli belgeler ise bu araştırmanın kapsamı dışında kalmaktadır.

Lee ve Yang, bu konuda yaptığı çalışmalarda Self-Organizing-Map (SOM) yöntemini kullanmışlardır. SOM genelde sınıflandırma ve dağılım yapmak için kullanılan yetenekli yapay sinir ağlarıdır [13].

Bu tez çalışması kapsamında kullanılacak algoritma yöntemleri olarak "doküman hizalama" ile ilgili çalışmalara bakıldığında bu konuda 1990 yıllardan itibaren çalışılmaya başlandığı görülmüştür [14-16].

Bu konuda yapılan araştırmalar ikiye ayrılmaktadır. Birinci grup araştırmacılar, hizalama işleminden sonra sözlük kullanımı ile karşılıklı çeviri başarı oranına bakılaraktan doküman benzerliğinin saptanması yöntemleri üzerlerinde durmuşlardır. İşlem gücü ve performans açısından sıkıntılar ile karşılaşılmıştır. İkinci grup araştırmacılar ise istatiksel hesaplamalar üzerinde durmuşlardır. Performans açısından başarılı olsa da dil benzerliği ve alfabe türleri kullanılan veri setlerinde önemli rol oynamıştır [17].

Brown ve arkadaşları [18] İngilizce ve Fransızca paralel metinler içeren Kanada Parlamentosu görüşmeleri üzerinde cümle hizalama yöntemi ile çalışmalar yapmışlardır. Cümlelerin karakter uzunlukları benzerliklerini çalışmışlardır. Fakat bir dildeki cümle uzunluğunun başka bir dilde çok kısa bir şekilde tanımlanması hata oranını yükseltmiştir.

Yine aynı yıllarda Kay ve arkadaşları tarafından paralel dokümanlarda "hizalama" çalışmaları yapılmıştır [15]. Yöntemleri istatistiksel hesaplamaların fazlalığından dolayı yavaş ve hata oranlarının verilmemesi dezavantaj teşkil etmektedir.

Gale ve Church, paralel dokümanlar arasında cümle hizalama yöntemi ile cümle ve karakter uzunlukları arasında istatistiksel hesaplarla çalışmışlardır. Bu yöntemde çeviri ne kadar başarılı ve test edilen dillerin morfolojik yapısının benzerliğinin sonucu pozitif anlamda etkilediklerini söylemişlerdir. Dillerin farklı yapılarının istatistiksel sınırları belirlediğini gözlemlemişlerdir. Örneğin İngilizce-Almanca testlerinin İngilizce-Fransızca çevirilerine göre daha başarılı olduğunu söylemişlerdir [14].

Literatürde tez çalışmamıza yakın yöntemler olarak Vu ve arkadaşları yaptıkları çalışmada paralel dokümanlardaki terim frekanslarının benzer şekilde bir ilgileşim göstereceğini belirtmişlerdir. İngilizce-Çince-Malay veri setlerinde yaptıkları çalışmada bir dokümandaki terim frekansları dağılımına Ayrık Fourier Dönüşümü uygulamışlar ve Pearson Korelasyon Katsayısı yöntemi ile Öznitelik vektörleri test edilmiştir. İstatistiksel hesaplamalara dayalı kalsalar da performans artırımı için sözlük kullanılmıştır [19].

Bir başka benzer çalışma Tao ve Zhai tarafından yapılmıştır. Kelimelerin dağılımının değişik dillerdeki paralel dokümanlarda da benzer dağılımda olacağını öne sürmüşlerdir. Sadece istatistiksel hesaplamaların yapıldığı bu çalışmada paralel dokümanlar arasındaki frekans korelasyonu, Pearson Korelasyon Katsayısı ile hesaplanmıştır. Hesaplamalarında Okapi BM25 ve IDF değerleri ile terimlere dair ağırlık verisi ile algoritmalarını güçlendirmişlerdir. Dillerin yapısındaki farklılıklar ve sınırlar istatiksel hesaplamalarda beklenmeyen sonuçlar çıkmasını sağlamıştır [20].

Başka bir çalışmada Munteanu [21], Romence-İngilizce dokümanlar üzerinde yaptığı çalışmada, derlemdeki bütün dokümanları LEMUR [22] yazılımı veri tabanına ekliyor ve daha sonra her kelimeyi birebir çevirerek, çevrilen kelimelerin Log-likelihood skoru ile eşdizimlilik yöntemleri hesaplanıyor ve bunun neticesinde hizalama hatalarını düzeltip benzerlik ölçümü yapıyor.

Bu yüksek lisans tezi çalışmasında farklı dillerdeki belgelerin benzerliğinin tespiti konusunda öznitelik vektörlerinin çıkarım yöntemlerine dair benzer çalışmalar bulunsa da belirli bir domain üzerinde lokal bazlı hesaplamalar yapan çalışmalara rastlanmamıştır. Ayrıca Öznitelik vektörlerinin kıyaslanması konusunda literatürde yer alan benzerlik ölçümlerinin yanı sıra literatürde olmayan yeni bir benzerlik ölçümü geliştirilmiştir. Yapılan benzerlik ölçümü ile lokal olarak sadece bir dokümana ait veriler hesaplanmış ve hesaplama süreleri bayağı düşürülmüştür.

3. MATERYAL VE YÖNTEM

3.1. İNTERNET DÖKÜMANLARI VE KULLANILAN DİLLER

3.1.1. Diller ve Veri Setleri

İnternet ortamında kullanılan değişik dillerin kullanım yüzdeleri Çizelge 3.1'de verilmiştir. Bu diller kodlama veya morfolojik olarak farklı yapılara sahip olabilirler. Bu çizelgeden anlaşıldığı üzere, İngilizce dağılımı diğer dillere göre belirgin bir üstünlük sağlasa da gün geçtikçe İngilizce olmayan dokümanların da arttığı gözlemlenmiştir. Bir diğer ilgi çekici nokta ise Türkçe dokümanların bu büyük uzayda çok küçük bir miktarda yer almasıdır. % 1.4'lük oranla Türkçe dokümanlar çok az yer kaplamaktadır.

Çizelge 3.1. 30 Temmuz 2013 itibariyle ile web sitelerinin kullandıkları dillerin yüzdesi [23].

Dil	Kullanım Yüzdesi	Dil	Kullanım Yüzdesi
İngilizce	55.4	İsveççe	0.6
Rusça	6.4	Endonezya Dili	0.5
Almanca	5.4	Vietnamca	0.4
Japonca	4.9	Romence	0.4
İspanyolca	4.3	Yunanca	0.4
Çince	4.1	Korece	0.4
Fransızca	3.8	Macarca	0.3
Portekizce	2.3	Danca	0.3
Lehçe	1.8	Tay Dili	0.3
İtalyanca	1.6	Fince	0.2
Türkçe	1.4	Slovakça	0.2
Felemenkçe	1.2	Bulgarca	0.2
Arapça	0.9	Norveççe	0.2
Persçe	0.8		
Çekçe	0.7		

3.1.2. Diller ve Yapısal Özellikleri

Bu tez çalışmasında kullanılan veriler için dillerin morfolojik yapıları göz önünde bulundurulmuştur. Morfoloji, kelimelerin yapısının tanımlanmasını, çözümlenmesini kelimelerin yapısının tanımlanmasını, çözümlemesini ve etiketlenmesini inceleyen bilim dalıdır.

Biçimbilim; modern dilbilimin bir alt alanıdır ve bir dilin anlam taşıyan en küçük parçalarının (biçimbirim) araştırmasını yapar. Biçimbirimler farklı biçimlerde kullanılır, anlam ayırıcı en küçük birimlerden oluşur ve bunların kelimelerini oluşturur. Biçimbirim kelimelerin içyapısındaki dil olgularına ilişkin kurallarla ilgilenir.

Kelimeler genellikle en küçük yazım birimlerine göre kabul edilmelerine rağmen, çoğu dilde kelimeler diğer kelimelere kurallar ile bağlı olabilmektedir. Örneğin; Türkçe ’de *göz, gözlük, gözlükçü, gözlükçülük* birbirine belirli bir kural ile bağlıdır. Göz yerine başka bir kök kelime konularak da aynı kural işletilebilir. Mesela k*ulak-kulaklık, diz-dizlik*, vb. Göz tekil anlamını da barındırırken çoğul yapılması için gözler denilir. Morfoloji dilbiliminde dillerdeki bu yapısal kuralların ve örnek yapıların anlaşılması için çalışan koludur.

3.1.3. Kullanılan Veri Setleri ve Türleri

Veri seti, NATO tarafından çıkarılan ‘NATO Review’ dergisinin 2000 ile 2008 yılları arasında çıkardığı makalelerden toplanan paralel metinlerden oluşmaktadır [24]. NATO Review dergisi, NATO üyesi ülkelerin siyaset, politika ve ekonomi konularında hazırlanmış makalelerinden oluşmaktadır. Dergideki makaleler 17 dilde paralel çeviri olarak yayınlanmaktadır.

Oluşturduğumuz web örümceği (crawler) yazılımı ile 247 adet makale 8 dilde indirildi. Almanca, İngilizce, İspanyolca, Fransızca, İtalyanca, Flemenkçe, Portekizce ve Türkçe dillerinden oluşan bu veri setinde 37 MB boyutunda 1991 adet paralel doküman elde edildi. Bu makaleler text dosya halinde yazıldığı dilde Çizelge 3.2 de belirtilen kod ile klasörler haline getirilmiştir.

Her dil için 1 doküman kullanılmıştır. Doküman uzantısı olarak metin dosya formatları ve UTF–8 kodlaması kullanılmıştır. UTF, Unicode Transformation Format kelimelerinin İngilizce kısaltmasıdır. Evrensel Kod karakterleri normal baytlardan oluşan metinlerde kullanabilmek için geliştirilmiş kodlama türüdür. 0–127 arası standart ASCII karakterleri 1 bayt ile gösterilirken, özel karakterleri de barındırıp, her türlü dili desteklemek için kullanılır. Her harf için kullanılan minimum bit sayısına göre 8–16–32 diye çeşitlendirilebilir. Bu metin dosyaları için isim olarak içerdiği dile ait kodlama kullanılmıştır.

Çizelge 3.2. Veri setinde kullanılan diller ve kısaltma kodları

Dil adı	Kullanım kodu
İngilizce	en
Almanca	de
Fransızca	fr
İtalyanca	it
İspanyolca	es
Portekizce	pt
Türkçe	tr
Flemenkçe	nl

Çizelge 3.3. Veri setinde kullanılan dokümanların özellikleri

Klasör Kodu	Türü
000-099	Avrupa Birliği Dokümanları
100-347	Nato Review dergisi
500	Melez dokümanlar
600	Türkçe-İngilizce melez metinler
700	Türkçe-İngilizce paralel metinler
800	Farklı üçlü benzerlikte dokümanlar
900	Farklı dökümanlar

Çizelge 3.4. 181 numaralı veri setinde doküman hizalaması ve terim dağılım örneği

İngilizce	Türkçe
Anton Tus became the first Croatian ambassador to NATO after Croatia joined the Partnership for Peace in May 2000. A professional soldier, hewas head of the Yugoslav Airforce until he resigned in June 1991. In September 1991, he became the first chief of staff of the Croatian Armed Forces, building them from scratch and leading the defence of Croatia during the 1991-92 war. A retired five-star general, he aims to guide Croatia into NATO. Why does Croatia aspire to join NATO? NATO membership is in Croatia's national interest. NATO, and NATO alone, offers my country the highest possible level of security, defence of the country's independence, territorial integrity, national and state identity. Here, it is worth pointing out that ten years ago, the fact that Croatia was a member of both the United Nations and the predecessor of the Organisation of Security and Cooperation in Europe failed to protect us from aggression and war. In addition, NATO today is the critical security and defence organisation in the Euro-Atlantic area and together with Russia and other Partners the basis for peace and security in the northern hemisphere. NATO membership can also help speed completion of democratic and economic reform within Croatia and, in this way, help the country enter the European Union. The alternative to NATO membership is far from ideal. It would mean less security, increased defence expenditure, an absence of allies and greater isolation. What can Croatia offer the Alliance?A stable and democratic Croatia can contribute to the stability of its surrounding region and the whole of Southeastern Europe.	Anton Tus, Hırvatistan'ın Mayıs 2000'da Barış İçin Ortaklık'a katılmasından sonra ülkenin NATO nezdindeki ilk büyükelçisi oldu. Profesyonel bir asker olan Tus, Haziran 1991'de istifa edene kadar Yugoslav Hava Kuvvetleri Başkanı olarak görev yaptı. Eylül 1991'de Hırvatistan Silahlı Kuvvetleri'nin ilk Genel Kurmay Başkanı oldu; bu görevi sırasında silahlı kuvvetleri sıfırdan başlayarak inşa etti ve 1991-92 savaşında Hırvatistan savunmasını yönetti. Beş yıldızlı bir emekli general olan Tus, Hırvatistan'ın NATO'ya girmesini amaçlamaktadır. Hırvatistan'ın NATO'ya girmek istemesinin nedeni nedir? NATO üyeliği Hırvatistan'ın ulusal çıkarları doğrultusundadır. Hırvatistan'a mümkün olan en üst düzeyde güvenliği, ülkenin bağımsızlığının, toprak bütünlüğünün, ulusal ve devlet kimliğinin korunmasını sadece NATO sunabilir. Bu noktada, on yıl önce Hırvatistan'ın hem Birleşmiş Milletler'e hem de Avrupa Güvenlik ve İşbirliği Konferansı'na üye olmasının bizi saldırganlık ve savaştan koruyamamış olduğuna işaret etmeliyim. Ayrıca NATO bugün Avrupa-Atlantik bölgesinin en önemli güvenlik örgütüdür ve Rusya ve diğer ortaklarla birlikte kuzey yarı kürenin barış ve güvenliğinin temelidir. NATO üyeliği ayrıca Hırvatistan'daki demokratik ve ekonomik reformların hızlanmasına ve ülkenin Avrupa Birliği'ne girmesine yardımcı olabilir. NATO üyeliğinin alternatifi pek hoş değildir. Bu daha az güvenlik, artan savunma harcamaları, hiç müttefikimizin olmaması ve soyutlanma demektir.Hırvatistan İttifak'a ne sunabilir?İstikrarlı ve demokratik bir Hırvatistan çevresindeki bölgenin ve güneydoğu Avrupa'nın tümünün

3.1.4. Etkisiz Kelime Listeleri

İngilizcede "Stop words" diye geçen Etkisiz kelimeler bir dilde çok sık kullanılan (Türkçe 'de "bir", "ama", "ve" "bu", "şu" gibi) kelimelerdir. Anlamsal olarak dokümana olan katkıları göz ardı edilir. Doküman vektörleri oluşturulurken tüm dokümanlarda genelde aynı ortalama sıklığa sahip olduklarından dolayı pozitif bir katkı sağlamazken aksine benzerlik ölçümlerinde benzer veya negatif sonuçlar göstermektedir. Etkisiz kelimelerin göz ardı edilmesi doküman karakteristiğini yansıtan kelimelerin de ön plana çıkmasını ve bu kelimelerin ağırlık değerlerinin yükselmesini sağlar. Etkisiz kelime listeleri dokümanlarda boyut indirgeme işlemi için de tercih edilir. Bu sayede işlem süresi azalır.

Çizelge 3.5. Diller ve kullanılan etkisiz kelime sayısı

Dil adı	Etkisiz kelime sayısı
İngilizce	571
Almanca	603
Fransızca	464
İtalyanca	399
İspanyolca	351
Portekizce	392
Türkçe	223
Felemenkçe	104

3.2. VERİ SETLERİNİN HAZIRLIĞI VE DÜZENLENMESİ

Veri setinde yer alan dokümanların öznitelik vektörleri oluşturulmadan önce bazı işlemlerden geçirilmesi gerekiyor. Öncelikle dokümanlardaki karakter setleri UTF–8 olarak düzenlenmektedir.

Daha sonra dokümanda bulunan özel noktalı karakterleri normalize ediyoruz. Mesela é harfi e harfine dönüştürülüyor. Bu karşılaştırma ve ölçümlerde benzer dokümanların değerlerini güçlendiriyor.

Bu işlemden sonra terimleştirme işlemine başlıyoruz. Terimleştirme bir metni bir dizi kelime, deyim, sembol veya anlamlı katarlar haline getirme işlemidir. Terimleştirme işleminden sonra oluşan vektörler dil bilimi veya veri madenciliğinde kullanılır.

Dokümanları terimleştirirken düzenli ifadelerden (regular expression) faydalanıyoruz. Terimleştirme adımlarımız şu şekildedir;

1.adımda tüm noktalama işlemlerinin yerine boşluk ekliyoruz.

2.adımda dokümandaki bütün fazla boşlukları ve beyaz boşlukları (white-space) kaldırıyoruz.

3.adımda dokümandaki özel karakterleri kaldırıyoruz. Alfa nümerik olmayan ve düzenli ifadeler ile her boşluk arasındaki bir kelimeyi bir terim olarak kabul ediyoruz.

4.adımda dokümandaki bütün harfleri küçük harflere çeviriyoruz.

Terimleştirme işlemi, kelime sınırı bulunmayan dillerde zor bir işlemdir. Bu yüzden, bu tez çalışmasında Çince gibi diller, veri seti ve benzerlik ölçümlerinde kullanılmamaktadır.

Kullanılan veri setlerinin öznitelik vektörlerine dönüştürülmeden önce bazı normalizasyon ve optimizasyon işlemlerinden geçirilmesi gerekmektedir. Bu nedenle öncelikle dokümanlar klasörlere göre kümelenmektedir. Bu klasörler içerisinde farklı türde dosyalar bir araya getirilmiştir. Bazı klasörlerin içerisinde yer alan dokümanların hepsi aynı içeriğe sahip iken bazı klasörlerdeki veriler ise karışık bir şekilde dağıtılmıştır. Bu sayede benzer ve benzemez dokümanlardan pozitif-negatif sonuçlar beklenmiştir.

Çizelge 3.6. 733 numaralı veri seti önişleme aşamaları

Doküman Parçası	NATO'nun 1995 yazında Bosna ve Hersek'teki müdahalesi İttifak için bir dönüm noktası olmuştur. Başlangıçta NATO Birleşmiş Milletlerin uyguladığı silah ambargosuna ve uçuşa kapalı hava sahası uygulamasına destek vermek ve askeri olası durum planlaması sağlamak üzere Bosna ve Hersek'te bulunuyordu. Alınan bu önlemler çatışmanın yumuşamasına ve hayatların kurtarılmasına yardımcı olmuş ama savaşı sona erdirmeye yetmemişti. Buna karşılık NATO'nun 12 günlük hava harekatı 20 Aralık 1995'ta imzalanan ve Bosna Savaşına son veren Dayton Barış Anlaşması'na giden yolu açmıştır. Bu anlaşma çerçevesinde NATO ilk kez kuvvet konuşlandırmış ve 60,000 kişilik Uygulama Gücü'nün (IFOR) liderliğini üstlenmiştir.
Önişleme Sonrası	natonun 1995 yazında bosna ve hersekteki müdahalesi ittifak için bir dönüm noktası olmuştur başlangıçta nato birleşmiş milletlerin uyguladığı silah ambargosuna ve uçuşa kapalı hava sahası uygulamasına destek vermek ve askeri olası durum planlaması sağlamak üzere bosna ve hersekte bulunuyordu alınan bu önlemler çatışmanın yumuşamasına ve hayatların kurtarılmasına yardımcı olmuş ama savaşı sona erdirmeye yetmemişti buna karşılık natonun 12 günlük hava harekatı 20 aralık 1995ta imzalanan ve bosna savaşına son veren dayton barış anlaşmasına giden yolu açmıştır bu anlaşma çerçevesinde nato ilk kez kuvvet konuşlandırmış ve 60000 kişilik uygulama gücünün ıfor liderliğini üstlenmiştir

Çizelge 3.6. (devamı) 733 numaralı veri seti önişleme aşamaları

Gövdeleme sonrası	natonun 1995 yaz bosna ve hersek müdahale ittifak için bir dönüm nokta ol başlangıç nato birleşmiş millet uygu silah ambargo ve uç kapalı hava saha uygu destek ver ve asker ol durum plan sağ üzeri bosna ve hersek bul al bu önlem çatışma yumuşak ve hayat kurtar yardım ol ama savaş son erdirme yetme buna karşı natonun 12 gün hava harekat 20 aralık 1995 imza ve bosna savaş son ver dayton barış anlaşma git yol aç bu anlaşma çerçeve nato ilk kez kuvvet konuş ve 60000 kişi uygu güç ıfor lider üst
Etkisiz kelimelerin çıkartılması sonrası	natonun 1995 yaz bosna hersek müdahale ittifak dönüm nokta başlangıç nato birleşmiş millet uygu silah ambargo uç kapalı hava saha uygu destek asker durum plan sağ bosna hersek önlem çatışma yumuşak hayat kurtar yardım savaş son erdirme yetme natonun 12 gün hava harekat 20 aralık 1995 imza bosna savaş son ver dayton barış anlaşma git yol aç anlaşma çerçeve nato kuvvet konuş 60000 kişi uygu güç ıfor lider üst

Terimleştirme işleminden sonra kelimelerden oluşan bir dizi elde ediyoruz. Daha sonra bu dizideki kelimelere Gövdeleme işlemi yapılıyor ve yeni düzenlemeden sonra Etkisiz Kelimeler çıkartılıyor. Etkisiz Kelimelerin dışında ayrıca dokümanda 2 karakterden az olan kelimeleri de çıkarıyoruz. Bu işlemlerden sonra vektörün indisi kelimenin doküman içerisinde aldığı pozisyonu, değeri ise terim değerini veriyor.

$$D_i = \{ t_1 , t_2 , t_3 , t_4 \dots t_n \} \quad (3.1)$$

Test edilen veri seti klasörü içerisinde bulunan tüm dil dokümanlarına bu işlem uygulanarak her dil için Eş.(3.1)'deki pozisyon ve terimlerden oluşan D_i vektörünü elde ediyoruz.

3.2.1. Küçük Harfe Dönüşüm

Dokümanlarda yer alan terimler öznitelik vektörlerimiz için anahtar değerler olmaktadır. Bu terimler vektör uzaylarında sayısal ifadelere çevrilmiştir. Bu yüzden hepsinde standart bir karakter düzenlemesi gerekmektedir. Örneğin; bir dokümanda geçebilecek "üniversite", "Üniversite", "ÜNİVERSİTE" kelimeleri aslında aynı değere (sıklığa) işaret etmektedir. Dolayısıyla bu anahtar değer tek bir terim (*t*) ile temsil edilmeli ve terim frekansları (*f*) tek bir değere atanmalıdır. Bunun için bütün terimlere genel bir dönüşüm için küçük-harfe-dönüşüm işlemi uygulanmıştır. Metinde yer alan harfleri küçük harfe dönüştürme işlemi vektör boyutlarının da azalmasını sağlamıştır. Bu aşamadan sonra standart bir forma dönüşen terimler, Karakter Normalizasyonu, Gövdeleme ve Etkisiz Kelimeler işlemine daha uygun bir hale gelir.

3.2.2. Karakter Normalizasyonu

Bazı dillerde yer alan özel karakterler dönüşüme uğratılarak karakter normalizasyon işlemi uygulanmıştır. Bu karakterler özellikle noktalama işaretli, aksanlı veya Latin alfabesinde yer almayan karakterlerdir. Örneğin bir terim içerisinde geçen "é" karakteri "e" karakterine dönüştürülmüştür.

Karakter normalizasyonu, sonuçların doğruluğunu artıran bir işlemdir. Bu sayede benzer bir dokümanda aynı anlama gelen birden fazla terim, tek indiste ifade edilmiş oluyorlar ve benzerlik ölçümlerindeki veri kirliliği (noise) veya yanlış sapmalar, azalmış oluyor.

Örneğin; Almanca bir dokümanda, örnek bir kelime, bazı dokümanlarda *"schoen"* diye geçerken, diğer dokümanlarda *"schön"* şeklinde geçiyor olabilir. Türkçe bir dokümanda bazı pozisyonlarda *"ağaçlar"* diye geçen bir kelime başka bir yerde benzer konudan bahsederken *"ağacı"* şeklinde geçiş yapabilir. Veya bazı dillerde, bir dokümanda *"Æ"* şeklinde geçen bir karakteri *"ae"* şeklinde ifade etmek gerekebilir. Anlamsal açıdan aynı vurguya sahip olabileceğinden dolayı boyut indirgeme işlemine uygun bir yöntemdir.

3.2.3. Gövdeleme

İngilizcede "*Stemming"* olarak adlandırılan metin-doküman önişleme safhalarından birisidir. Kelimelerin dokümanda kullanım şeklinden ziyade, sözlükteki kök kullanımına indirgenmiş şeklidir.

Gövdeleme işlemi Türkçe gibi eklemeli dillerle yapılan aramalardaki katkısı büyüktür. Türkçedeki kelimelerin morfolojik yapısı çoğunlukla "kök+yapım eki+çekim eki" şeklindedir. Gövdeleme, çekim eklerinin kelimeden çıkartılması olarak da düşünülebilir.

İnternet gibi çok büyük sayılarda belgenin bulunduğu bir ortamda gövdeleme, işlem süresini kısaltan bir uygulamadır. Gövdeleme yapılması benzer içeriğe sahip dokümanlar için mesafe ölçümlerini düşürmekte veya benzerlik skorunu yükseltmektedir.

3.2.4. Etkisiz Kelimeler

Bu tez çalışmasında amaçlanan yönteme uygun olarak, dokümanlarda etkisiz kelimelerin kaldırılması yapılan testler sonunda pozitif neticeler vermiştir. Bir dokümana ait özellikli kelimelerin, o dokümanın karakteristiğini yansıttığını düşünecek olursak, bu karakteristik kelimelerin etkisiz kelimeler dışında olması çalışmayı test ve hesaplamalar aşamasında daha kararlı bir hale getirmektedir. Tez çalışmasında öne sürülen algoritma gereği yapılan hesaplamalar ile o dokümanda en sık geçen kelimelerin yerleşimi, bu kelimelere ait ağırlık değerleri ve komşuluk ağırlıkları ile benzer manaya sahip farklı dildeki doküman için de aynı olduğunu düşünecek olursak, etkisiz kelimelerin çıkartılması burada çok önemli bir rol oynamaktadır. Aksi takdirde test edilen ve karşılaştırılan her dokümanda sık geçen kelimeler için etkisiz kelimeler dominant olacak ve kıyaslamalar neticesinde her dokümanın birbirine benzer olduğu sonucu çıkabilir.

Etkisiz kelimelerin çıkartılması doküman uzunluğunu azaltırken aynı zamanda dokümana ait özellikli kelimeleri ön plana çıkartarak dokümanların birbirine benzerliğini/benzeşmezliğini daha güçlü vurgulamaktadır. Vektör

boyutunun etkisiz kelimelerin çıkartılması ile işlem karmaşıklığı ve süresi azalmaktadır.

Örneğin; veri setinde yer alan "000" numaralı klasörde, İngilizce, Almanca ve Fransızca 3 doküman bulunmaktadır. Bu üç doküman "Avrupa Birliği Hakkında Genel Bilgi" içeriğine sahiptirler [25].

Çizelge 3.7 000-Numaralı veri seti için "Avrupa" kelimesine dair geçiş sıklıkları

Almanca	İngilizce	Fransızca
europaischen 6 europaische 3 europaerinnen 1 europaer 1	european 3 europeans 1 europe 1	europeens 8 europeenne 5 europeen 4 europeennes 1
Toplam: 11	Toplam: 5	Toplam: 18
Toplam terim sayısına oranı	Toplam terim sayısına oranı	Toplam terim sayısına oranı
0.0506	0.2057	0.0567

Bu dokümanlarda küçük harf dönüşümü, etkisiz kelimelerin çıkarılması ve doküman uzay vektörüne çevrildikten sonra Çizelge 3.7'de sıklık listesine bakıldığında; Almanca dokümanda içinde "Avrupa" kelimesi anlamı geçen 11, İngilizcede 5, Fransızcada ise 18 adet kelime görülmektedir. Rakamlar arasındaki farklılık terim sayısına orantılığında Almanca ve Fransızca dokümanlarda yakın bir oran olarak çıkmaktadır.

Çizelge 3.8 NATO veri setinde 101 numaralı veri seti için gövdeleme ve etkisiz kelime çıkartımı işlemi sonunda doküman istatistikleri

Dil	Toplam kelime sayısı	Gövdeleme ve etkisiz kelime çıkartımı sonrası	Yüzde
de	2455	839	34.17
en	2187	774	35.39
es	2491	959	38.49
fr	2597	952	36.65
it	2276	893	39.23
pt	2376	939	39.52
tr	1932	1035	53.57
nl	2418	1757	72.66

Çizelge 3.8 de görüldüğü gibi Almanca (de) doküman 2455 terimden 839 terime düşmüştür. Etkisiz kelimelerin çıkartılması neticesinde geriye kalan 839 kelime bu dokümana özgü kelimeler olarak kalmaktadır ve öznitelik olarak bu dokümanı yansıtmaktadır.

3.3. ÖZNİTELİK VEKTÖRÜ

Dokümanları nicel veya nitel olarak tanımlayan verilere *"öznitelik"* denir. Oluşturulan, Öznitelik Vektörleri (ÖV) ile dokümanlarının birbirine benzerlik veya benzeşmezlik ölçümü yapılabilir.

Bu tez çalışmasında yer alan özgün yöntem, Öznitelik Vektörlerinin oluşturulma yöntemidir. Uygun seçilmiş ve oluşturulmuş öznitelikler, dokümanların karakteristik özelliklerini barındırırlar. Burada amaç dokümana ait bilgiyi mümkün olan en küçük boyuta indirgeyerek ÖV'ye taşımaktır. Bu sayede m>n şeklinde düşünecek olursak, doküman m-boyutlu bir durumdan n-boyutlu bir duruma indirgenerek, işlem süresi azaltılır. ÖV karakteristiği boyut indirgeme işlemleri ile dokümanı daha iyi tanımlar hale gelir.

Oluşturulan ÖV, dokümanı ifade edişi, diğer dokümanlar içinde sağladığı karakter mesafe ölçümleri ile doğrusal olmalı ve benzeyen dokümanlar için benzer vektörler bulundururken, benzemeyen dokümanlar için mesafe ölçümleri ile ayırt edilebilir ve uzak vektörler oluşturulmalıdır.

ÖV bir nevi dokümanımızın kelime dizilimi gibi oluyor ve onu benzer dokümanlarla bu şekilde kıyaslayabiliyoruz. Öznitelik vektörünün dokümanın karakteristiğini, uç ve sınır noktaları temsil etmesi, algoritmanın başarısı açısından öenmli olacaktır.

3.3.1. Öznitelik Vektörü Ölçütleri

Bu tez çalışmasında anlamsal benzerlikten ziyade yapısal benzerlik ölçümleri yapılacaktır. Bu yüzden dokümanlar öznitelik vektörlerine çevrilirken sayısal verilerle ifade edilerek benzerlik aranacaktır. Bu sayısal ifadelerin anlamlı ve doküman karakterini yansıtabilmesi için kelimeler yerine temsil ettiği sayısal değerler alınacaktır. Örneğin dokümanda terim frekansları, terim ağırlığı, MI Skor, Log-Likelihood Oranı (*LLR*) değeri gibi ifadelerle bir pozisyonda bulunan kelimenin gücünün anlamsal benzeri diğer doküman içinde aynı olduğu varsayılmaktadır.

3.3.1.1 . Terim Frekansları

Spesifik bir konudan bahseden bir doküman içerisinde bulunan özel kelimelerin sayısı başka bir dildeki paralel çevirisinde-benzerinde doküman içerisinde de baskın olacaktır [19]. Bu da iki doküman arasında benzerlik açısından bağ kurmaya yeterli bir parametre olacaktır.

Tez çalışması içerisinde dokümanlar düzenlendikten ve vektörler elde edildikten sonra her doküman için bir sözlük oluşturulmaktadır. Hangi pozisyonda, hangi kelimenin olduğu ve ilgili kelimenin doküman içerisindeki terim frekansı hesaplanmaktadır.

Terim frekansı için önişlemden geçirilmiş bir dokümanda tüm kelimeler terim dizisine atıldıktan sonra her kelimenin doküman içerisinde kaç kere geçtiği hesaplanır. Terim frekansı (*f*) değeri ÖV için bir öznitelik değeridir.

Terim frekanslarının uygulama için birden fazla yararı vardır; öncelikle frekansı düşük terimler göz ardı edilerek, boyut indirgenmesine gidilmekte ve bu sayede işlem süresi kısalmaktadır, ayrıca herhangi bir doküman içerisinde çok nadir geçen bir kelimenin doküman karakterini pek yansıtmayacağı düşünülmüştür. İkinci olarak belirlenmiş bir limitin üzerindeki frekansa sahip kelimelerin hesaplanması ölçümler açısından daha faydalı neticeler vermektedir. Burada dikkat edilmesi gereken husus Etkisiz Kelime optimizasyonun düzgün ve etkili bir şekilde uygulanmış olmasıdır. Aksi takdirde sık geçen bir kelime tüm dokümanlar içerisinde baskın hale gelecektir. Bağlaçlar gibi her dokümanda fazlasıyla yer alan kelimeler sanki o dokümanın karakterini yansıtacak ve teoride bütün dokümanlar birbirine benzer çıkacaktır. Etkisiz kelimelerin çıkartım işlemi her dokümanda baskın geçen bu kelimeleri temizler ve geriye sadece o dokümana has kelimeler kalır.

3.3.1.2. Terim Ağırlıkları

Terim frekanslarının belirlenmesi dokümanda terimlere ağırlık verilmesi [26] için de gereken bir hesaplamadır. Salton tarafından geliştirilen Vektör Uzay Modeli'ne göre herhangi bir dokümanı, bir uzay içerisinde vektörel olarak göstermek mümkündür.

İki boyutlu uzayda birinci boyut pozisyon, 2.boyut ise terim sıklığı olabilir. Vektör uzay modeline göre ikiden fazla veyahut sonsuz sayıda boyutlarda ekleyip dokümanı uzayda daha güçlü temsil edebilmek mümkündür.

Vektör Uzayının kullanılmasının avantajları; doğrusal cebir kullanılarak işlenebilen veri yapılarının elde edilmesi, ikilik tabandaki sayılar yerine ağırlıkların hesaba katılabilir olması, vektörler arasında tanımlı olan bütün fonksiyonların metinler arasında da tanımlanabilir olması (örneğin kosinüs benzerliği, metinler üzerinde sıralama fonksiyonlarının çalıştırılabilir olması, metnin tamamı yerine bir parçası üzerinde çalışabilir olması şeklinde sayılabilir.

Bunun yanında bazı dezavantajlarına bakacak olursak; genelde bu tip özellik vektörlerinin çıkarılması sonucunda çok yüksek miktarda özellik içeren veriyle uğraşmak gerekir. Vektör uzay modelinin en büyük dezavantajlarından birisi, metin boyutu uzadıkça kullanışsız hale gelmesidir. Çünkü uzayan metinler birbirine benzemeye başlar ve metinleri ayırt etmede önemli rol oynayan kelime farklılıkları azalır [27].

Çizelge 3.9. Temel terim ağırlık hesaplamaları

Ağırlık Sistemi	Terim Ağırlığı
İkili (Binary)	1
t	tf
w	$0,5+0,5\frac{tf}{\max(tf)}$ (3.2)
idf	$\log\frac{n}{N}$ (3.3)

3.3.1.3. Eşdizimlilik Skorları

Eşdizimlilik, bir dokümanda veya derlemde bulunan iki veya daha fazla teriminin yan yana bulunma eğilimidir. Kullanım sıklığına göre bazı sözcüklerin metinlerde birlikteliği tesadüfî değildir, dokümandaki anlamı birlikte

kuvvetlendirirler. Bu birliktelikler deyim olabileceği gibi terim veya eş anlamlı kelimeler de olabilir.

D dokümanında t_1 *ve* t_2 sözcüklerinin eşdizimlilik ölçümleri, birliktelik tablosu çıkarımından sonra değişik algoritmalar ile hesaplanabilir.

Çizelge 3.10. t1 ve t2 terimleri için birliktelik tablosu

	t_2	$\neg t_2$	
t_1	a	b	a+b
$\neg t_1$	c	d	c+d
	a+c	b+d	a+b+c+d = N

a : t_1 *ve* t_2 sözcük birimlerinin yan yana görülme sözcük sayısı : $t_1 \wedge t_2$

b : t_1 in t_2 ile yan yana görülmediği sözcük sayısı : $t_1 \wedge \neg t_2$

c : t_2 in t_1 ile yan yana görülmediği sözcük sayısı : $\neg t_1 \wedge t_2$

d : t_1 ve t_2 nin yan yana görülmediği sözcük sayısı : $\neg t_1 \wedge \neg t_2$

N : dokümandaki toplam sözcük sayısı : $\neg t_1 \wedge \neg t_2$

Eşdizimlilik ölçümlerinde, birliktelik tablosu genellikle bütün derlem üzerinden hesaplanarak çıkartılır, ama bu tez çalışmasında belirli bir veri seti, online benzerlik ölçümüne tutulacağı için, hesaplamalar sadece ilgili doküman için hesaplanmıştır.

Karşılıklı Bilgi Miktarı (Mutual Information)

Karşılıklı Bilgi Miktarı (MI) iki kelime arasındaki eşdizimliliği ölçmekte kullanılan metriklerden birisidir. Bir dokümanda veya derlemde yer alan iki kelimenin ne kadar ilişkili ve anlam olarak birbirlerini desteklediklerini belirtir. [28]

D dokümanında t_1 *ve* t_2 sözcüklerinin MI skoru için; Çizelge 3.10'daki eşdizimlilik tablosuna göre;

$$MI(t_1, t_2) = log_2\left(N \frac{a}{(a+c)(a+b)}\right) \quad (3.4)$$

Eş.(3.4)'de MI skorunun 10 üzeri çıkması anlamlı bir birliktelik olduğunu gösterir.

Log-Likelihood Oranı

MI 'ya benzer şekilde, Log-Likelihood bir dokümanda yer alan kelimeleri sıklık ve yakınlık ilişkisine göre ilişkisel açıdan ölçen bir algoritmadır. MI 'ya göre daha doğrusal sonuçlar üretir ve baskın olmayan ama spesifik kelimeleri de ön plana çıkartır [29].

$$LLR = -2\log(\lambda) \tag{3.5}$$

D dokümanında t_1 *ve* t_2 sözcüklerinin Log-Likelihood Oranı (*LLR*) için; Çizelge 3.10'daki eşdizimlilik tablosuna göre;

$$LLR\,(t_1, t_2) = 2\,(\,a.log_2(a) + b.log_2(b) + c.log_2(c) + d.log_2(d) - (a+b).log_2(a+b) - (a+c).log_2(a+c) - (b+d).log_2(b+d) - (c+d).log_2(c+d) + (a+b+c+d).log_2(a+b+c+d)$$

(3.6)

Ki-Kare Testi

Ki-Kare Testi ($\chi 2$) iki sözcük birimin bir doküman içindeki frekanslarının dağılımında gözlenen ve beklenen frekanslar arasında bir anlam olup olmadığını test eder.

$$\chi 2 = \sum \frac{(G-B)^2}{B} \tag{3.7}$$

$\chi 2$ test ölçümlerine Yate's Doğrulaması uygulandığında daha doğru bir ölçüm alınır [20].

Çizelge 3.10'daki eşdizimlilik tablosuna göre;

$$\chi 2(t_1, t_2) = \begin{cases} \text{eğer } a < 5 & \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)} \\ a \geq 5 & \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \end{cases} \quad (3.8)$$

şeklinde hesaplanabilir.

3.3.1.4. Entropi

Entropi bir sistemin düzensizliğini ifade eden terimdir. Entropi terimi ilk kez Shannon tarafından bilgi kazanımı ve veri madenciliği alanlarında kullanılmıştır. Dolayısıyla literatürde Shannon Entropisi (Shannon's Entropy) [30] olarak da geçen kavrama göre aslında "bilgi" ölçülebilen bir nesnedir. Termodinamikten geçişi ile anlatılacak olursa bir sistemin hareketi ne kadar ölçülebilir ise o kadar az bilgi içerir mantığı ile rastgeleliğin yüksek miktarda bilgi içerdiğini söyleyen ifadedir.

Bu tez çalışması kapsamında veri setlerindeki farklı dillerdeki paralel dillerdeki dokümanların anlamına bakılmadan kelime dağılımına bakılmıştır. Dolayısıyla kelimelerin rastgele dağılımından elde edilen Entropi değerlerinin de farklı dildeki çevirisi veya benzerinde de yakın bir Entropi değeriyle karşılaşılması beklenmektedir.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \cdot \log p(x_i) \quad (3.9)$$

Ayrıca Entropi kavramı ile bir sistemin içerdiği bilgi kaybedilmeden ne kadar sıkıştırılabileceği ve boyutunun indirgenebileceği ölçülmüştür.

3.3.1.5. K-Ortalama (K-Means)

Kümeleme de kullanılan bir algoritmadır. ÖV için boyut indirgeme ve güçlendirme amaçlı kullanılabilir.

K-Ortalama yönteminin uygulanabilmesi için en önemli koşul, veri setindeki değişkenlerin en azından aralık ölçekte bulunmasıdır. Çünkü küme merkezleri oluşturulurken her bir iterasyonda oluşan kümeler için değişkenlerin ortalamaları

alınır. İkinci önemli koşul ise, oluşturulacak olan küme sayısının başlangıçta biliniyor olmasıdır.

$$\arg\min_{Di} \sum_{i=1}^{k} \sum_{p_i \in D_i} \| p_i - \mu_i \|^2 \tag{3.10}$$

K-ortalamalar yönteminin kullandığı algoritma aşağıdaki gibidir:

- K adet birim başlangıç küme merkezleri olarak rasgele seçilir.
- Küme merkezi olmayan birimler belirlenen uzaklık ölçütlerine başlangıç küme merkezlerinin ait oldukları kümelere atlanır
- Yeni küme merkezleri oluşturulan k adet başlangıç kümesindeki değişkenlerin ortalamaları alınarak oluşturulur.
- Birimler en yakın oldukları oluşturulan yeni küme merkezlerine birimlerin uzaklıkları hesaplanarak kümeye atlanır.
- Bir önceki küme merkezlerine olan uzaklıklar ile yeni oluşturulan küme merkezlerine olan uzaklıklar karşılaştırılır.
- Uzaklıklar makul görülebilir oranda azalmış ise 4. adıma dönülür.
- Eğer çok büyük bir değişiklik söz konusu olmamış ise, iterasyon sona erdirilir.

3.3.2. Önerilen Öznitelik Vektörleri Çıkarım Yöntemleri

Aynı veya benzer içeriğe sahip ama farklı dillerdeki dokümanların semantiğine bakılmadan veya çift dilli (bilingual) sözlüğe başvurulmadan, kelimelerin diziliş sıralarının da birbirine benzer veya yakın olduğu düşünülmektedir. Her doküman belirli bir anlamsal ifade üzerinden bazı temel sıralara ve sayısal anlamlara sahiptir. Anlatılmak istenen ise o dile ait kelimelerle ifade edilir. Dolayısıyla metni anlatan özellikli kelimelerin ilgili dokümanda baskın olması kaçınılmazdır. Aynı dizilim veya dağılım, aynı içeriğe sahip başka dildeki bir doküman için geçerli olacaktır. Bu durumda iki dokümandaki belirli sıklıkta geçen kelimelerin yerleri saptanıp, pozisyon değeri olarak alınıp, vektörel mesafe ölçümleri ile karşılaştırıldığında yakın değerler çıkması beklenmektedir.

Doküman Benzerlikleri için 6 adet yöntem uygulanmıştır. Her yöntem ÖVÇY kodu ile birlikte isimlendirilmiştir. Bu çalışmada kullanılan özgün yöntem ÖV

oluşturulma yöntemleridir. Kullanılan yöntemler, bazı yöntemlerin bir sonraki güncellemesi şeklinde uygulanmıştır.

Dokümanlar vektörel hale getirilmeden önce belirli önişlemlerden geçirilmiştir. Her klasörde farklı türde, dilde ve anlamsal benzerlikte dokümanlar bulunmaktadır. Dokümanlar yüklendikten sonra her doküman için dil saptaması yapılıp, ilgili dokümana karakter normalizasyonu yapılmaktadır. Karakter normalizasyonu işleminden sonra dokümandaki bütün harfler küçük harflerine çevrilmektedir. Benzer anlama sahip ama farklı köklere sahip yakın kelimelerin tek bir vektörel değerde toplanması için gövdeleme işlemi uygulanmaktadır. Gövdeleme işlemi için Porter Stemming Algoritmaları kullanılmıştır [31,32]. Porter çeşitli dillerde desteklenen hazır, esnek bir gövdeleme algoritmasıdır. Gövdeleme işleminden sonra dokümandaki etkisiz kelimeler kaldırılmaktadır. Dokümandaki etkisiz kelimelerin temizlenmesi işlem süresini ve doküman boyutunu azaltmaktadır. Sonraki aşamada terim oluşturma aşamasına geçilmiştir. Dokümandaki noktalama işareti ve özel karakterler kaldırılmıştır. Burada düzenli ifadelerden (regular expressions) faydalanılmıştır. Boşluklar ayraç kabul edilerekten, iki boşluk arası bir terim olarak kabul edilmiş ve terim dizisine atanmıştır.

Dokümanların vektörel hale getirilmesinden önce 2 adet hizalama yöntemi tecih edilmiştir. 1. yöntemde dokümanların oluşturduğu Öznitelik Vektörleri ağırlıksız normal pozisyonları ile hizalanmış ve benzerlik ölçümlerine tabi tutulmuştur. 2. yöntemde ise öncelikle doküman içerisinde en güçlü kelimenin pozisyonu belirlenmiştir ve bu noktaya göre dokümanlar hizalanmıştır. 2. yöntem için pozisyon seçiminde Eşdizimlilik skorlarından yararlanılmıştır.

Daha sonra belirlenen yöntemlere göre Öznitelik Vektörleri oluşturulmuştur. ÖV'lerin boyutu her yönteme göre değişkenlik göstermektedir.

Oluşturulan ÖV'ler

- Öklid Mesafesi
- Kosinüs Benzerliği
- Mahalonobis Uzaklığı
- Pearson Çarpım-Moment Korelâsyon Katsayısı
- MeanOfMinMax Benzerlik Ölçümü

ile ölçülmüştür.

Mesafe Ölçümleri neticesinde elde edilen sonuç matriksleri global normalizasyon işlemine tabi tutulmuştur. Her mesafe ölçüm yöntemi için Ortalama Hata (MSE) değerleri hesaplanmıştır.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Yi} - \bar{\mathrm{Yi}}\right)^2 \qquad (3.11)$$

Bunun için her yöntemde elde edilen değerlerin ortalaması alınmış, simetri matristeki her değer bu ortalama değerinden çıkartılmış, karesi alınmıştır ve toplanmıştır. Elde edilen farkların karesi toplamı, çerçeve (doküman setindeki dil dosyası sayısı) değerine bölünmüştür. Sonuçlar MSE ve kendi öz değerleri ile benzerlik-yakınlık açısından sıralanarak sıfıra yakınlıkları temel alınmıştır.

Yöntemlerde, önişlemden geçen her hangi bir doküman D_i olsun;

- $D_i = \{ t_1, t_2, t_3, \ldots t_n \}$, vektörü sıralı terimlerden (t_i) oluşan bir vektördür.
- S_i, D_i içinde en çok tekrar eden L adet terimin kümesidir.
- $\ddot{O}V_i$, D_i' ye karşılık gelen öznitelik vektörüdür.
- p, pozisyonu temsil eder.
- f, frekansı temsil eder.

3.3.2.1. Öznitelik Vektörü Çıkarma Yöntemi–1 (ÖVÇY-1)

Bu yöntemde her bir D_i, dokümanın içinde en sık geçen terimlerin (S_i) pozisyonlarından (p) oluşan öznitelik vektörü $\ddot{O}V_i$ ile temsil edilmiştir. D_i ye karşılık gelen öznitelik vektörü $\ddot{O}V_i$ aşağıdaki işlem neticesinde elde edilir.

$$\ddot{O}V_i = \{ p \mid \mathrm{D_i}(p) \in S_i \} = (p_1, p_2, p_3 \ldots, p_M), \quad , \; m \leq n \qquad (3.12)$$

3.3.2.2. Öznitelik Vektörü Çıkarma Yöntemi–2 (ÖVÇY-2)

Bu yöntemde her bir D_i, dokümanın içinde en sık geçen terimlerin (S_i) pozisyonları (p) ve frekanslarından (f) oluşan öznitelik vektörü $\ddot{O}V_i$ ile temsil edilmiştir. $\ddot{O}V_i$ aşağıdaki işlem neticesinde elde edilir.

$$\ddot{O}V_i = \{ (p, \mathrm{f}) \mid \mathrm{D_i}(p) \in S_i \text{ ve } f=f_p \} = \{(p_1, f_1) \ldots (p_M, f_M)\} , \; m \leq n \qquad (3.13)$$

f_p , pozisyonu p olan terimin frekansıdır ve Eş.(3.2)'ye göre hesaplanmıştır.

3.3.2.3. Öznitelik Vektörü Çıkarma Yöntemi–3 (ÖVÇY-3)

Bu yöntemde her bir D_i, dokümanın içinde en sık geçen terimlerin (S_i) pozisyonları (p) ve frekansları (f) çarpımından ($p * f$) oluşan öznitelik vektörü $\ddot{O}V_i$ ile temsil edilmiştir. $\ddot{O}V_i$ aşağıdaki işlem neticesinde elde edilir.

$$\ddot{O}V_i = \{ (p*\mathrm{f}) \mid \mathrm{D_i}(p) \in S_i \text{ ve } f=f_p \} = \{(p_{1*} f_1) \ldots (p_{M*} f_M)\} , m \leq n \qquad (3.14)$$

f_p , pozisyonu p olan terimin frekansıdır ve Eş.(3.2)'ye göre hesaplanmıştır.

3.3.2.4. Öznitelik Vektörü Çıkarma Yöntemi–4 (ÖVÇY-4)

Bu yöntemde her bir D_i, doküman içinde en çok tekrar eden L adet terimin (S_i) pozisyonlarının (p) k-ortalama algoritması ile elde edilen k adet sendroidlerinin (c_i) oluşturduğu öznitelik vektörü $\ddot{O}V_i$ ile temsil edilmiştir. $\ddot{O}V_i$ aşağıdaki sıralı işlemler neticesinde elde edilir.

$$V_i = \{ p \mid \mathrm{D_i}(p) \in S_i \} = (p_1, p_2, p_3 \ldots , p_M) , \; m \leq n \qquad (3.15)$$

$$\ddot{O}V_i = \underset{iterasyon}{\arg\min} \sum_{i=1}^{k} \sum_{p_i \in V_i} \| p_i - \mu_i \|^2 = (c_1, \ldots , c_k) , \qquad (3.16)$$

3.3.2.5. Öznitelik Vektörü Çıkarma Yöntemi–5 (ÖVÇY-5)

Bu yöntemde, önce en çok tekrar eden L adet terimlerin (S_i) pozisyonları k-ortalama algoritması ile k adet kümelere (C_i) ayrılmıştır. Bu kümelerin merkezleri k adet sendroittir (c_i) . Sonra her bir küme içerisinde yer alan terimlerin frekanslarının ortalaması alınmıştır. Bu ortalamalar, ağırlık olarak (w_i) ile temsil edilmiştir. Öznitelik Vektörü $\ddot{O}V_i$, aşağıdaki sıralı işlemler neticesinde elde edilmiştir.

$$V_i = \{ p \mid \mathrm{D_i}(p) \in S_i \} = (p_1, p_2, p_3 \ldots , p_M) , \; m \leq n \qquad (3.17)$$

$$V_i = \underset{iterasyon}{\arg\min} \sum_{i=1}^{k} \sum_{p_i \in V_i} \| p_i - \mu_i \|^2 = (c_1, \ldots, c_k) , \quad (3.18)$$

$$w_i = \text{mean}(\{ f_p \mid t_p \in C_i \}) \quad (3.19)$$

$$\ddot{O}V_i = \{(c_1, w_1) , (c_2, w_2) \ldots (c_K, w_K)\} \quad (3.20)$$

3.3.2.6. Öznitelik Vektörü Çıkarma Yöntemi–6 (ÖVÇY-6)

Bu yöntemde her bir doküman D_i, (p, w) ikililerinden oluşan bir vektörle temsil edilmiştir. p en çok tekrar eden L adet terimin (S_i) pozisyonlarına karşılık gelmektedir. w her bir terimin k komşuluğunda (k adet solda ve k adet sağda) elde edilen eşdizimlilik skorlarının toplamı olarak düşünülmüştür.

$$V_i = \{ p_j \mid D_i (p) \in S_i \} = (p_1, p_2, p_3 \ldots , p_M) , \; m \leq n \quad (3.21)$$

$$\ddot{O}V_i = \{ (p,w(p)) \} = \{(p_1, w_1) , (p_2, w_2) \ldots (p_M, w_M)\} \quad (3.22)$$

$w(p)$ değeri Eş.(3.23)'ye göre hesaplanmıştır

$$w(p) = \sum_{j=p-k}^{p+k} \begin{cases} log_2 \frac{P(t_p, t_{p-j})}{P(t_p).P(t_{p-j})} , & MI\ Skor \\ -2\log(\lambda) , & LLR \\ \frac{(G\text{-}B)^2}{B} , & \chi 2 \end{cases} \quad (3.23)$$

Eşdizimlilik skorlarının hesaplama yöntemleri bölüm 3.3.1.3'de verilmiştir.

Bu yöntemde bir doküman için anlam ifade eden kelimelerin f dışında yeni bir ağırlık değeri ile ilişkilendirilmesi hedeflenmiştir. Her kelimenin o doküman için ifade ettiği anlamı, kelimeye ait sol ve sağ komşuları ile eşdizimliliklerine bakarak ilişkilendirme ile yeni bir ağırlık değeri hesaplanmıştır. Bu yeni ağırlık veya skor değeri için Ki-Kare Testi, Log-Likelihood değeri, MI skor eşdizimlilik hesaplamaları kullanılmıştır. Araştırmalarda eşdizimlilik skor hesaplamaları genelde derlemin tümü için hesaplanmaktadır. Fakat bu çalışmada online veriler üzerinde hesaplama yapılmıştır. Yani var olan global skor hesaplamaları, bu çalışmada lokal bir

hesaplama yöntemine dönüştürülmüş ve (ÖVÇY-6) içerisinde yeni ağırlık değerleri hesaplaması için kullanılmıştır.

Bunun için her yöntemde elde edilen değerlerin ortalaması alınmış, simetri matriksteki her değer bu ortalama değerinden çıkartılmış, karesi alınmıştır ve toplanmıştır. Elde edilen farkların karesi toplamı, çerçeve değerine bölünmüştür.

Ayrıca bu metotta *w(p)* değeri en yüksek olan kelime doküman içerisinde anlamı en iyi ifade eden kelime olarak alınabilir. Böylece Etiket Bulutunda dokümanı ifade edecek bir skor hesaplaması da oluşabilir.

3.3.3. Öznitelik Vektörlerinin Benzerlik Ölçümleri

Test kümesindeki her bir dokümanın öznitelik vektörleri elde edilerek, öznitelik vektörleri kümesi V elde edilmiştir;

$V = \{ V_1, V_2, \ldots V_N \}$,

elde edilen V_i vektörlerine global normalizasyon işlemi uygulanmıştır. Böylelikle nitelikler [0,1] aralığında nitelenmiş olur.

$$\text{Norm}(V_i) = \max[\cup_{\forall Di}(\max(Vi)], \quad (3.24)$$

$$V_i = V_i / \text{Norm}(V_i), \quad (3.25)$$

Öznitelik vektörlerinin benzerliğinin saptanması konusunda Bölüm 3,4'te yer alan yer alan mesafe ve benzerlik ölçümleri uygulanmıştır. Öklid ve Mahalonobis mesafe ölçümleri olduğundan dolayı [0,∞] aralığında sonuç üretirken, Pearson Korelasyon Katsayısı [-1,1] aralığında sınırlanmış, Kosinüs ise benzerlik ölçümü olduğundan dolayı [0,1] aralığında değer vermektedir.

Benzer dokümanları 0 değerine yaklaştırmak için Pearson Korelâsyon Katsayısı sonucu = *r* olmak üzere;

$$pearson(Di, Dj) = \frac{(2-(r+1))}{2} \quad (3.26)$$

dönüşümü uygulanmıştır.

Kosinüs Benzerliği sonucunun dönüşümü için Eş.(3.27) deki açısal benzerlik formülünden yararlanılmıştır.

$$\cos(Di, Dj) = 1-\left(\frac{2\cdot\cos^{-1}(\theta)}{\pi}\right) \quad (3.27)$$

$$sim(Di, Dj) = \begin{cases} euclidian(Vi,Vj),\ BY1\ \text{Öklid} \\ \cos ine(Vi,Vj),\ BY2 Kosinüs \\ mahalanobis(Vi,Vj),\ BY3 Mahalonobis \\ pearson(Vi,Vj),\ BY4\ Pearson \\ meanofminmax(Vi,Vj),\ BY5\ MeanOfMinMax \end{cases} \quad (3.28)$$

3.4. KULLANILAN BENZERLİK ÖLÇÜMLERİ

Önişlemlerden geçirilen, çok dilli veri setindeki dokümanların benzerliği için, belirlenmiş yöntemler ile oluşturulan Öznitelik Vektörleri ile mesafe ölçüm algoritmalarını kullanıyoruz. Oluşturduğumuz ÖV'lerin boyutuna ve uzunluğuna göre değişik yöntemler test ettik. Bu yöntemlerden elde edilen ÖV'lerin benzerlik ölçümleri Sonuçlar bölümünde incelenmiştir.

ÖV'ler sayısal değerlere sahip diziler olarak temsil edilmişti. Bu vektörlerin uzayda belirli bir alanı temsil ettiği düşünülecek olursa, bir başka vektöre yakınlığını veya benzerliğini mesafe ölçümleri tespit edebiliriz. Oluşturulan ÖV'ler ile dokümanları vektörel uzayda temsil ederek, temel mesafe ölçüm algoritmaları kullanılacaktır.

Her mesafe ölçümü metrik kabul edilemez. Bir ölçümün metrik kabul edilebilmesi için aşağıdaki 4 koşulu içermesi gerekir [33].

x ve *y* birbirine uzaklığı olan iki nokta olsun, *x* ile *y* arasındaki mesafe *d(x,y)* olmak üzere

1. iki nokta arasındaki mesafe negatif olamaz, $d(x, y) \geq 0$ (3.29)

2. iki nokta arasındaki mesafe ancak *x=y* durumunda sıfır olabilir, eğer *x=y* ise $d(x, y) = 0$ (3.30)

3. iki nokta arası mesafe simetrik olmalıdır.

$$d(x, y) = d(y, x) \tag{3.31}$$

4. *x, y, z* adlı üç nokta arasındaki mesafe ölçümünde *x* ile *z* arasındaki mesafe ölçümü x'in y'ye ve y'nin z'ye uzaklıkları toplamından büyük olamaz; (üçgen eşitsizliği)

$$d(x, z) \leq d(x, y) + d(y, z) \tag{3.32}$$

3.4.1. Öklid Uzaklığı Temelli Benzerlik Ölçüm Yöntemi

Öklid Uzaklığı (Euclidian Distance) en çok kullanılan ölçümlerden birisidir. Di ve Dj ' ye ait iki öznitelik vektörü x=Vi ve y=Vj olarak gösterilirse; iki döküman arasındaki benzerlik şu şekilde hesaplanır;

$$\mathrm{d}(\mathrm{x}, y) = \left(\sum_{k=1}^{n} \left(x_k - y_k \right)^2 \right)^{1/2} \tag{3.33}$$

İki boyutlu bir düzlemde yer alan,

X = (x_x,x_y) ve Y = (y_x,y_y) noktaları için ise Öklid uzaklığı şu şekilde hesaplanır

$$d(x,y)=\left(\sum_{k=1}^{n}\left(x_{xk}-y_{xk}\right)^2+\left(x_{yk}-y_{yk}\right)^2\right)^{1/2} \tag{3.34}$$

3.4.2. Kosinüs Uzaklığı Temelli Benzerlik Ölçüm Yöntemi

Bu tez çalışmasında Kosinüs Benzerliği (Cosine Similarity) iki doküman arasındaki yakınlığı saptamak için kullanılacaktır.

Belge kümelemede çok kullanılan vektör tabanlı bir ölçüt olan Kosinüs Benzerliği ile iki vektör arasındaki açının kosinüs değeri hesaplanarak vektörlerin benzerliği bulunur. Vektör boyutundan etkilenmemesi, kosinüs benzerliğinin güçlü bir özelliğidir. Farklı çok sayıda kelimeler içeren benzer içerikteki belgeleri kolaylıkla tespit eder.

$$\cos(\theta)=\frac{\sum_{k=1}^{n}x_k\cdot y_k}{\sqrt{\sum_{k=1}^{n}\left(x_k\right)^2}+\sqrt{\sum_{k=1}^{n}\left(y_k\right)^2}} \tag{3.35}$$

Kosinüs Benzerliği 1'e yaklaştıkça dokümanların birbirine benzer olduğuna işaret eder. Öklid Mesafe ölçümünde ise 0'a yaklaştıkça doküman benzerlikleri birbirine benzer, bu tip global bir normalizasyon işlemi için kosinüs değerini 0 ile 1 arasında bir aralığa sıkıştırmak gereklidir.

Bu dönüşüm için Açısal benzerlik formülünden faydalanılır;

$$1-\left(\frac{2\cdot\cos^{-1}(\theta)}{\pi}\right) \tag{3.36}$$

3.4.3. Mahalanobis Uzaklığı Temelli Benzerlik Ölçüm Yöntemi

Uzayda iki nokta arasındaki mesafeyi ölçmek için kullanılan en temel uzaklık ölçütü Öklid uzaklığıdır.

Öklid Uzaklığı Eş(3.33)' deki gibi hesaplanır. Bu uzaklık, bahsi geçen iki nokta dışındaki bilgileri hesaba katmaz. Mahalanobis uzaklığı iki noktanın arasındaki mesafeyi hesaplamak için veriden hesaplanan kovaryans (özdeğişinti) matrisini hesaba katar. Dolayısıyla diğer noktaların davranışını da hesaba katmış olur. Matematiksel olarak ifade edecek olursak C veriden elde edilen kovaryans matrisini ifade etmek üzere, herhangi x ve y noktaları (vektörleri) arasındaki Mahalanobis uzaklığı [41]

$$\mathrm{d}(\mathrm{x}, y) = (x - y)'\mathrm{C}^{-1}(\mathrm{x} - \mathrm{y}) \tag{3.37}$$

şeklinde hesaplanır. Aradaki kovaryans matrisinin tersi ile çarpma işlemi, elde edilen uzaklığın o doğrultuda hesaplanan değişinti (varyans, standart sapmanın karesi) ile bölünmesi gibi yorumlanabilir. Yani elde edilen uzaklığın birimi, iki nokta arasından geçen doğrunun doğrultusu boyunca olan standart sapma cinsindendir. Bu da elbette veriye bağlı bir değerdir [34].

3.4.4. Pearson Çarpım-Moment Korelâsyon Katsayısı Temelli Benzerlik Ölçüm Yöntemi

Korelâsyon katsayısı, bağımsız değişkenler arasındaki ilişkinin yönü ve büyüklüğünü belirten katsayıdır. Bu katsayı, (-1) ile (+1) arasında bir değer alır. Pozitif değerler direk yönlü doğrusal ilişkiyi; negatif değerler ise ters yönlü bir doğrusal ilişkiyi belirtir. Korelâsyon katsayısı 0 ise söz konusu değişkenler arasında doğrusal bir ilişki yoktur.

Matematik beklenti değerleri μ_X ve μ_Y, standart sapmaları σ_X ve σ_Y olan iki bağımsız değişken *x* ve *y* arasındaki Pearson'un çarpım-moment korelasyon katsayısı ($\rho_{X,Y}$), şu şekilde tanımlanır:

$$d(\mathrm{x}, \mathrm{y}) = r = \frac{\sum_{k=1}^{n}(x_k - \bar{x})\cdot(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})}\sqrt{\sum_{k=1}^{n}(y_k - \bar{y})^2}} \tag{3.38}$$

3.4.5. MeanOfMinMax Benzerlik Ölçümü

MeanOfMinMax benzerlik ölçümü, çıkartılan öznitelik vektörleri arasındaki benzerliği yüzdelik ölçüde ifade etmeye çalıştığımız test çalışmasına ait özgün bir

benzerlik ölçümüdür. Bu benzerlik ölçümünde vektöre ait noktalar arası mesafe ölçümü ile hesaplanan benzerlik ölçümünün yanı sıra, doküman boyutu ve entropi hesaplamaları da eklenmiştir. Benzer dokümanlar için doküman boyutları da benzer olmaktadır. Bu durumda doküman boyutlarının benzerlik ölçümüne katılması, benzemeyen dokümanlar için cezalandırma skoru olmaktadır. Aynı zamanda dağılımları da benzerlik gösterdiğinden dolayı entropi hesaplamaları da benzerlik ölçüm skoruna eklenmiştir.

x ve *y*, uzunlukları değişken olabilir iki vektör olsun, bu vektörleri aşağıdaki gibi ifade edecek olursak.

$x = \{ x_1, x_2, x_3, \ldots\ldots, x_m \}$,

$y = \{ y_1, y_2, y_3, \ldots\ldots, y_q \}$

$$\text{MeanOfMinMax}\ (x, y) = \frac{\sum_{k=1}^{\min(m,q)} \left(\frac{min(x_k, y_k)}{\max(x_k, y_k)}\right)}{\max(|x_k|, |y_k|)} * \frac{\min(|x|, |y|)}{max(|x|, |y|)} * \frac{H(x)}{H(y)} * 100 \quad (3.39)$$

H(x) ve H(y) değerleri Eş.(3.9)'da tanımlanmış entropi değerleridir.

4. BULGULAR VE TARTIŞMALAR

Performans ölçümleri için Nato Review dergisi veri setleri kullanılmıştır. Ölçümler öncesinde her dilden 1 adet makale seçilmiştir. Bunlardan 2 tanesi benzer içeriğe sahip çeviri metinler olarak seçilmiş diğer 5 makale ise farklı dillerde ve anlamlarda rastgele seçilmiştir. Seçilen 7 adet makale klasörlere konularak kodlanmıştır. Elde edilen bu yeni veri setleri 6 adet Öznitelik Vektörü Çıkarım yöntemi ve 2 farklı hizalama metodu ile benzerlik testlerine tabi tutulmuştur.

Birinci hizalama metodunda, Öznitelik vektörleri hizalanmadan, sayısal pozisyonları her doküman için başlangıç noktası alınaraktan test edilmiştir.

İkinci hizalama metodunda ise her doküman için dokümanı en iyi ifade eden pozisyon belirlenmiş ve her dokümanın pozisyon verileri, en güçlü olduğu noktaya göre yeniden düzenlenmiştir.

Bulgular için veri setinde 990, 987, 983, 972, 935 ve 887 numaralı veri setlerinden faydalanılmıştır. Her klasörde bulunan makalelerden oluşturulan 7 adet Öznitelik Vektörleri; Öklid, Kosinüs, Mahalonobis, Pearson ve bu tez çalışmasında geliştirilen “MeanOfMinMax ” yöntemleri ile benzerlik açısından test edilmiştir. Öklid, Kosinüs, Mahalonobis ve Pearson değerleri her karşılaştırma için sıfıra yaklaştıkça anlam ve benzerlik kazanmaktadır. “MeanOfMinMax ” ölçümü ise yüzdesel bir benzerlik yaklaşımı sunmaktadır.

Her sonuç tablosunda sol en üstte karşılaştırılan ana dokümanın terim sayısı, entropi dağılım değeri ve önişlemlerden geçtikten sonra geriye kalan terim sayısı verilmiştir. Bir sonraki satırda ise ilk kolonda karşılaştırılan dokümanın dil kodu ve sırası ile BY kodu ile benzerlik yöntemleri verilmiştir. Karşılaştırılan ana dokümandan sonra gelen satırlarda gri renk ile verilen satıra ait dokümanın ilgili dokümana anlamsal olarak benzerliğinin gösterimi için satır gri renk ile gösterilmiştir. Ayrıca her ölçüm için ilgili benzerlik yöntemindeki en iyi sonuç açık mavi renk ile vurgulanmıştır. Yöntemlerin başarısının tespiti için açık mavi renklerin gri hücrelerde olması gerekliliği hedeflenmiştir.

4.1. DOKÜMANIN BAŞLANGIÇ NOKTASINA GÖRE HİZALANMIŞ ÖZNİTELİK VEKTÖRÜ PERFORMANS ÖLÇÜMLERİ

4.1.1. ÖVÇY-1 Performans Ölçümleri

Çizelge 4.1. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 227					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	0.105	0.032	0.292	0.049	68.93
it	0.065	0.045	0.873	0.475	2.83
pt	0.348	0.033	0.849	0.441	2.83
tr	0.95	0.049	0.832	0.39	3.5
fr	0.12	0.041	0.834	0.448	5.34
es	0.135	0.059	0.826	0.448	4.17

Çizelge 4.2. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 8 terim tekrar sayısı : 250					
de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.105	0.032	0.292	0.049	68.93
it	0.058	0.041	0.887	0.477	2.61
pt	0.219	0.033	0.866	0.446	3.2
tr	0.759	0.06	0.851	0.401	3.99
fr	0.036	0.042	0.854	0.455	5.06
es	0.06	0.069	0.844	0.454	3.96

Çizelge 4.3. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 8 terim tekrar sayısı : 924					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.471	0.062	0.601	0.067	67.61
en	0.678	0.053	0.732	0.364	12.84
pt	0.331	0.039	0.702	0.353	16.89
de	0.087	0.116	0.759	0.483	0.86
es	0.98	0.043	0.57	0.014	64.86
it	0.915	0.115	0.715	0.097	54.66

Çizelge 4.4. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 979					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.471	0.062	0.601	0.067	67.61
en	0.923	0.048	0.751	0.379	8.91
pt	0.671	0.059	0.726	0.373	12.36
de	0.094	0.135	0.762	0.476	1.27
es	0.513	0.046	0.6	0.057	65.14
it	0.537	0.08	0.77	0.026	71.8

Çizelge 4.5. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 8 terim tekrar sayısı : 2398					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.709	0.028	0.692	0.020	80.03
fr	1.355	0.018	0.694	0.055	67.45
it	0.594	0.071	0.722	0.284	18.47
en	0.628	0.060	0.723	0.260	22
pt	3.290	0.042	0.691	0.044	42.88
de	0.431	0.072	0.844	0.455	1.19

Çizelge 4.6. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 8 terim tekrar sayısı : 2711					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.709	0.028	0.692	0.02	80.03
fr	2.024	0.027	0.69	0.039	71.75
it	0.939	0.044	0.707	0.26	17.96
en	1.023	0.035	0.709	0.238	21.51
pt	3.897	0.061	0.686	0.061	42.16
de	0.293	0.047	0.856	0.463	1.49

Çizelge 4.7. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 978

en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.188	0.030	0.607	0.092	65.07
it	0.478	0.061	0.734	0.362	12.9
pt	0.169	0.072	0.732	0.399	9.1
es	0.463	0.039	0.61	0.087	64.01
de	1.225	0.074	0.715	0.268	20.4
tr	0.289	0.063	0.691	0.328	18.31

Çizelge 4.8. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 1147

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.188	0.03	0.607	0.092	65.07
it	0.392	0.074	0.708	0.341	18.82
pt	0.193	0.063	0.739	0.412	6.62
es	0.37	0.036	0.561	0.01	80.86
de	1.135	0.083	0.694	0.237	25.47
tr	0.353	0.074	0.665	0.306	25.83

Çizelge 4.9. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 8 terim tekrar sayısı : 270					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.030	0.033	0.108	0.004	92.06
de	0.391	0.056	0.916	0.389	7.76
en	0.039	0.039	0.957	0.468	2.98
tr	0.483	0.09	0.868	0.326	10.71
es	0.104	0.051	0.905	0.411	9.9
pt	0.062	0.071	0.997	0.478	2

Çizelge 4.10. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741 Entropi H(it) : 11.178 En sık geçen 8 terim tekrar sayısı : 272					
it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.03	0.033	0.108	0.004	92.06
de	0.386	0.071	0.933	0.39	7.33
en	0.049	0.049	0.974	0.468	2.82
tr	0.477	0.097	0.884	0.327	10.16
es	0.101	0.058	0.922	0.411	9.58
pt	0.069	0.077	1.015	0.478	1.88

Çizelge 4.11. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543 Entropi H(it) : 9.909 En sık geçen 8 terim tekrar sayısı : 1301					
it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.408	0.031	0.631	0.014	80.73
fr	0.281	0.026	0.624	0.075	68.1
pt	1.387	0.04	0.731	0.405	10.32
tr	1.846	0.042	0.648	0.147	34.81
de	0.471	0.023	0.655	0.177	47.63
en	0.337	0.026	0.665	0.174	44.98

Çizelge 4.12. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621 Entropi H(es) : 10.557 En sık geçen 8 terim tekrar sayısı : 1353					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.408	0.031	0.631	0.014	80.73
fr	0.249	0.041	0.64	0.065	74.93
pt	1.713	0.037	0.753	0.408	9.84
tr	1.47	0.058	0.669	0.16	40.47
de	0.252	0.036	0.674	0.185	51.5
en	0.208	0.032	0.683	0.182	48.63

Çizelge 4.13. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 8 terim tekrar sayısı : 1235					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.281	0.026	0.624	0.075	68.1
es	0.249	0.041	0.64	0.065	74.93
pt	1.432	0.048	0.74	0.421	8.43
tr	1.218	0.042	0.668	0.213	32.63
de	0.189	0.028	0.67	0.229	42.38
en	0.191	0.034	0.678	0.222	40.96

4.1.2. ÖVÇY-2 Performans Ölçümleri

Çizelge 4.14. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 48					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	1.699	0.205	0.106	0.071	64.26
it	1.383	0.156	0.618	0.478	0.51
pt	1.725	0.184	0.637	0.442	0.83
tr	1.717	0.184	0.73	0.402	1.86
fr	1.65	0.186	0.571	0.456	1.39
es	1.636	0.16	0.652	0.454	0.99

Çizelge 4.15. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 45

de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	1.699	0.205	0.106	0.071	64.26
it	1.564	0.184	0.598	0.469	0.43
pt	1.756	0.196	0.612	0.446	0.78
tr	1.602	0.178	0.705	0.406	1.71
fr	1.717	0.199	0.544	0.46	1.15
es	1.575	0.161	0.622	0.455	0.87

Çizelge 4.16. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 127

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	1.286	0.081	0.161	0.072	64.06
en	1.509	0.087	0.385	0.37	5.3
pt	2.071	0.12	0.423	0.353	7.44
de	0.517	0.09	0.571	0.489	0.16
es	1.424	0.082	0.153	0.019	71.02
it	1.736	0.112	0.294	0.103	54.63

Çizelge 4.17. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 116					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	1.286	0.081	0.161	0.072	64.06
en	1.517	0.098	0.317	0.385	3.65
pt	2.007	0.13	0.366	0.373	5.22
de	0.421	0.075	0.592	0.485	0.23
es	1.318	0.076	0.227	0.063	61.39
it	1.414	0.093	0.153	0.03	74.14

Çizelge 4.18. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 8 terim tekrar sayısı : 258					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	3.059	0.122	0.116	0.031	75.87
fr	3.155	0.123	0.3	0.063	62.52
it	2.355	0.139	0.388	0.287	11.17
en	2.44	0.136	0.364	0.26	14.16
pt	4.42	0.185	0.378	0.051	50.29
de	1.356	0.148	0.663	0.446	0.3

Çizelge 4.19. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 8 terim tekrar sayısı : 265					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	3.059	0.122	0.116	0.031	75.87
fr	3.705	0.142	0.452	0.046	69.08
it	2.157	0.129	0.431	0.27	11.08
en	2.566	0.143	0.438	0.245	13.73
pt	4.904	0.21	0.453	0.072	48.93
de	1.067	0.118	0.668	0.46	0.33

Çizelge 4.20. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 8 terim tekrar sayısı : 118					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	1.972	0.13	0.23	0.099	56.33
it	2.188	0.158	0.465	0.369	5.83
pt	1.521	0.17	0.512	0.4	4.25
es	2.213	0.129	0.408	0.091	57.51
de	2.772	0.148	0.667	0.269	12.95
tr	2	0.136	0.404	0.322	9.3

Çizelge 4.21. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 130

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	1.972	0.13	0.23	0.099	56.33
it	2.41	0.152	0.452	0.354	8.92
pt	1.212	0.125	0.546	0.407	2.79
es	1.446	0.085	0.181	0.015	80.67
de	2.012	0.104	0.499	0.241	18.24
tr	1.698	0.108	0.308	0.307	14.83

Çizelge 4.22. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 8 terim tekrar sayısı : 49

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	1.771	0.203	0.024	0.031	85.61
de	1.534	0.176	0.601	0.392	3.45
en	1.487	0.153	0.565	0.47	0.52
tr	1.569	0.172	0.612	0.328	5.62
es	1.55	0.154	0.564	0.423	3.48
pt	1.796	0.191	0.611	0.486	0.29

Çizelge 4.23. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 8 terim tekrar sayısı : 49

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	1.771	0.203	0.024	0.031	85.61
de	1.657	0.192	0.595	0.389	3.29
en	1.57	0.161	0.576	0.466	0.49
tr	1.682	0.185	0.614	0.321	5.32
es	1.761	0.178	0.574	0.418	3.3
pt	1.598	0.168	0.617	0.477	0.29

Çizelge 4.24. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar sayısı : 191

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	2.204	0.106	0.254	0.023	78.46
fr	1.868	0.097	0.026	0.081	60.08
pt	2.562	0.117	0.367	0.409	4.85
tr	3.715	0.184	0.459	0.147	31.33
de	1.957	0.094	0.236	0.18	36.09
en	3.235	0.154	0.377	0.185	33.31

Çizelge 4.25. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 8 terim tekrar sayısı : 188

es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	2.204	0.106	0.254	0.023	78.46
fr	2.074	0.102	0.272	0.072	66.38
pt	2.752	0.146	0.272	0.419	4.75
tr	3.337	0.176	0.332	0.167	33.79
de	2.219	0.1	0.341	0.193	36.95
en	2.803	0.148	0.216	0.198	35.27

Çizelge 4.26. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827
Entropi H(fr) : 10.586
En sık geçen 8 terim tekrar sayısı : 171

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	1.868	0.097	0.026	0.081	60.08
es	2.074	0.102	0.272	0.072	66.38
pt	2.346	0.113	0.381	0.425	3.67
tr	3.224	0.167	0.47	0.212	25.08
de	1.796	0.092	0.254	0.23	28.18
en	2.96	0.147	0.39	0.229	26.48

4.1.3. ÖVÇY-3 Performans Ölçümleri

Çizelge 4.27. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 4069
Entropi H(en) : 11.86
En sık geçen 8 terim tekrar sayısı : 227

en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	0.16	0.22	0.36	0.15	62.35
it	0.12	0.16	0.81	0.48	2.45
pt	0.37	0.21	0.78	0.45	2.68
tr	0.82	0.17	0.79	0.41	3.29
fr	0.17	0.18	0.77	0.45	4.67
es	0.21	0.18	0.77	0.46	3.66

Çizelge 4.28. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 250

de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.16	0.22	0.36	0.15	62.35
it	0.13	0.19	0.83	0.48	2.25
pt	0.27	0.19	0.8	0.45	2.96
tr	0.68	0.17	0.81	0.42	3.73
fr	0.14	0.2	0.79	0.46	4.06
es	0.15	0.19	0.78	0.46	3.52

Çizelge 4.29. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 8 terim tekrar sayısı : 924					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.47	0.1	0.58	0.08	68.19
en	0.75	0.1	0.72	0.36	12.07
pt	0.71	0.14	0.69	0.35	15.17
de	0.08	0.17	0.67	0.48	0.87
es	0.95	0.08	0.57	0.03	65.11
it	0.79	0.16	0.68	0.12	57.35

Çizelge 4.30. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 979					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.47	0.1	0.58	0.08	68.19
en	0.89	0.11	0.73	0.38	8.96
pt	0.8	0.14	0.7	0.37	11.53
de	0.08	0.17	0.67	0.48	1.28
es	0.61	0.09	0.59	0.07	62.48
it	0.58	0.13	0.69	0.05	71.28

Çizelge 4.31. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 8 terim tekrar sayısı : 2398					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	1.71	0.15	0.67	0.07	72.51
fr	1.44	0.12	0.69	0.09	68.69
it	0.88	0.16	0.7	0.3	17.29
en	0.86	0.15	0.69	0.28	21.36
pt	2.93	0.15	0.65	0.08	44.95
de	0.39	0.13	0.81	0.46	1.17

Çizelge 4.32. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 8 terim tekrar sayısı : 2711					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	1.71	0.15	0.67	0.07	72.51
fr	1.58	0.12	0.69	0.07	76.11
it	0.98	0.14	0.69	0.28	18.27
en	0.97	0.14	0.68	0.26	22.56
pt	3.09	0.16	0.66	0.1	45.71
de	0.22	0.11	0.83	0.47	1.6

Çizelge 4.33. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 978

en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.53	0.14	0.59	0.13	56.47
it	0.65	0.17	0.7	0.37	12.59
pt	0.18	0.17	0.65	0.4	8.82
es	0.76	0.14	0.61	0.11	55.72
de	1.49	0.17	0.7	0.3	17.62
tr	0.56	0.17	0.68	0.35	16.45

Çizelge 4.34. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 1147

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.53	0.14	0.59	0.13	56.47
it	0.74	0.19	0.68	0.36	18.24
pt	0.21	0.14	0.68	0.42	6.1
es	0.53	0.1	0.58	0.03	75.37
de	1.27	0.14	0.68	0.26	23.91
tr	0.62	0.16	0.65	0.33	23.85

Çizelge 4.35. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 8 terim tekrar sayısı : 270					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.13	0.19	0.34	0.11	77.24
de	0.32	0.17	0.78	0.41	7.82
en	0.11	0.15	0.83	0.47	2.55
tr	0.41	0.17	0.77	0.36	9.95
es	0.17	0.15	0.79	0.42	8.67
pt	0.14	0.19	0.87	0.48	1.7

Çizelge 4.36. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741 Entropi H(it) : 11.178 En sık geçen 8 terim tekrar sayısı : 272					
it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.13	0.19	0.34	0.11	77.24
de	0.33	0.2	0.81	0.42	7.45
en	0.11	0.16	0.86	0.47	2.47
tr	0.43	0.2	0.8	0.37	9.9
es	0.2	0.18	0.82	0.43	8.29
pt	0.13	0.18	0.9	0.48	1.69

Çizelge 4.37. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543 Entropi H(it) : 9.909 En sık geçen 8 terim tekrar sayısı : 1301					
it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.62	0.11	0.61	0.04	78.2
fr	0.54	0.1	0.62	0.1	63.3
pt	1.46	0.11	0.72	0.41	9.79
tr	1.47	0.16	0.62	0.16	40.15
de	0.7	0.11	0.65	0.2	45.33
en	0.89	0.16	0.65	0.2	40.83

Çizelge 4.38. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621 Entropi H(es) : 10.557 En sık geçen 8 terim tekrar sayısı : 1353					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.62	0.11	0.61	0.04	78.2
fr	0.54	0.11	0.62	0.09	69.84
pt	1.56	0.12	0.73	0.42	9.89
tr	1.4	0.18	0.64	0.19	41
de	0.68	0.11	0.66	0.21	47
en	0.9	0.16	0.65	0.21	42.94

Çizelge 4.39. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 8 terim tekrar sayısı : 1235					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.54	0.11	0.62	0.1	63.3
es	0.54	0.11	0.62	0.09	69.84
pt	1.43	0.11	0.73	0.43	7.95
tr	1.06	0.17	0.65	0.24	35.88
de	0.49	0.1	0.67	0.24	39.61
en	0.79	0.16	0.66	0.23	35.41

4.1.4. ÖVÇY-4 Performans Ölçütleri

Çizelge 4.40. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=12

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 227					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	0.06	0.05	0.16	0.01	81.96
it	1.24	0.08	0.63	0.1	13.69
pt	1.53	0.07	1.11	0	9.72
tr	1.59	0.05	0.7	0.07	9.11
fr	1.55	0.09	0.97	0.01	10.93
es	1.06	0.04	0.65	0.08	12.53

Çizelge 4.41. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=12

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 250

de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.06	0.05	0.16	0.01	81.96
it	1.19	0.06	0.59	0.1	16.76
pt	1.48	0.09	1.23	0.01	12.04
tr	1.54	0.06	0.71	0.07	11.21
fr	1.5	0.11	1.03	0.02	12.39
es	1.01	0.06	0.66	0.08	15.46

Çizelge 4.42. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 924

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.22	0.1	0.73	0.02	74.64
en	0.65	0.07	0.63	0.01	51.29
pt	1.31	0.08	0.62	0.09	29.55
de	0.24	0.03	0.65	0.36	6.47
es	0.5	0.09	0.61	0.02	58.88
it	0.19	0.05	0.64	0	79.96

Çizelge 4.43. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 979					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.22	0.1	0.73	0.02	74.64
en	0.5	0.07	0.69	0.01	62.5
pt	1.13	0.03	0.66	0.05	36
de	0.41	0.07	0.6	0.26	5.14
es	0.33	0.04	0.85	0	71.11
it	0.08	0.05	0.72	0	85.12

Çizelge 4.44. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 8 terim tekrar sayısı : 2398					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.41	0.13	0.72	0.09	55.37
fr	0.48	0.02	0.53	0.12	54.78
it	0.67	0.14	0.69	0.11	47.4
en	0.65	0.06	0.58	0.09	41.11
pt	0.66	0.14	0.72	0.1	56.15
de	1.03	0.05	0.61	0.12	11.34

Çizelge 4.45. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151
Entropi H(tr) : 11.986
En sık geçen 8 terim tekrar sayısı : 2711

tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.41	0.13	0.72	0.09	55.37
fr	0.79	0.13	1.08	0.01	51.22
it	1	0.02	0.7	0	40.8
en	1.01	0.07	1.16	0	36.89
pt	0.98	0.02	0.67	0	48.07
de	1.35	0.08	0.7	0.15	6.77

Çizelge 4.46. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 978

en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.18	0.11	1.01	0.02	69.38
it	1.2	0.07	0.91	0.01	39.91
pt	0.76	0.07	0.53	0.02	17.82
es	0.3	0.2	0.96	0.06	62.16
de	0.98	0.11	1.24	0.01	59.85
tr	0.55	0.2	0.96	0.08	45.55

Çizelge 4.47. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 1147

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.18	0.11	1.01	0.02	69.38
it	1.15	0.05	1.35	0	41.47
pt	0.8	0.08	0.66	0.01	17.72
es	0.19	0.1	0.73	0.01	73.74
de	0.92	0.04	0.68	0	54.51
tr	0.47	0.1	0.65	0.07	58.06

Çizelge 4.48. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 8 terim tekrar sayısı : 270

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.08	0.1	0.61	0.28	31.48
de	1.03	0.07	0.65	0.01	18.8
en	1.16	0.14	1.38	0.01	28.13
tr	0.8	0.04	0.42	0.01	23.55
es	0.65	0.06	0.55	0	32.3
pt	1.68	0.07	1.24	0	18.57

Çizelge 4.49. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 8 terim tekrar sayısı : 272

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.08	0.1	0.61	0.28	31.48
de	0.59	0.08	0.81	0.19	11.31
en	0.35	0.06	0.77	0.33	20.24
tr	0.41	0.08	0.8	0.23	14.86
es	0.28	0.07	0.76	0.24	20.67
pt	0.77	0.08	0.78	0.27	13.25

Çizelge 4.50. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim Sayısı: 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar Sayısı: 1301

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.2	0.11	1.03	0.01	65.34
fr	0.12	0.05	0.88	0	79.56
pt	1.34	0.08	1.37	0.01	32.45
tr	0.94	0.06	1.38	0.01	39
de	0.47	0.05	0.65	0.01	63.3
en	0.49	0.02	0.66	0.05	47.8

Çizelge 4.51. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621 Entropi H(es) : 10.557 En sık geçen 8 terim tekrar sayısı : 1353					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.2	0.11	1.03	0.01	65.34
fr	0.11	0.07	0.84	0	78.75
pt	1.5	0.18	1.27	0.03	25.83
tr	1.09	0.16	1.33	0.02	30.68
de	0.6	0.1	1.36	0	48.37
en	0.65	0.11	0.68	0.07	34.2

Çizelge 4.52. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 8 terim tekrar sayısı : 1235					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.12	0.05	0.88	0	79.56
es	0.11	0.07	0.84	0	78.75
pt	1.44	0.13	1.24	0.02	30
tr	1.04	0.1	1.22	0.01	35.87
de	0.56	0.07	0.79	0.01	57.68
en	0.59	0.06	0.65	0.06	42.92

4.1.5. ÖVÇY-5 Performans Ölçütleri

Çizelge 4.53. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=10

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 12 terim tekrar sayısı : 227					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	0.28	0.06	0.21	0.01	82.7
it	1.44	0.27	1.2	0.03	50.03
pt	1.54	0.3	1.12	0.01	44.48
tr	1.82	0.33	1.12	0.01	47.33
fr	1.15	0.23	1.03	0.01	52.33
es	1.24	0.22	1.41	0.01	46.29

Çizelge 4.54. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=10

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 12 terim tekrar sayısı : 250					
de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.28	0.06	0.21	0.01	82.7
it	1.4	0.27	0.84	0.01	55.14
pt	1.53	0.3	1.04	0.01	46.46
tr	1.79	0.33	0.92	0	51.27
fr	1.14	0.23	0.85	0.01	50.4
es	1.2	0.22	1.04	0.01	50.88

Çizelge 4.55. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)
k=10

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 12 terim tekrar sayısı : 924

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.14	0.02	0.43	0	89.57
en	0.89	0.18	0.83	0	64.1
pt	1.08	0.22	1.06	0.01	63.37
de	0.54	0.15	0.72	0.24	33.95
es	0.64	0.12	0.48	0.03	66.78
it	0.28	0.06	0.93	0.02	84.3

Çizelge 4.56. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945
Entropi H(tr) : 9.653
En sık geçen 12 terim tekrar sayısı : 979

tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.14	0.02	0.43	0	89.57
en	0.9	0.17	0.49	0	62.32
pt	1.08	0.22	0.63	0	62.79
de	0.56	0.16	0.78	0.23	34.94
es	0.66	0.12	0.83	0.02	63.44
it	0.24	0.05	0.31	0.01	82.68

Çizelge 4.57. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181
Entropi H(es) : 12.795
En sık geçen 12 terim tekrar sayısı : 2398

es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.2	0.04	0.58	0	86.98
fr	0.35	0.06	1.74	0	85.63
it	0.82	0.15	0.37	0.01	59.6
en	0.68	0.13	0.27	0.02	66.32
pt	0.69	0.13	0.3	0.01	72.58
de	1.57	0.33	0.7	0	41.43

Çizelge 4.58. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151
Entropi H(tr) : 11.986
En sık geçen 12 terim tekrar sayısı : 2711

tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.2	0.04	0.58	0	86.98
fr	0.5	0.09	2.23	0	87.02
it	0.93	0.18	0.55	0	60.92
en	0.79	0.16	0.56	0.02	68.13
pt	0.81	0.16	0.77	0	74.56
de	1.69	0.36	0.81	0.01	43.37

Çizelge 4.59. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 12 terim tekrar sayısı : 978

en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.4	0.06	0.74	0.01	71.87
it	1.18	0.24	0.5	0.01	58.46
pt	0.41	0.1	0.67	0.12	46
es	0.66	0.1	1.68	0.03	64.57
de	1.48	0.23	1.79	0.04	54.27
tr	1.13	0.21	0.82	0.02	48.98

Çizelge 4.60. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 12 terim tekrar sayısı : 1147

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.4	0.06	0.74	0.01	71.87
it	0.96	0.2	1.1	0	70.94
pt	0.67	0.14	0.9	0.08	36.18
es	0.37	0.07	0.41	0.02	75.02
de	1.18	0.18	0.88	0.02	59.99
tr	0.86	0.17	0.53	0.01	63.15

Çizelge 4.61. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 12 terim tekrar sayısı : 270

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.23	0.05	0.17	0	87.09
de	0.75	0.18	0.66	0.08	49.91
en	1.4	0.28	1.19	0.08	46.05
tr	0.55	0.12	0.73	0.11	49.93
es	0.53	0.1	0.63	0.11	58.33
pt	1.46	0.29	0.84	0.08	48.92

Çizelge 4.62. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 12 terim tekrar sayısı : 272

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.23	0.05	0.17	0	87.09
de	0.74	0.18	0.65	0.08	46.45
en	1.41	0.28	1.62	0.07	42.78
tr	0.55	0.11	1	0.11	46.27
es	0.56	0.09	0.93	0.12	55.18
pt	1.47	0.29	1.28	0.08	44.77

Çizelge 4.63. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 12 terim tekrar sayısı : 1301

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.29	0.05	0.94	0	84.32
fr	0.15	0.03	0.11	0	83.78
pt	1.17	0.22	0.99	0.01	56.75
tr	1	0.21	1.73	0.01	52.93
de	0.52	0.1	0.41	0.01	72.88
en	0.79	0.16	1.5	0.01	59.19

Çizelge 4.64. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 12 terim tekrar sayısı : 1353

es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.29	0.05	0.94	0	84.32
fr	0.35	0.06	1.1	0	81.75
pt	1.03	0.18	0.5	0	64.13
tr	0.8	0.17	0.81	0	62.38
de	0.49	0.06	1.09	0	79.14
en	0.59	0.13	0.69	0.01	68.94

Çizelge 4.65. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 12 terim tekrar sayısı : 1235					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.15	0.03	0.11	0	83.78
es	0.35	0.06	1.1	0	81.75
pt	1.2	0.22	1.28	0.01	57
tr	1.05	0.22	1.96	0.01	53.39
de	0.56	0.1	0.4	0	73.66
en	0.84	0.17	1.67	0.01	59.36

4.1.6. ÖVÇY-6 Performans Ölçütleri

Çizelge 4.66. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 25 terim tekrar sayısı : 96					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
de	0.105	0.501	0.073	0.225	65.73
it	0.271	0.607	0.597	0.486	0.01
pt	0.312	0.639	0.567	0.512	0.03
tr	0.24	0.66	0.652	0.427	0.03
fr	0.214	0.586	0.571	0.528	0.05
es	0.257	0.521	0.579	0.463	0.04

Çizelge 4.67. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 25 terim tekrar sayısı : 91					
de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.105	0.501	0.073	0.225	65.73
it	0.28	0.689	0.587	0.48	0.01
pt	0.329	0.729	0.552	0.495	0.02
tr	0.246	0.707	0.628	0.445	0.03
fr	0.221	0.641	0.557	0.46	0.04
es	0.27	0.622	0.566	0.469	0.03

Çizelge 4.68. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 25 terim tekrar sayısı : 567					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.284	0.56	0.088	0.48	44.4
en	0.537	0.54	0.35	0.479	3.29
pt	0.538	0.536	0.422	0.49	2.78
de	0.084	0.708	0.588	0.514	0
es	0.298	0.537	0.216	0.421	19.22
it	0.275	0.528	0.117	0.428	44.78

Çizelge 4.69. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 25 terim tekrar sayısı : 526					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.284	0.56	0.088	0.48	44.4
en	0.554	0.576	0.377	0.51	2.44
pt	0.534	0.513	0.439	0.486	2.05
de	0.074	0.625	0.584	0.536	0
es	0.268	0.528	0.175	0.474	15.47
it	0.256	0.549	0.056	0.436	37.63

Çizelge 4.70. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 25 terim tekrar sayısı : 1765					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.525	0.504	0.076	0.437	50.76
fr	0.557	0.525	0.081	0.521	34.02
it	0.481	0.594	0.346	0.472	4.81
en	0.452	0.555	0.32	0.476	4.22
pt	0.483	0.548	0.356	0.404	7.24
de	0.278	0.683	0.605	0.454	0.0

Çizelge 4.71. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 25 terim tekrar sayısı : 1545					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.525	0.504	0.076	0.437	50.76
fr	0.542	0.525	0.153	0.501	31.02
it	0.438	0.574	0.392	0.437	7.74
en	0.406	0.537	0.365	0.453	6.81
pt	0.489	0.582	0.422	0.427	11.22
de	0.23	0.657	0.63	0.431	0.01

Çizelge 4.72. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 25 terim tekrar sayısı : 568					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.259	0.527	0.117	0.393	66.91
it	0.481	0.545	0.438	0.461	1.23
pt	0.197	0.627	0.525	0.476	0.2
es	0.273	0.526	0.191	0.404	12.4
de	0.412	0.565	0.447	0.495	4.44
tr	0.42	0.551	0.378	0.416	3.13

Çizelge 4.73. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 25 terim tekrar sayısı : 642

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.259	0.527	0.117	0.393	66.91
it	0.491	0.534	0.384	0.457	2.18
pt	0.188	0.639	0.548	0.446	0.11
es	0.295	0.556	0.078	0.491	18.8
de	0.402	0.514	0.371	0.53	6.58
tr	0.414	0.517	0.31	0.456	5.48

Çizelge 4.74. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 25 terim tekrar sayısı : 109

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.11	0.441	0.045	0.221	77.67
de	0.166	0.63	0.617	0.483	0.2
en	0.292	0.592	0.581	0.494	0.03
tr	0.175	0.662	0.577	0.419	0.46
es	0.212	0.617	0.539	0.43	0.2
pt	0.37	0.595	0.561	0.528	0.01

Çizelge 4.75. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 25 terim tekrar sayısı : 109

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.11	0.441	0.045	0.221	77.67
de	0.19	0.695	0.616	0.435	0.19
en	0.298	0.634	0.566	0.519	0.02
tr	0.185	0.665	0.573	0.379	0.46
es	0.211	0.609	0.534	0.457	0.19
pt	0.377	0.649	0.556	0.508	0.01

Çizelge 4.76. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 25 terim tekrar sayısı : 973

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.326	0.571	0.035	0.47	77.07
fr	0.284	0.553	0.044	0.472	70.86
pt	0.586	0.547	0.427	0.519	2.99
tr	0.37	0.546	0.329	0.464	10.34
de	0.335	0.535	0.235	0.525	19.24
en	0.343	0.517	0.242	0.483	16.5

Çizelge 4.77. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 25 terim tekrar sayısı : 1022

es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.326	0.571	0.035	0.47	77.07
fr	0.284	0.526	0.067	0.524	68.93
pt	0.583	0.536	0.401	0.499	3.76
tr	0.395	0.571	0.311	0.471	12.45
de	0.361	0.557	0.21	0.468	23.37
en	0.358	0.531	0.209	0.517	20.44

Çizelge 4.78. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827
Entropi H(fr) : 10.586
En sık geçen 25 terim tekrar sayısı : 868

fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.284	0.553	0.044	0.472	70.86
es	0.284	0.526	0.067	0.524	68.93
pt	0.588	0.563	0.443	0.521	2.1
tr	0.357	0.557	0.332	0.488	7.31
de	0.322	0.538	0.247	0.5	14.89
en	0.329	0.515	0.262	0.51	12.13

4.1.7. MeanOfMinMax Benzerlik Ölçümü Sonuçları

Çizelge 4.79. 990 numaralı doküman setindeki İngilizce dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İngilizce (en), Almanca (de)

Terim sayısı : 4069
Entropi H(en) : 11.86
En sık geçen 8 terim tekrar sayısı : 227

en	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
de	68.93	64.26	62.35	81.96	82.7	65.73
it	2.83	0.51	2.45	13.69	50.03	0.01
pt	2.83	0.83	2.68	9.72	44.48	0.03
tr	3.5	1.86	3.29	9.11	47.33	0.03
fr	5.34	1.39	4.67	10.93	52.33	0.05
es	4.17	0.99	3.66	12.53	46.29	0.04

Çizelge 4.80. 987 numaralı doküman setindeki Fransızca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 924

fr	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
tr	67.61	64.06	68.19	74.64	89.57	44.4
en	12.84	5.3	12.07	62.5	64.1	3.29
pt	16.89	7.44	15.17	36	63.37	2.78
de	0.86	0.16	0.87	5.14	33.95	0
es	64.86	71.02	65.11	71.11	66.78	19.22
it	54.66	54.63	57.35	85.12	84.3	44.78

Çizelge 4.81. 983 numaralı doküman setindeki İspanyolca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İspanyolca (es), Türkçe (tr)

Terim sayısı : 11181
Entropi H(es) : 12.795
En sık geçen 25 terim tekrar sayısı : 1765

es	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
tr	80.03	75.87	72.51	55.37	86.98	50.76
fr	67.45	62.52	68.69	54.78	85.63	34.02
it	18.47	11.17	17.29	47.4	59.6	4.81
en	22	14.16	21.36	41.11	66.32	4.22
pt	42.88	50.29	44.95	56.15	72.58	7.24
de	1.19	0.3	1.17	11.34	41.43	0.0

Çizelge 4.82. 972 numaralı doküman setindeki İngilizce dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İngilizce (en), Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 25 terim tekrar sayısı : 568

en	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
fr	65.07	56.33	56.47	69.38	71.87	66.91
it	12.9	5.83	12.59	39.91	58.46	1.23
pt	9.1	4.25	8.82	17.82	46	0.2
es	64.01	57.51	55.72	62.16	64.57	12.4
de	20.4	12.95	17.62	59.85	54.27	4.44
tr	18.31	9.3	16.45	45.55	48.98	3.13

Çizelge 4.83. 935 numaralı doküman setindeki Fransızca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 8 terim tekrar sayısı : 270

fr	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
it	92.06	85.61	77.24	31.48	87.09	77.67
de	7.76	3.45	7.82	18.8	49.91	0.2
en	2.98	0.52	2.55	28.13	46.05	0.03
tr	10.71	5.62	9.95	23.55	49.93	0.46
es	9.9	3.48	8.67	32.3	58.33	0.2
pt	2	0.29	1.7	18.57	48.92	0.01

Çizelge 4.84. 887 numaralı doküman setindeki İtalyanca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), İtalyanca (it), İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar sayısı : 1301

it	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
es	80.73	78.46	78.2	65.34	84.32	77.07
fr	68.1	60.08	63.3	79.56	83.78	70.86
pt	10.32	4.85	9.79	32.45	56.75	2.99
tr	34.81	31.33	40.15	39	52.93	10.34
de	47.63	36.09	45.33	63.3	72.88	19.24
en	44.98	33.31	40.83	47.8	59.19	16.5

4.2. DOKÜMANIN MI SKOR İLE BULUNMUŞ EN GÜÇLÜ KELİMESİNİN POZİSYONUNA GÖRE HİZALANMIŞ ÖZNİTELİK VEKTÖRÜ PERFORMANS ÖLÇÜMLERİ

4.2.1. ÖVÇY-1 Performans Ölçümleri

Çizelge 4.85. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 227					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.004	0.109	0	0.049	55.8
it	0.008	0.267	0.021	0.475	1.58
pt	0.265	0.851	0.015	0.441	0.24
tr	0.155	0.724	0.002	0.39	0.91
fr	0.165	0.837	0.006	0.448	0.5
es	0.031	0.693	0.017	0.448	1.33

Çizelge 4.86. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 8 terim tekrar sayısı : 250					
de	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.004	0.109	0	0.049	55.8
it	0.009	0.327	0.024	0.477	1.29
pt	0.259	0.864	0.017	0.446	0.24
tr	0.154	0.751	0.002	0.401	0.89
fr	0.162	0.854	0.006	0.455	0.47
es	0.027	0.68	0.02	0.454	1.33

Çizelge 4.87. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 8 terim tekrar sayısı : 924					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.139	0.713	0.041	0.067	24.91
en	0.413	0.473	0.092	0.364	3.41
pt	0.201	0.964	0.139	0.353	5.6
de	0.088	1.275	0.002	0.483	0.04
es	0.246	0.953	0.087	0.014	23.08
it	0.247	1.135	0.075	0.097	25.24

Çizelge 4.88. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 979					
tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.139	0.713	0.041	0.067	24.91
en	0.527	1.108	0.112	0.379	2.64
pt	0.063	0.332	0.098	0.373	8.62
de	0.045	1.227	0.001	0.476	0.11
es	0.096	0.301	0.036	0.057	31.23
it	0.112	0.427	0.038	0.026	36.05

Çizelge 4.89. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181
Entropi H(es) : 12.795
En sık geçen 8 terim tekrar sayısı : 2398

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.227	0.27	0.071	0.02	39.79
fr	0.13	0.238	0.044	0.055	42.74
it	0.397	1.338	0.089	0.284	9.94
en	0.191	0.459	0.025	0.26	11.56
pt	0.201	0.496	0.037	0.044	26.27
de	0.216	1.194	0.009	0.455	0.15

Çizelge 4.90. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151
Entropi H(tr) : 11.986
En sık geçen 8 terim tekrar sayısı : 2711

tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.227	0.27	0.071	0.02	39.79
fr	0.114	0.047	0.041	0.039	58.81
it	0.572	1.404	0.136	0.26	10.33
en	0.371	0.538	0.063	0.238	7.59
pt	0.416	0.736	0.082	0.061	20.75
de	0.332	1.221	0.017	0.463	0.1

Çizelge 4.91. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 978

en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.135	0.841	0.041	0.092	19.85
it	0.152	0.902	0.095	0.362	4.63
pt	0.1	1.383	0.007	0.399	2.89
es	0.067	0.442	0.025	0.087	34.43
de	0.284	0.426	0.04	0.268	7.52
tr	0.091	0.629	0.066	0.328	6.84

Çizelge 4.92. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 1147

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.135	0.841	0.041	0.092	19.85
it	0.024	0.097	0.087	0.341	16.82
pt	0.018	0.624	0.008	0.412	3.48
es	0.074	0.33	0.026	0.01	42.63
de	0.438	1.252	0.111	0.237	12.43
tr	0.057	0.209	0.057	0.306	14.75

Çizelge 4.93. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 8 terim tekrar sayısı : 270					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.007	0.251	0.001	0.004	47.86
de	0.179	1.199	0.012	0.389	1.2
en	0.064	1.125	0.011	0.468	1.01
tr	0.141	1.145	0.011	0.326	2.44
es	0.006	0.107	0.01	0.411	7.58
pt	0.368	1.28	0.02	0.478	0.1

Çizelge 4.94. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741 Entropi H(it) : 11.178 En sık geçen 8 terim tekrar sayısı : 272					
it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.007	0.251	0.001	0.004	47.86
de	0.172	0.962	0.01	0.39	1.01
en	0.058	0.888	0.01	0.468	0.9
tr	0.134	0.909	0.009	0.327	1.98
es	0.009	0.206	0.01	0.411	6.07
pt	0.361	1.042	0.016	0.478	0.08

Çizelge 4.95. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar sayısı : 1301

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.22	0.89	0.047	0.014	25.15
fr	0.149	0.82	0.028	0.075	19.46
pt	0.082	0.468	0.061	0.405	6.18
tr	0.156	0.605	0.051	0.147	16.42
de	0.198	0.457	0.028	0.177	19.54
en	0.143	0.409	0.019	0.174	19.51

Çizelge 4.96. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 8 terim tekrar sayısı : 1353

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.22	0.89	0.047	0.014	25.15
fr	0.062	0.173	0.019	0.065	43.98
pt	0.159	0.423	0.08	0.408	3.84
tr	0.091	0.289	0.047	0.16	29.29
de	0.416	1.335	0.087	0.185	22.38
en	0.362	1.288	0.074	0.182	19.23

Çizelge 4.97. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-1 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 8 terim tekrar sayısı : 1235					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.149	0.82	0.028	0.075	19.46
es	0.062	0.173	0.019	0.065	43.98
pt	0.09	0.338	0.057	0.421	4.56
tr	0.04	0.172	0.029	0.213	27.68
de	0.339	1.205	0.059	0.229	16.61
en	0.284	1.162	0.049	0.222	15.51

4.2.2. ÖVÇY-2 Performans Ölçümleri

Çizelge 4.98. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 48					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.061	0.205	0.008	0.071	58.13
it	0.051	0.159	0.588	0.478	0.4
pt	0.341	0.647	0.573	0.442	0.52
tr	0.203	0.489	0.256	0.402	1.44
fr	0.218	0.525	0.251	0.456	0.75
es	0.071	0.193	0.74	0.454	0.65

Çizelge 4.99. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 45

de	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.061	0.205	0.008	0.071	58.13
it	0.058	0.187	0.581	0.469	0.32
pt	0.334	0.652	0.566	0.446	0.44
tr	0.202	0.496	0.237	0.406	1.24
fr	0.215	0.531	0.243	0.46	0.58
es	0.067	0.189	0.727	0.455	0.57

Çizelge 4.100. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 127

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.162	0.283	0.455	0.072	44.56
en	0.47	0.483	0.515	0.37	3.39
pt	0.239	0.403	0.786	0.353	4.9
de	0.101	0.43	0.377	0.489	0.1
es	0.28	0.441	0.678	0.019	50.46
it	0.285	0.5	0.848	0.103	41.53

Çizelge 4.101. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945
Entropi H(tr) : 9.653
En sık geçen 8 terim tekrar sayısı : 116

tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.162	0.283	0.455	0.072	44.56
en	0.598	0.73	0.754	0.385	2.49
pt	0.098	0.177	0.535	0.373	4.45
de	0.053	0.257	0.116	0.485	0.15
es	0.116	0.181	0.37	0.063	45.65
it	0.135	0.237	0.394	0.03	56.73

Çizelge 4.102. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181
Entropi H(es) : 12.795
En sık geçen 8 terim tekrar sayısı : 258

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.342	0.276	0.324	0.031	56.28
fr	0.209	0.191	0.324	0.063	51.1
it	0.58	0.865	0.597	0.287	8.74
en	0.287	0.378	0.164	0.26	10.97
pt	0.306	0.358	0.309	0.051	42.46
de	0.315	0.69	0.265	0.446	0.21

Çizelge 4.103. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151
Entropi H(tr) : 11.986
En sık geçen 8 terim tekrar sayısı : 265

tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.342	0.276	0.324	0.031	56.28
fr	0.192	0.149	0.481	0.046	62.94
it	0.831	1.019	0.894	0.27	8.96
en	0.543	0.537	0.508	0.245	9.58
pt	0.61	0.588	0.66	0.072	39.11
de	0.483	0.817	0.562	0.46	0.21

Çizelge 4.104. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 118

en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.187	0.344	0.646	0.099	35.81
it	0.209	0.401	0.714	0.369	4.02
pt	0.139	0.406	0.398	0.4	2.96
es	0.115	0.202	0.5	0.091	43.98
de	0.375	0.423	0.57	0.269	9.25
tr	0.137	0.259	0.585	0.322	6.3

Çizelge 4.105. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 130

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.187	0.344	0.646	0.099	35.81
it	0.09	0.146	0.444	0.354	8.43
pt	0.049	0.143	0.381	0.407	2.2
es	0.107	0.176	0.281	0.015	61.71
de	0.565	0.744	0.759	0.241	14.11
tr	0.094	0.151	0.202	0.307	11.63

Çizelge 4.106. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 8 terim tekrar sayısı : 49

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.066	0.205	0.155	0.031	63.52
de	0.269	0.624	0.451	0.392	2.22
en	0.109	0.302	0.372	0.47	0.34
tr	0.215	0.527	0.52	0.328	3.76
es	0.057	0.155	0.537	0.423	3.07
pt	0.546	0.798	0.415	0.486	0.14

Çizelge 4.107. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 8 terim tekrar sayısı : 49

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.066	0.205	0.155	0.031	63.52
de	0.261	0.598	0.418	0.389	2.1
en	0.102	0.281	0.408	0.466	0.31
tr	0.206	0.502	0.482	0.321	3.48
es	0.066	0.18	0.584	0.418	2.67
pt	0.536	0.765	0.411	0.477	0.15

Çizelge 4.108. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar sayısı : 191

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.3	0.482	0.772	0.023	51.1
fr	0.206	0.358	0.464	0.081	38.31
pt	0.125	0.207	0.585	0.409	3.98
tr	0.227	0.372	0.56	0.147	23.85
de	0.27	0.351	0.323	0.18	25.45
en	0.212	0.33	0.385	0.185	23.43

Çizelge 4.109. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 8 terim tekrar sayısı : 188

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.3	0.482	0.772	0.023	51.1
fr	0.101	0.167	0.384	0.072	52.31
pt	0.221	0.345	0.149	0.419	3.51
tr	0.148	0.227	0.142	0.167	29.31
de	0.56	0.816	0.918	0.193	26.09
en	0.49	0.782	0.803	0.198	24.04

Çizelge 4.110. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-2 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827
Entropi H(fr) : 10.586
En sık geçen 8 terim tekrar sayısı : 171

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.206	0.358	0.464	0.081	38.31
es	0.101	0.167	0.384	0.072	52.31
pt	0.131	0.217	0.36	0.425	2.94
tr	0.102	0.169	0.36	0.212	23.28
de	0.455	0.681	0.669	0.23	19.44
en	0.389	0.65	0.674	0.229	17.64

4.2.3. ÖVÇY-3 Performans Ölçümleri

Çizelge 4.111. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 4069
Entropi H(en) : 11.86
En sık geçen 8 terim tekrar sayısı : 227

en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.08	0.22	0.02	0.07	51.43
it	0.13	0.32	0.5	0.48	1.56
pt	3.85	0.85	0.48	0.44	0.22
tr	2.21	0.74	0.05	0.4	0.87
fr	2.32	0.86	0.2	0.45	0.49
es	0.47	0.69	0.57	0.45	1.27

Çizelge 4.112. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 250

de	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.08	0.22	0.02	0.07	51.43
it	0.14	0.35	0.5	0.48	1.26
pt	3.76	0.87	0.48	0.45	0.23
tr	2.19	0.76	0.05	0.41	0.84
fr	2.29	0.87	0.2	0.46	0.45
es	0.42	0.69	0.57	0.45	1.24

Çizelge 4.113. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 924

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	1.92	0.71	0.9	0.07	24.51
en	5.53	0.49	0.61	0.38	3.58
pt	2.7	0.96	0.78	0.35	5.52
de	1.21	1.27	0.28	0.48	0.04
es	3.43	0.95	2.39	0.02	22.65
it	3.32	1.13	1.32	0.11	24.68

Çizelge 4.114. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945
Entropi H(tr) : 9.653
En sık geçen 8 terim tekrar sayısı : 979

tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	1.92	0.71	0.9	0.07	24.51
en	7.07	1.11	0.71	0.39	2.71
pt	0.88	0.35	0.53	0.37	8.42
de	0.61	1.23	0.16	0.48	0.11
es	1.35	0.31	0.62	0.07	31.1
it	1.47	0.43	1.06	0.04	37.15

Çizelge 4.115. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181
Entropi H(es) : 12.795
En sık geçen 8 terim tekrar sayısı : 2398

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	6.28	0.3	0.79	0.04	42.21
fr	4.66	0.26	0.51	0.06	39.17
it	11.21	1.32	0.77	0.29	9.98
en	5.49	0.48	0.22	0.26	11.84
pt	5.87	0.51	0.66	0.06	26.81
de	6.25	1.18	0.3	0.45	0.15

Çizelge 4.116. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151
Entropi H(tr) : 11.986
En sık geçen 8 terim tekrar sayısı : 2711

tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	6.28	0.3	0.79	0.04	42.21
fr	2.54	0.11	0.5	0.05	66.16
it	15.64	1.41	1.06	0.29	10.13
en	9.97	0.56	0.53	0.25	8.72
pt	11.31	0.75	1.11	0.08	21.02
de	8.68	1.21	0.54	0.47	0.1

Çizelge 4.117. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241
Entropi H(en) : 9.611
En sık geçen 8 terim tekrar sayısı : 978

en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	1.81	0.84	0.98	0.11	19.41
it	1.96	0.91	0.68	0.37	4.38
pt	1.29	1.37	0.46	0.41	2.85
es	0.93	0.46	0.59	0.09	36.39
de	4.32	0.44	0.38	0.27	6.62
tr	1.2	0.63	0.56	0.34	7.17

Çizelge 4.118. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621
Entropi H(fr) : 11.211
En sık geçen 8 terim tekrar sayısı : 1147

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	1.81	0.84	0.98	0.11	19.41
it	0.66	0.21	0.52	0.36	15.88
pt	0.25	0.63	0.42	0.41	3.54
es	1.06	0.35	0.97	0.02	43.87
de	6.3	1.24	0.88	0.24	11.85
tr	0.92	0.26	0.41	0.32	14.12

Çizelge 4.119. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332
Entropi H(fr) : 10.859
En sık geçen 8 terim tekrar sayısı : 270

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.15	0.32	0.21	0.03	45.72
de	2.97	1.18	0.45	0.39	1.28
en	1.17	1.11	0.31	0.47	0.93
tr	2.49	1.12	0.58	0.32	2.45
es	0.14	0.16	0.44	0.41	6.82
pt	6.69	1.27	0.38	0.48	0.09

Çizelge 4.120. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 8 terim tekrar sayısı : 272

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.15	0.32	0.21	0.03	45.72
de	2.86	0.95	0.36	0.39	1.03
en	1.06	0.88	0.27	0.47	0.78
tr	2.38	0.88	0.46	0.31	1.92
es	0.21	0.28	0.44	0.42	5.47
pt	6.58	1.04	0.32	0.48	0.07

Çizelge 4.121. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 8 terim tekrar sayısı : 1301

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	3.93	0.89	2	0.03	25.24
fr	2.79	0.82	0.95	0.08	19.77
pt	1.58	0.48	0.48	0.41	6.04
tr	2.69	0.61	0.66	0.15	16.61
de	3.81	0.46	0.42	0.18	19.22
en	2.33	0.45	0.31	0.2	22.13

Çizelge 4.122. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 8 terim tekrar sayısı : 1353

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	3.93	0.89	2	0.03	25.24
fr	1.07	0.19	0.61	0.08	46.22
pt	2.83	0.43	0.59	0.42	3.84
tr	1.83	0.33	0.55	0.18	26.27
de	7.65	1.32	1.19	0.19	21.34
en	6.11	1.29	1.01	0.21	19.16

Çizelge 4.123. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-3 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 8 terim tekrar sayısı : 1235					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	2.79	0.82	0.95	0.08	19.77
es	1.07	0.19	0.61	0.08	46.22
pt	1.73	0.35	0.46	0.43	4.61
tr	0.95	0.24	0.34	0.22	25.53
de	6.41	1.2	0.83	0.23	16.34
en	4.86	1.17	0.7	0.24	15.29

4.2.4. ÖVÇY-4 Performans Ölçütleri

Çizelge 4.124. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=12

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 8 terim tekrar sayısı : 227					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.01	0.09	0	0	68.64
it	0.46	0.54	0.41	0.09	17.89
pt	0.49	0.79	0.48	0.1	6.88
tr	0.31	0.15	0.05	0.01	8.16
fr	0.25	0.53	0.17	0.09	16.9
es	0.45	0.64	0.45	0.01	9.86

Çizelge 4.125. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=12

Terim sayısı : 2994
Entropi H(de) : 11.344
En sık geçen 8 terim tekrar sayısı : 250

de	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.01	0.09	0	0	68.64
it	0.46	0.57	0.47	0.08	19.78
pt	0.49	0.74	0.5	0.1	7.97
tr	0.31	0.16	0.08	0.01	13.43
fr	0.24	0.47	0.17	0.08	17.16
es	0.45	0.68	0.52	0.01	11.41

Çizelge 4.126. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766
Entropi H(fr) : 10.015
En sık geçen 8 terim tekrar sayısı : 924

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.09	0.36	0.05	0.11	39.42
en	0.49	0.55	1.17	0.03	20.94
pt	0.61	0.75	1.47	0.03	17.28
de	0.13	0.75	0.04	0.15	10.76
es	0.31	0.74	0.92	0.06	13.92
it	0.25	0.64	0.77	0.04	20.1

Çizelge 4.127. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 8 terim tekrar sayısı : 979					
tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.09	0.36	0.05	0.11	39.42
en	0.48	0.57	1.02	0.07	19.58
pt	0.52	0.75	1.21	0.08	23.17
de	0.12	0.99	0.05	0.31	10.45
es	0.3	0.69	1.17	0.02	29.41
it	0.21	0.62	0.78	0.07	27.69

Çizelge 4.128. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 8 terim tekrar sayısı : 2398					
es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.33	0.34	2.47	0	38.33
fr	0.36	0.44	1.1	0.04	41.78
it	0.44	0.9	1.57	0.05	17.75
en	0.16	0.08	0.14	0	41.27
pt	0.24	0.41	1.04	0.01	37.39
de	0.33	0.55	0.3	0.05	13.01

Çizelge 4.129. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 8 terim tekrar sayısı : 2711					
tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.33	0.34	2.47	0	38.33
fr	0.12	0.13	0.59	0.02	69.25
it	0.76	1.2	2.26	0.06	19.44
en	0.44	0.37	2.02	0	22.22
pt	0.55	0.73	3.18	0	22.12
de	0.59	0.85	0.67	0.06	9.26

Çizelge 4.130. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 8 terim tekrar sayısı : 978					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.26	1.02	0.68	0.11	16.89
it	0.67	1.07	2.28	0.01	18.27
pt	0.13	0.75	0.11	0.03	21.23
es	0.16	0.61	0.6	0.01	36.97
de	0.46	0.32	0.82	0.03	20.54
tr	0.44	1.14	0.89	0.15	12.95

Çizelge 4.131. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621 Entropi H(fr) : 11.211 En sık geçen 8 terim tekrar sayısı : 1147					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.26	1.02	0.68	0.11	16.89
it	0.37	0.19	0.74	0.06	21.96
pt	0.2	0.24	0.09	0.16	7.36
es	0.12	0.38	0.45	0.07	35.3
de	0.69	1.26	1.52	0.11	22.17
tr	0.22	0.11	0.39	0.09	32.39

Çizelge 4.132. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 8 terim tekrar sayısı : 270					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.01	0.11	0.01	0.1	57.91
de	0.34	0.78	0.47	0.01	20.18
en	0.51	0.37	0.4	0.02	9.99
tr	0.3	0.84	0.43	0.01	20.49
es	0.24	0.35	0.2	0.01	18.37
pt	0.61	0.77	0.77	0.02	1

Çizelge 4.133. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741 Entropi H(it) : 11.178 En sık geçen 8 terim tekrar sayısı : 272					
it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.01	0.11	0.01	0.1	57.91
de	0.34	0.76	0.37	0.07	14.2
en	0.46	0.42	0.42	0.05	8.25
tr	0.3	0.83	0.34	0.09	12.76
es	0.2	0.42	0.22	0.08	15.04
pt	0.61	0.76	0.56	0.08	12.36

Çizelge 4.134. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543 Entropi H(it) : 9.909 En sık geçen 8 terim tekrar sayısı : 1301					
it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	1.4	1.02	4.67	0.01	30.6
fr	1.06	0.93	2.68	0.02	53.61
pt	1.98	0.94	7.88	0	22.44
tr	1.53	0.82	3.19	0.01	24.43
de	0.6	0.13	0.27	0.01	29.11
en	0.39	0.19	0.38	0.02	50.2

Çizelge 4.135. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 8 terim tekrar sayısı : 1353

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	1.4	1.02	4.67	0.01	30.6
fr	0.35	0.1	1.78	0	62.93
pt	0.75	0.08	1.47	0	53.72
tr	0.58	0.21	2.32	0	50.23
de	1.8	1.05	6.91	0	28.98
en	1.42	0.89	5.44	0	31.87

Çizelge 4.136. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-4 benzerlik ölçümleri

Benzer dokümanlar: Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827
Entropi H(fr) : 10.586
En sık geçen 8 terim tekrar sayısı : 1235

fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	1.06	0.93	2.68	0.02	53.61
es	0.35	0.1	1.78	0	62.93
pt	1.04	0.07	0.51	0.01	45.4
tr	0.73	0.14	0.78	0.01	47.38
de	1.5	0.96	3.98	0.01	31.91
en	1.11	0.8	3.44	0.01	35.61

4.2.5. ÖVÇY-5 Performans Ölçütleri

Çizelge 4.137. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=10

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 12 terim tekrar sayısı : 227					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.3	0.07	0.07	0.01	74.71
it	2.07	0.41	1.17	0.02	51.86
pt	0.88	0.19	0.72	0.01	49.34
tr	0.93	0.19	0.72	0	46.93
fr	1.32	0.29	1.08	0	50.36
es	1.21	0.24	1.26	0.02	45.64

Çizelge 4.138. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Almanca (de)
k=10

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 12 terim tekrar sayısı : 250					
de	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.3	0.07	0.07	0.01	74.71
it	2.05	0.4	1.04	0.01	54.57
pt	0.85	0.18	0.59	0.01	54.39
tr	0.87	0.18	0.58	0.01	54.54
fr	1.3	0.28	1.02	0.01	49.15
es	1.21	0.24	1.07	0.03	48.32

Çizelge 4.139. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 12 terim tekrar sayısı : 924					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	0.22	0.04	0.49	0.01	86.52
en	2.63	0.52	1.22	0.05	62.85
pt	2.28	0.46	1.06	0.01	61.9
de	1.39	0.33	0.86	0.25	32.79
es	2.53	0.5	1.21	0.01	59.91
it	2.21	0.48	1.46	0.1	59.23

Çizelge 4.140. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-5 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 12 terim tekrar sayısı : 979					
tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.22	0.04	0.49	0.01	86.52
en	2.52	0.5	0.91	0.02	58.84
pt	2.16	0.44	0.75	0.02	59.15
de	1.36	0.33	0.74	0.19	34.19
es	2.45	0.48	1.4	0	55.19
it	2.11	0.47	1.14	0.06	56.61

Çizelge 4.141. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 12 terim tekrar sayısı : 2398					
es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
tr	2.48	0.35	0.66	0	61.88
fr	1.13	0.17	1.56	0.1	64.57
it	2.09	0.41	0.48	0.01	50.78
en	1.07	0.2	0.16	0.04	49.55
pt	1.42	0.27	0.4	0.01	59.86
de	1.58	0.33	0.18	0.1	38.93

Çizelge 4.142. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 12 terim tekrar sayısı : 2711					
tr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	2.48	0.35	0.66	0	61.88
fr	1.47	0.2	2.11	0.01	68.43
it	4.5	0.75	1.06	0.01	48.51
en	2.98	0.44	0.76	0.04	46.09
pt	3.78	0.61	1.12	0.01	55.07
de	3.49	0.56	0.75	0.11	35.92

Çizelge 4.143. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 12 terim tekrar sayısı : 978					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.44	0.09	1.03	0.09	52.87
it	1.04	0.24	0.24	0.07	51.18
pt	0.47	0.12	0.42	0.03	50.73
es	0.62	0.11	2.34	0.07	52.81
de	1.63	0.29	1.66	0.07	45.04
tr	1.35	0.3	0.93	0.08	38.82

Çizelge 4.144. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621 Entropi H(fr) : 11.211 En sık geçen 12 terim tekrar sayısı : 1147					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
en	0.44	0.09	1.03	0.09	52.87
it	1.32	0.29	1.11	0.01	61.05
pt	0.47	0.1	0.58	0.09	37.5
es	0.26	0.05	0.47	0.02	79.79
de	1.88	0.35	0.93	0.02	53.9
tr	1.16	0.23	0.47	0.02	56.85

Çizelge 4.145. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 12 terim tekrar sayısı : 270					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.28	0.07	0.55	0.06	58.41
de	0.79	0.19	0.53	0.07	43.98
en	1.32	0.26	1.21	0.1	45.5
tr	0.71	0.16	0.72	0.07	49.98
es	0.77	0.15	1.13	0.08	48.84
pt	1.14	0.23	0.79	0.06	46.78

Çizelge 4.146. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741 Entropi H(it) : 11.178 En sık geçen 12 terim tekrar sayısı : 272					
it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
fr	0.28	0.07	0.55	0.06	58.41
de	0.78	0.17	0.37	0.1	54.36
en	1.42	0.25	1.17	0.01	54.51
tr	0.87	0.19	0.88	0.1	56.59
es	0.91	0.17	1.3	0.01	59.9
pt	1.22	0.23	0.84	0.01	56.19

Çizelge 4.147. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 12 terim tekrar sayısı : 1301

it	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
es	0.3	0.05	1.1	0	72.73
fr	0.78	0.16	0.57	0	60.74
pt	1.73	0.34	1.08	0.01	48.43
tr	0.93	0.2	1.65	0.01	51.88
de	0.79	0.16	0.53	0.01	59.76
en	0.9	0.19	1.73	0.02	53.89

Çizelge 4.148. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621
Entropi H(es) : 10.557
En sık geçen 12 terim tekrar sayısı : 1353

es	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.3	0.05	1.1	0	72.73
fr	0.97	0.2	1.19	0.01	61.42
pt	1.86	0.37	0.77	0.01	52.93
tr	0.96	0.21	0.73	0.01	54.07
de	0.97	0.19	1.38	0.01	61.38
en	0.66	0.15	0.8	0.01	65.37

.Çizelge 4.149. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-5 benzerlik ölçümleri

k=10
Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 12 terim tekrar sayısı : 1235					
fr	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
it	0.78	0.16	0.57	0	60.74
es	0.97	0.2	1.19	0.01	61.42
pt	1.05	0.2	1.02	0.01	65.1
tr	0.7	0.14	1.84	0.03	59.65
de	0.28	0.06	0.14	0.01	80.51
en	1.6	0.34	1.92	0.03	54.93

4.2.6. ÖVÇY-6 Performans Ölçütleri

Çizelge 4.150. 990 numaralı doküman setindeki İngilizce doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 En sık geçen 25 terim tekrar sayısı : 96					
en	BY1 Öklid	BY2 Kosinüs	BY3 Mahalanobis	BY4 Pearson	BY5 MeanOfMinMax
de	0.105	0.501	0.073	0.225	65.46
it	0.271	0.607	0.597	0.486	0.01
pt	0.312	0.639	0.567	0.512	0.03
tr	0.24	0.66	0.652	0.427	0.03
fr	0.214	0.586	0.571	0.528	0.05
es	0.257	0.521	0.579	0.463	0.04

Çizelge 4.151. 990 numaralı doküman setindeki Almanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar: İngilizce (en) ve Almanca (de)

Terim sayısı : 2994 Entropi H(de) : 11.344 En sık geçen 25 terim tekrar sayısı : 91					
de	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.105	0.501	0.073	0.225	65.46
it	0.28	0.689	0.587	0.48	0.01
pt	0.329	0.729	0.552	0.495	0.02
tr	0.246	0.707	0.628	0.445	0.03
fr	0.221	0.641	0.557	0.46	0.04
es	0.27	0.622	0.566	0.469	0.03

Çizelge 4.152. 987 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 25 terim tekrar sayısı : 567					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.284	0.56	0.088	0.48	44.36
en	0.537	0.54	0.35	0.479	3.29
pt	0.538	0.537	0.422	0.49	2.78
de	0.084	0.708	0.588	0.514	0
es	0.298	0.537	0.216	0.421	19.2
it	0.275	0.528	0.117	0.428	44.73

Çizelge 4.153. 987 numaralı doküman setindeki Türkçe doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve Türkçe (tr)

Terim sayısı : 1945 Entropi H(tr) : 9.653 En sık geçen 25 terim tekrar sayısı : 526					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.284	0.56	0.088	0.48	44.36
en	0.554	0.576	0.377	0.51	2.44
pt	0.534	0.514	0.439	0.486	2.05
de	0.074	0.625	0.584	0.536	0
es	0.268	0.528	0.175	0.474	15.45
it	0.257	0.549	0.056	0.436	37.59

Çizelge 4.154. 983 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 25 terim tekrar sayısı : 1765					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
tr	0.525	0.505	0.076	0.437	59.81
fr	0.557	0.525	0.081	0.521	30.41
it	0.481	0.594	0.346	0.472	4.8
en	0.452	0.555	0.32	0.476	4.22
pt	0.483	0.548	0.356	0.404	7.24
de	0.278	0.684	0.605	0.454	0.01

Çizelge 4.155. 983 numaralı doküman setindeki Türkçe doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İspanyolca (es) ve Türkçe (tr)

Terim sayısı : 7151 Entropi H(tr) : 11.986 En sık geçen 25 terim tekrar sayısı : 1545					
tr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.525	0.505	0.076	0.437	59.81
fr	0.542	0.525	0.153	0.501	24.13
it	0.439	0.574	0.392	0.437	7.74
en	0.406	0.537	0.365	0.453	6.81
pt	0.489	0.583	0.422	0.427	11.21
de	0.23	0.658	0.63	0.431	0.01

Çizelge 4.156. 972 numaralı doküman setindeki İngilizce doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 25 terim tekrar sayısı : 568					
en	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.259	0.527	0.117	0.393	56.88
it	0.481	0.545	0.438	0.461	1.23
pt	0.197	0.627	0.525	0.476	0.2
es	0.273	0.526	0.191	0.404	12.38
de	0.412	0.565	0.447	0.495	4.44
tr	0.42	0.551	0.378	0.416	3.13

Çizelge 4.157. 972 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : İngilizce (en) ve Fransızca (fr)

Terim sayısı : 4621 Entropi H(fr) : 11.211 En sık geçen 25 terim tekrar sayısı : 642					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
en	0.259	0.527	0.117	0.393	56.88
it	0.491	0.534	0.384	0.457	2.18
pt	0.188	0.639	0.548	0.446	0.11
es	0.295	0.557	0.078	0.491	18.78
de	0.403	0.515	0.371	0.53	6.57
tr	0.414	0.518	0.31	0.456	5.48

Çizelge 4.158. 935 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 25 terim tekrar sayısı : 109					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.11	0.441	0.045	0.221	77.56
de	0.167	0.63	0.617	0.483	0.2
en	0.292	0.592	0.581	0.494	0.03
tr	0.175	0.662	0.577	0.419	0.46
es	0.212	0.617	0.539	0.43	0.2
pt	0.371	0.596	0.561	0.528	0.01

Çizelge 4.159. 935 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr) ve İtalyanca (it)

Terim sayısı : 2741
Entropi H(it) : 11.178
En sık geçen 25 terim tekrar sayısı : 109

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
fr	0.11	0.441	0.045	0.221	77.56
de	0.19	0.695	0.616	0.435	0.19
en	0.298	0.634	0.566	0.519	0.02
tr	0.185	0.666	0.573	0.379	0.46
es	0.211	0.61	0.534	0.457	0.19
pt	0.377	0.649	0.556	0.508	0.01

Çizelge 4.160. 887 numaralı doküman setindeki İtalyanca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 2543
Entropi H(it) : 9.909
En sık geçen 25 terim tekrar sayısı : 973

it	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
es	0.326	0.571	0.035	0.47	67.04
fr	0.284	0.553	0.044	0.472	60.83
pt	0.586	0.547	0.427	0.519	2.99
tr	0.371	0.546	0.329	0.464	10.33
de	0.335	0.535	0.235	0.525	19.23
en	0.343	0.517	0.242	0.483	16.49

Çizelge 4.161. 887 numaralı doküman setindeki İspanyolca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3621 Entropi H(es) : 10.557 En sık geçen 25 terim tekrar sayısı : 1022					
es	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.326	0.571	0.035	0.47	67.04
fr	0.284	0.526	0.067	0.524	58.91
pt	0.583	0.536	0.401	0.499	3.76
tr	0.395	0.571	0.311	0.471	12.44
de	0.361	0.558	0.21	0.468	23.36
en	0.358	0.531	0.209	0.517	20.43

Çizelge 4.162. 887 numaralı doküman setindeki Fransızca doküman için ÖVÇY-6 benzerlik ölçümleri

Benzer dokümanlar : Fransızca (fr), İtalyanca (it) ve İspanyolca (es)

Terim sayısı : 3827 Entropi H(fr) : 10.586 En sık geçen 25 terim tekrar sayısı : 868					
fr	BY 1 Öklid	BY 2 Kosinüs	BY 3 Mahalanobis	BY 4 Pearson	BY 5 MeanOfMinMax
it	0.284	0.553	0.044	0.472	60.83
es	0.284	0.526	0.067	0.524	58.91
pt	0.588	0.563	0.443	0.521	2.1
tr	0.357	0.557	0.332	0.488	7.3
de	0.322	0.539	0.247	0.5	14.88
en	0.329	0.516	0.262	0.51	12.12

4.2.7. MeanOfMinMax Benzerlik Ölçümü Sonuçları

Çizelge 4.163. 990 doküman setindeki İngilizce dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İngilizce (en). Almanca (de)

Terim sayısı : 4069 Entropi H(en) : 11.86 Ensık geçen 8 terim tekrar sayısı: 227						
en	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
de	55.8	58.13	51.43	68.64	74.71	65.46
it	1.58	0.4	1.56	17.89	51.86	0.01
pt	0.24	0.52	0.22	6.88	49.34	0.03
tr	0.91	1.44	0.87	8.16	46.93	0.03
fr	0.5	0.75	0.49	16.9	50.36	0.05
es	1.33	0.65	1.27	9.86	45.64	0.04

Çizelge 4.164. 987 doküman setindeki Fransızca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), Türkçe (tr)

Terim sayısı : 2766 Entropi H(fr) : 10.015 En sık geçen 8 terim tekrar sayısı : 924						
fr	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
tr	24.91	44.56	24.51	39.42	86.52	44.36
en	3.41	3.39	3.58	20.94	62.85	3.29
pt	5.6	4.9	5.52	17.28	61.9	2.78
de	0.04	0.1	0.04	10.76	32.79	0
es	23.08	50.46	22.65	13.92	59.91	19.2
it	25.24	41.53	24.68	20.1	59.23	44.73

Çizelge 4.165. 983 doküman setindeki İspanyolca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İspanyolca (es). Türkçe (tr)

Terim sayısı : 11181 Entropi H(es) : 12.795 En sık geçen 25 terim tekrar sayısı : 1765						
es	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
tr	39.79	56.28	42.21	38.33	61.88	59.81
fr	42.74	51.1	39.17	41.78	64.57	30.41
it	9.94	8.74	9.98	17.75	50.78	4.8
en	11.56	10.97	11.84	41.27	49.55	4.22
pt	26.27	42.46	26.81	37.39	59.86	7.24
de	0.15	0.21	0.15	13.01	38.93	0.01

Çizelge 4.166. 972 doküman setindeki İngilizce dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : İngilizce (en). Fransızca (fr)

Terim sayısı : 2241 Entropi H(en) : 9.611 En sık geçen 25 terim tekrar sayısı : 568						
en	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
fr	19.85	35.81	19.41	16.89	52.87	56.88
it	4.63	4.02	15.88	18.27	51.18	1.23
pt	2.89	2.96	3.54	21.23	50.73	0.2
es	34.43	43.98	43.87	36.97	52.81	12.38
de	7.52	9.25	11.85	20.54	45.04	4.44
tr	6.84	6.3	14.12	12.95	38.82	3.13

Çizelge 4.167. 935 doküman setindeki Fransızca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), İtalyanca (it)

Terim sayısı : 2332 Entropi H(fr) : 10.859 En sık geçen 8 terim tekrar sayısı : 270						
fr	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
it	47.86	63.52	45.72	57.91	58.41	77.56
de	1.2	2.22	1.28	20.18	43.98	0.2
en	1.01	0.34	0.93	9.99	45.5	0.03
tr	2.44	3.76	2.45	20.49	49.98	0.46
es	7.58	3.07	6.82	18.37	48.84	0.2
pt	0.1	0.14	0.09	1	46.78	0.01

Çizelge 4.168. 887 doküman setindeki İtalyanca dokümanın 6 adet Öznitelik Vektörü ile MeanOfMinMax Benzerlik Ölçümü yönteminde karşılaştırılması

Benzer dokümanlar : Fransızca (fr), İtalyanca (it), İspanyolca (es)

Terim sayısı : 2543 Entropi H(it) : 9.909 En sık geçen 8 terim tekrar sayısı : 1301						
it	ÖVÇY-1	ÖVÇY-2	ÖVÇY-3	ÖVÇY-4	ÖVÇY-5	ÖVÇY-6
es	25.15	51.1	25.24	30.6	72.73	67.04
fr	19.46	38.31	19.77	53.61	60.74	60.83
pt	6.18	3.98	6.04	22.44	48.43	2.99
tr	16.42	23.85	16.61	24.43	51.88	10.33
de	19.54	25.45	19.22	29.11	59.76	19.23
en	19.51	23.43	22.13	50.2	53.89	16.49

4.3. BULGULAR VE DEĞERLENDİRME

NATO veri setinde bulunan dokümanlar üzerinde test edilen belge benzerliği ölçümlerinde, 6 farklı yöntem test edilmiştir ve sonuçlara bakıldığında "*MeanOfMinMax* " yöntemi ile yapılan benzerlik ölçümlerinde en başarılı yöntemler ÖVÇY-5 ve ÖVÇY-6 gözükmektedir.

ÖVÇY-1 vektörleri tek boyutlu olarak, normalleştirilmiş, pozisyon değerleri taşımaktadır. Çizelge 4.80 incelendiğinde 990 numaralı veri setinin özelliği gereği içinde bulunan benzeyen ve benzemeyen dosyalar içerisinde hem İngilizce hem de Almanca için dokümanların paralel metinlerini benzerlik yüzdesinde 1. sırada getirmesi algoritmayı başarılı kılmaktadır. Diğer benzemeyen dokümanlar için ise düşük bir yüzde çıkması metinlerde anlama ve sözlüğe bakılmadan istatistiksel yollarla hesaplama yapan yöntemi başarılı göstermektedir ve algoritmamızı desteklemektedir. ÖVÇY-1 yönteminin zayıf yanı Çizelge 4.82'de görüldüğü üzere, tek boyutlu vektör olmasından dolayı, benzerlik-mesafe ölçümlerinde dosya boyutunun (aynı zamanda ÖV boyutu) artması ile vektörlerin uzayda birbirine benzemeye başlaması ve ölçümlerin başarısız veya tesadüfi benzerliklere yol açmasıdır. Ayrıca bir doküman içindeki her kelimenin o dokümanın özniteliğine olan katkısı farklıdır. Bu yüzden ağırlık değerlerinin göz ardı edilmesi benzemez dokümanlarda da tesadüfi yüksek sonuçlara neden olmaktadır.

ÖVÇY-2 yöntemi, Bölüm 4.1.7 incelendiğinde başarılı sonuçlar göstermiştir. Çizelge 4.15-4.25 incelendiğinde pozisyon değerinin yanı sıra terim sıklığının ağırlık olarak verilmesi, Mahalonobis değerlerini daha başarılı göstermiştir. ÖVÇY-1'de olduğu gibi diğer ilginç nokta ise birbirine benzeyen dokümanların test sonuçlarının başka benzemeyen dokümanlar ile karşılaştırıldığında yakın değerlerde görünmesidir. Dolayısıyla oluşturduğumuz ÖV yönteminin paralel metinler için benzer sonuçlar göstermesi, aynı anlamdaki dokümanların karakteristiğinin öznitelik vektörlerine başarılı bir şekilde aktarıldığını göstermektedir. ÖVÇY-2 yönteminde başarılı sonuçlar alınmasına rağmen benzemeyen dokümanlarda da yüksek değerlere rastlanmıştır. Bu durum doküman frekans dağılımının karakteristiğinin daha da güçlendirilmesi gerekliliğini doğurmuştur.

ÖVÇY-3 yönteminde başarı oranı %56'ya düşmüştür, dokümanların kayması nedeniyle, pozisyon ve ağırlık çarpımı, nitelik değerini düşürmüştür. ÖVÇY-1 ve

ÖVÇY-2 değerlerinden düşük olsa da paralel metinler için diğerlerine göre daha pozitif sonuçlar çıkmıştır. ÖVÇY-3 yöntemindeki sonuçlardan ÖV oluşturulurken bütün yöntemlerde temel olarak baz alınan önişleme-En sık geçen kelimelerin pozisyonlarının baz alınması ile oluşturulan hesaplamaların işe yaradığını görüyoruz. ÖVÇY-1 ve ÖVÇY-2'de olduğu gibi benzeyen dokümanların, benzemeyen dokümanlara mesafeleri de benzer çıkmıştır.

ÖVÇY-4 yöntemi, ÖVÇY-1 yönteminin k-ortalama ile kümelendirilmiş ve boyutu indirgenmiş halidir. 990 numaralı veri seti için ÖVÇY-1 ile oluşturulan ÖV k-ortalama, 12 adet sendroide dağıtılmış ve bu noktalar dokümanın pozisyon merkezleri olarak belirlenmiştir. Her dilin kendine has morfolojisinden dolayı paralel metinlerde geçen benzer kelimelerin pozisyonları arasında kaymalar olabilir. K-ortalama algoritması bu şekilde dokümanları birbirine daha doğru hizalamakta ve elde edilen pozisyonlar normalize edilerek ölçümlerin ÖVÇY-1, ÖVÇY-2, ÖVÇY-3'e göre daha yüksek çıktığı gözlemlenmiştir. ÖVÇY-4 neticesinde elde edilen ölçüm sonuçları ÖVÇY-2'den az miktarda yüksek olsa da diğer mesafe ve benzerlik ölçümünde yaklaşık %500'lük bir iyileştirme sağlamıştır.

ÖVÇY-4 başarısından dolayı ÖVÇY-5 yöntemi geliştirilmiştir. ÖVÇY-5 yöntemi ÖVÇY-2 yönteminin k-ortalama uygulanmış halidir. Elde edilen sendroidlere ağırlık ortalama değerleri verilmiştir. ÖVÇY-5 yönteminde; k sayısının düşüklüğü ve öznitelik vektörünün boyutunu indirgemekte ve işlem süresini azaltmaktadır. Sonuç olarak önceki 4 yöntemden pozitif bir sonuç çıkmıştır. ÖVÇY-5 yöntemi uzun dokümanlarda daha başarılı sonuçlar vermektedir. ÖVÇY-5 yönteminde bir diğer dikkat çeken nokta mesafe ve benzerlik ölçümlerinin paralel dokümanlarda diğer dokümanlara göre belirgin farklar göstermesidir.

ÖVÇY-6 yönteminde doküman içerisinde en değerli kelimelerin pozisyonları sendroid olarak kabul edilmiş ve bu noktalar ağırlığı, yapılan ölçümler neticesinde MI Skor ile verilmiştir. ÖVÇY-6 ile elde edilen sonuçlar ÖVÇY-5'den daha düşük olsa da, ÖVÇY-6 ile belirlenen sendroid değerleri paralel dokümanlarda benzer anlamalar içermektedir ve bu yüzden daha doğrusal sonuçlar vermektedir. Ayrıca ÖVÇY-6 ölçümlerinde birbirine paralel olmayan dokümanların benzerlik yüzdeleri diğer ölçümlere göre daha düşük çıkmaktadır. Dolayısıyla, başarı ve hata oranları ele alındığında ÖVÇY-6 en doğru sonuçları veren yöntem olarak belirlenmiştir.

5. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında farklı dillerdeki paralel dokümanların benzerliğinin saptanmasına dair yeni öznitelik vektörü çıkarım yöntemleri ve benzerlik ölçümü geliştirilmiş ve bu yöntemler test edilmiştir.

Farklı dillerdeki dokümanların benzerliğinin saptanmasına dair literatürde yapılan araştırmalar iki gruba ayrılmıştır. Birinci grup dilbilgisi ve sözlükten faydalanarak, çift-dilli sözlükler kullanmış ve dilin anlamından faydalanarak karşılıklı benzerlik ölçümleri yapmışlardır. Başarılı araştırmalarda bile en sık rastlanan problem performans ve değişik morfolojideki yapısal farklılıklar olmuştur. İkinci grup araştırmacılar ise anlamdan ziyade istatistiksel ölçümler ile doküman benzerliklerini saptamaya çalışmışlardır. İstatistiksel ölçümler birinci gruba göre performans açısından başarılı olsa da dillerin farklı yapısından sonuçlarda hata oranlarına rastlanmıştır.

Bu tez ağırlıklı olarak birinci grubun izlediği istatistiksel yöntemler izlense de dokümanlar üzerinde önişlemler, normalizasyon, gövdeleme, etkisiz kelimelerin kaldırılması, dokümanı ifade eden süper kelimelerin seçilip bu kelimeler üzerinden işlem yapılması ve bu kelimelere eşdizimlilik ölçümlerinin eklenmesi ile semantiğe bakılmadan anlam sayısal verilere yüklenmiştir.

Yapılan literatür araştırmasında ve ÖVÇY-6 ölçümleri neticesinde paralel dokümanlarda kelime yerleşimlerinin benzerlik gösterdiği saptanmış, doküman içeriğini ve özetini yansıtan, o dokümana has kelimelerin sıklığının benzer dokümanlarda yaklaşık olarak aynı lokasyonlarda olduğuna, bu lokasyonlarda eşdizimlilik skorlarının da benzer çıktığı ve pozisyon değerlerinin hizalanması ile yapılan mesafe ve benzerlik ölçümlerinde %85'lik bir benzerlik ölçüsüne ulaşılmıştır. Benzerliğin yanı sıra benzemeyen dokümanlarda düşük bir benzerlik yüzdesi ve hataya rastlanmıştır. Bazı benzemeyen dokümanlarda ortaya çıkan tesadüfî benzerliklerin ise doküman hizalarının en güçlü kelimeye göre alınması ile düşürüldüğü gözlemlenmiştir. Doküman hizalama işlemlerinde referans noktası olarak dokümana ait en güçlü kelimenin pozisyonun alınması ciddi anlamda başarılı neticeler vermiştir.

Ayrıca oluşturulan yeni bir benzerlik ölçümü ile bu öznitelik vektörlerinin benzerliği daha doğru bir şekilde ifade edilmiştir. Bu benzerlik ölçümü ile

oluşturulan dokümana ait noktasal pozisyon bilgilerinin yanı sıra doküman boyutu ve doküman dağılımı yani entropisi, benzerlik ölçümünde parametre olarak kullanılıp başarılı neticeler alınmıştır.

Benzerlik ölçümlerinde dokümanlar çeşitli önişlemlerden geçirilmiş, dokümanlardaki etkisiz kelimeler kaldırılmış, terimlerin kökleri gövdeleme yöntemi ile atılmıştır. Bu önişlemlerin her birisi için literatürde bulunan ama dilden dile farklılık gösteren listelerden faydalanılmıştır. Literatür çalışmalarında İngilizce, Fransızca ve İspanyolca dokümanların başat olmasından dolayı bu dillere ait etkisiz kelime ve gövdeleme algoritmaları daha geniş kapsamlı oluşturulmuştur. Bunun neticesinde bu dillerde yapılan ölçümlerde başarı oranı daha yüksek sağlanmıştır.
Kullanılan veri setindeki tüm dillerin gövdeleme ve etkisiz kelimeler listelerinin daha kapsamlı ve paralel zenginlikte oluşması durumunda başarılı oranlarının artabileceği gözlemlenmiştir.

Bu çalışma ile paralel dildeki doküman benzerliği konusunda bu konuda çalışma yapan akademik çevrelere ve mühendislere performansı yüksek, işlem hızı düşük bir öznitelik vektörü yöntemi ve benzerlik ölçümü önerilmiştir. Bu sayede farklı dillerdeki belgelerin benzerliğinin içerik olarak tespit edilmesi ve bunun yanısıra farklı dillerde de olsa intihal belgelerinin saptanmasına yardımcı olması hedeflenmiştir.

KAYNAKLAR

[1] ITU: Commited to connect the world, http://www.itu.int/en/default.aspx (23.12.2012)

[2] Lazarinis, F., Vilares, J. ve Tait, J. I. " Improving non-English web searching", ACM SIGIR Forum 41(2), (2007).

[3] The size of the World Wide Web (The Internet), http://www.worldwidewebsize.com/ (07.08.2013)

[4] "Languages on the web" New Media Trend Watch, http://www.newmediatrendwatch.com/world-overview/92-languages-on-the-web (26.01.212)

[5] Lee, C., Yang, H., Chen, T. ve Ma, S., "A Comparative Study on Supervised and Unsupervised Learning Approaches for Multilingual Text Categorization" First International Conference on Innovative Computing, Information and Control - Volume II, (ICICIC'06) Başlılı Bildiri Özetleri kitabı , Beijing, 511-514, (2006).

[6] Salton, G., ve McGill, M., "Introduction to Modern Information Retrieval," McGraw Hill, New York, 448 s., (1983).

[7] Savoy, J. "Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach", Evaluation of Cross-Language Information Retrieval Systems Lecture Notes in Computer Science, 2406: 27-43, (2002).

[8] Eurovoc (1995). *Thesaurus EUROVOC - Volume 2: Subject-Oriented Version*. Ed. 3/English Language. Annex to the index of the Official Journal of the EC. Luxembourg, Office for Official Publications of the European Communities. http://eurovoc.europa.eu/ (06.12.2012)

[9] Mathieu, B., Besancon, R. ve Fluhr, C. "Multilingual document clusters discovery" In: RIAO 2004, Avignon, France, (2004).

[10] Steinberger, R., Pouliquen, B., ve Ignat, C. "Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications", 4th Slovenian Language Technology Conference, Information Society Başlıklı Bildiri Kitabı, Slovenia (2004).

[11] Bruno, P., Steinberger, R., Ignat C. ve Groeve, T. "Geographical Information Recognition and Visualisation in Texts Written in Various Languages", 19th Annual ACM Symposium on Applied Computing (SAC'2004), Special Track on Information Access and Retrieval (SAC-IAR) Başlıklı Bildiri Özetleri Kitabı, Nicosia, Cyprus, 1051-1058 (2004).

[12] Montalvo, S., Martinez, R., Casillas, A. ve Fresno, V., "Multilingual document clustering: a heuristic approach based on cognate named entities" 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL Başlıklı Bildiri Özetleri Kitabı, (2004).

[13] Lee, C. ve Yang, H. "A multilingual text mining approach based on self-organizing maps." Applied Intelligence 18(3): 295-310, (2003).

[14] Gale, W. A. ve Church, K. W. "A program for aligning sentences in bilingual corpora." Computational linguistics 19 (1): 75-102, (1993).

[15] Kay, M. ve Röscheisen, M. "Text-translation alignment"Computational Linguistics 19 (1): 121-142, (1993).

[16] Wu, D., "Aligning a parallel English-Chinese corpus statistically with lexical criteria." 32nd annual meeting on Association for Computational Linguistics Başlıklı Bildiri Özetleri Kitabı, New Mexico, USA, 80-87, (1994).

[17] Salton G. "Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer" Addison-Wesley Pub., 530 s., (1989).

[18] Brown, P. F., Lai, J. C. ve Mercer, R. L., "Aligning sentences in parallel corpora." 29th annual meeting on Association for Computational Linguistics Başlıklı Bildiri Özetleri Kitabı, Stroudsburg, USA, 169-176 (1991).

[19] Vu, T., Aw, A. T. ve Zhang, M. "Feature-based method for document alignment in comparable news corpora" 112th Conference of the European Chapter of the ACL Başlıklı Bildiri Özetleri Kitabı, 843-851, (2009).

[20] Tao T. ve Zhai, C.. "Mining comparable bilingual text corpora for cross-language information integration." , The Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining Başlıklı nBildiri Özetleri Kitabı. Chicago, USA, 691-696, (2005).

[21] Munteanu, D. S. ve Marcu, D. "Extracting parallel sub-sentential fragments from non-parallel corpora", 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL Başlıklı Bildiri Özetleri Kitabı, 81–88, Sydney, (2006).

[22] Ogilvie, P. ve Callan, J. P. "Experiments using the lemur toolkit", Tenth Text Retrieval Conference (TREC) Başlıklı Bildiri Özetleri Kitabı, 103-108, (2001).

[23] W3 Techs Web Technology Surveys, http://w3techs.com/technologies/overview/content_language/all (12.12.2012)

[24] Nato Review, http://www.nato.int/docu/review/index_EN.htm

[25] Basic information on the European Union , http://europa.eu/about-eu/basic-information/ (12.12.2012)

[26] Salton, G. ve Buckley, C. "Term-weighting approaches in automatic text retrieval", Information Processing & Management 24(5): 513-523, (1998).

[27] Bilgisayar Kavramları - Vektör Uzay Modeli (Vector Space Model) http://www.bilgisayarkavramlari.com/2012/12/10/vektor-uzay-modeli-vector-space-model/ (03.03.2013)

[28] Manning, C. D., Raghavan, P. ve Schütze, H. "Introduction to information retrieval", Cambridge University Press, Cambridge, 496 p., (2008).

[29] Dunning, T. "Accurate methods for the statistics of surprise and coincidence", Computational Linguistics, 19(1):61-74, (1993).

[30] Wikipedia, The Free Encylopedia, Entrophy (information theory), http://en.wikipedia.org/wiki/Entropy_(information_theory) (21.11.2012).

[31] Porter, M.F., "Stemming algorithms for various European languages", http://www.snowball.tartarus.org/texts/stemmersoverview.html (03.11.2012)

[32] PHP PECL Porter Stemmer Extension, http://pecl.php.net/package/stem (02.11.2012).

[33] Huang, A. "Similarity measures for text document clustering." New Zealand Computer Science Research Student Conference (NZCSRSC) Başlıklı Bildiri Özetleri Kitabı, Christchurch, New Zealand, 49-56, (2008).

[34] İsmail Arı, Mahalanobis Uzaklığı, http://ismailari.com/blog/mahalanobis-uzakligi/ (29.04.2011).

ÖZGEÇMİŞ VE ESERLER LİSTESİ

Adı Soyadı: Hakan YILMAZER

Doğum Tarihi: 17/04/1979

Öğrenim Durumu: Yüksek Lisans

Derece	Bölüm/Program	Üniversite	Yıl
Lise	Fen Bilimleri	Gaziantep Tekerekoğlu Anadolu Lisesi	1994-1997
Lisans	Bilgisayar Mühendisliği	Mersin Üniversitesi	1998-2003
Yüksek Lisans	Bilgisayar Mühendisliği ABD	Mersin Üniversitesi	2010-2013

(Varsa) Görevler:

Görev Unvanı	Görev Yeri	Yıl
Uzman	Mersin Üniversitesi Bilgi İşlem Daire Başkanlığı	2003-2009
Bil. Müh.	Mersin Üniversitesi Bilgi İşlem Daire Başkanlığı	2009-…