T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLİMSEL YAYINLARDAN ANAHTAR KELİME ÇIKARIMI

Ahmet Sina BİRDEVRİM

Danışman
Dr. Öğr. Üyesi Ali BOYACI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
İSTANBUL - 2018

KABUL VE ONAY SAYFASI

Ahmet Sina BİRDEVRİM tarafından hazırlanan **"Bilimsel Yayınlarda Anahtar Kelime Çıkarımı"** adlı tez çalışması 14/08/2018 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Dr. Öğretim Üyesi Ali BOYACI** İstanbul Ticaret Üniversitesi	.....
Jüri Üyesi	**Dr.Öğretim Üyesi Mustafa Cem KASAPBAŞI** İstanbul Ticaret Üniversitesi	.....
Jüri Üyesi	**Dr.Öğretim Üyesi Oğuzhan ÖZTAŞ** İstanbul Üniversitesi	.....

Onay Tarihi : 17/10/2018

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

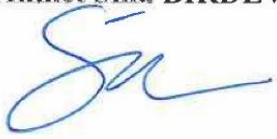
İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Tarih
14.08.2018
İmza

Ahmet Sina BİRDEVRİM



İÇİNDEKİLER

Sayfa

ÖZET

Yüksek Lisans Tezi

Bilimsel Yayınlardan Anahtar Kelime Çıkarımı

Ahmet Sina BİRDEVRİM

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ali BOYACI

2018, 44 sayfa

Akademik alanlarda yapılan çalışmalar okuyucuya, bilim ve teknoloji alanındaki ilerlemeler hakkında bilgi vermeyi amaçlamaktadır. Öte yandan okuyucuların bu bilimsel çalışmalarda aramakta oldukları bilgilere doğru ve hızlı bir şekilde ulaşabilmeleri gerekmektedir. Bu tür ihtiyaçlardan dolayı makale gibi belgelerde anahtar kelimelerin kullanılması, okuyucunun aranan bilgiye kolay ulaşmasını sağlar. Anahtar kelimeler çok sayıda metin tabanlı materyalin analizini kolaylaştırdığı gibi, istenen bilgiye hızlı ve kolay erişimi de sağlar. Bu verileri çıkarmak için otomatik anahtar kelime çıkarma algoritmaları kullanılabilir. Otomatik anahtar kelime çıkarma algoritmaları, belirli bir metinde yer alan en açık kelimeleri veya cümleleri ayıklamayı amaçlar. Bu amaç için en sık kullanılan algoritmalar TF-IDF ve RAKE dır. Fakat bu yaklaşımlar, pek çok metinde kullanılan aynı anlamı taşıyan farklı kelimeleri göz önüne alamamaktadır. Bu tezde geliştirilen algoritma ile verilen bir metindeki eşanlamlı kelimler tek bir çatı altında toplanarak bu kelimelerin sıklığı arttırılarak algoritmaların başarım oranları iyileştirmeye çalışılmıştır. Uygulanan bu yöntem diğer algoritmalar karşılaştırılmıştır ve sonuçlar ortaya konulmuştur.

Anahtar Kelimeler: Anahtar kelime çıkarımı, Anahtar kelime çıkarma algoritmaları, BRAKE.

ABSTRACT

M.Sc. Thesis

Keyword Extractions from Scientific Publications

Ahmet Sina BİRDEVRİM

İstanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Ali BOYACI

2018, 44 pages

The academic studies and publications aim to give insights to the reader about the progress in Science & Technology. On the other hand, readers should be able to reach the information they seek both promptly and accurately. For this reason, Keywords are used as search tools in obtaining the desired information from the articles and publications. Keywords facilitate the analysis of a large number of textual materials as well as provides rapid and easy access to desirable information. Automatic keyword algorithms, the method based on extracting the most explicit words or phrases in a given text, can be used to extract this data. TF-IDF and RAKE are the most commonly used algorithms for this purpose. The working principle of these methods do not take into account various form of words that have the same meaning. This study introduces a further algorithm that gathers synonym words in a text under one cluster and hence provides these words with increased frequencies. The proposed algorithm is compared with the other algorithms and some implications are presented.

Keywords: Keyword extraction, Keyword extraction algorithm, BRAKE.

TEŞEKKÜR

Bu çalışma esansında beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli danışman hocam Dr. Öğr. Üyesi Ali BOYACI'ya, değerli katkıları ve destekleri için Doç. Dr. Serhan YARKAN'a, değerli arkadaşım Alper Kaan YILDIRIM'a ve maddi ve manevi her türlü bana desteği sağlayan aileme teşekkürü borç bilirim.

Ahmet Sina BİRDEVRİM
İSTANBUL, 2018

ŞEKİLLER

ÇİZELGE

SİMGE VE KISALTMALAR

ACM	Association for Computer Machinery
BRAKE	Better Rapid Automatic Keyword Extraction
DB	Database
FN	False Negative
FP	False Positive
KEA	Key-phrase Extraction Algorithm
NLP	Natural Language Processing
PDF	Portable Document Format
POS	Part of Speech
RAKE	Rapid Automatic Keyword Extraction
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
TF-RR	Term Frequency Realized Relations
TP	True Positive
TN	True Negative
Deg	Derece
Freq	Frekans
w	Word (Kelime)

1. GİRİŞ

Bilimsel yayınlar, araştırmacılar arasında bir iletişim aracı olarak hizmet etmekle birlikte aynı zamanda akademik hayatta ve teknolojinin gelişimi gibi alanlarda bir çok fayda sağlamaktadır. Yazarlar ve araştırmacılar yeni bir bilimsel makaleye başlamadan önce yapılmış çalışmaları geliştirmek ve onları daha güncel hale getirmek üzere literatür taraması yapmaktadırlar.

Teknolojinin hızla büyüdüğü günümüzde artan bilgisayarların kapasiteleri ile birlikte depolanan verilerde çoğalmıştır. Bu yüzden okuyucuların aradıkları bilgilere doğru ve hızlı bir şekilde ulaşmaları, dijital metin tabanlı medyadaki en büyük zorluklardan biridir. Bilgisayarlar bu arşivleri aramak için yeterince hızlı olsa da, konuları üzerinden gruplanmış metinler üzerinde aramalar daha doğru sonuçlar üretmektedir. Bu metinleri gruplamak için genellikle anahtar kelimeler kullanılır.

Anahtar kelimeler, bir metni temsil eden en açık kelime veya kelime grubudur. Anahtar kelimeler okuyucuya aramakta olduğu veri hakkında bir ön fikir verir. Aynı zamanda anahtar kelimelerin kullanımı, okuyucunun bilgiye hızlı ve doğru bir şekilde ulaşmasını sağlar. Bu sayede okuyucu tüm metni gözden geçirmek yerine metin içeresinde geçen ve metni temsil eden anahtar kelimeler üzerinden metni okuyup okumayacağını karar verir.

Genellikle, anahtar kelimeler bir belgenin yazarları veya yayıncıları tarafından el ile seçilir. Bu durum çıkarılan anahtar kelimelerde doğruluk oranında insan faktörünü ön plana çıkararak oluşabilecek hata olasılığını arttırmakta olup okuyucunun aramakta olduğu bilgiye erişememesine sebep olabilir. Bunun yanı sıra el ile anahtar kelime çıkarma işlemleri harcanan zaman bakımından da maliyetli bir iştir. Bu tür sebeplerden dolayı otomatik anahtar kelime çıkarma algoritmalarından yararlanılabilir. Otomatik anahtar kelime çıkarma algoritmaları, bir yazıda konuyu en açık biçimde yansıtan kelime veya kelime gruplarını belirli bir algoritma yapısında çıkarması olarak tanımlanabilir. Otomatik anahtar kelime çıkarmada yaygın olarak TF-IDF, RAKE gibi algoritmalar kullanılmaktadır. Bu algoritmaların temelinde,

yüksek entropi değerine sahip metinlerden anlamlı anahtar kelimeler çıkarma amaçlanmıştır.

Bu tezde, metin içerisinde geçen eş anlamlı kelimeleri tek bir çatıda toplayarak entropi değerini düşürerek doğruluk oranını yükseltmek ve BRAKE algoritması çıkarılıp diğer algoritmalar ile karşılaştırmak amaçlanmıştır.

2. LİTERATÜR ÖZETİ

Bu bölümde anahtar kelime çıkarımı ile ilgili terimler açıklanmış olup bu alanda daha önceden çalışılmış yöntemler araştırılmıştır.

Anahtar kelime çıkarma, metin madenciliği alanında önemli bir kavramdır. Denetlenen ve denetlenmeyen makine öğrenimi, istatistiksel ve dilbilimsel yöntemler gibi anahtar kelime çıkarmanın gerçekleştirilebileceği birçok yaklaşım vardır.

2.1 İstatistik

G. Salton ve arkadaşları 1975 yılında, belgelerdeki metinleri birbirinden, ne kadar iyi ayırt edebildiklerine göre sıralayan bir yöntem olan ayırt edici değer analizi önermiştir. Bu yaklaşımdaki bir terimin değeri, belirli terimlerin içerik tanımlaması için atandığında ortaya çıkan tekil belgeler arasındaki ortalama farklardaki değişime bağlıdır. Fark oranı en büyük olan kelimelerin en iyi kelimeler olması beklenmektedir (Siddiqi ve Sharan, 2015).

1995 yılında J.D. Cohen, indis terimlerini metinden çıkarmak için bir yaklaşım önermiştir. Bu yöntemde n-gram sayılarında yararlanılmakta olunup herhangi bir etkisiz kelime listesi (Stop Words) veya dil ve alana özgü bir bileşen kullanılmamaktadır (Siddiqi ve Sharan, 2015).

2002 yılında Ortuño, bir metnin önemli kelimelerinin birbirlerini çekmeye ve kümeler oluşturmaya eğilimi olduğunu göstermiştir. Bir kelimenin ardışık oluşları arasındaki mesafenin standart sapmasının, kendi kendine çekmeyi ölçmek için böyle bir parametrenin olduğunu iddia eder. (Siddiqi ve Sharan, 2015)

2008 yılında J.P. Herrera ve arkadaşları, sözcüğün uzaysal kullanımlarına atıfta bulunan istatistiki bilgileri kullanarak bir belgenin ilgili kelimelerini bulma ve sıralama problemini ele almıştır. Shannon'ın entropisi, otomatik anahtar kelime çıkarımı için rastgele karıştırılan metin bir standart olarak kullanılmış ve orijinal belgede kullanılan metninde çeşitli ölçümler ile bu metne karşılık gelen ölçümleri normalleştirilmiştir (Siddiqi ve Sharan, 2015).

P. Carpena ve arkadaşları, kuantum düzensiz sistemlerin düzey istatistik analizinin genelleştirilmesi yoluyla edebi metinlerden anahtar kelimeleri otomatik olarak çıkarmayı önermiştir. Bu önerme, metin boyunca uzamsal dağılımları ile birlikte sözcüklerin frekanslarını göz önünde bulundurur ve önemli kelimelerin büyük ölçüde kümelenmiş olduğu, ancak metinde alakasız sözcüklerin rastgele dağıtıldığı gözlemine dayanır. Bu yöntemde referans bir yapıya ihtiyaç yoktur (Siddiqi ve Sharan, 2015).

2.2 Denetimli

Denetimli yaklaşımlar, herhangi bir eğitim dokümanı veya veri kaynağı gerektiren bir yaklaşım türüdür. Denetlenen yöntemlerin amacı, anahtar sözcük çıkarmayı bir ikili sınıflandırma yapısına dönüştürmektir.

Turney ilk olarak, denetlenen bir öğrenme problemi olan anahtar sözcük çıkarımını formülize etmiştir. Bir belgenin tüm ifadelerinin olası anahtar sözcükler olduğunu, ancak yalnızca insan tarafından atanan anahtar sözcüklerle eşleşen ifadelerin doğru anahtar sözcükler olduğunu savunmaktadır. Turney, bu araştırmasında genetik algoritma ve anahtar sözcük çıkarımı için bir dizi parametrik sezgisel kurallar kullanmıştır (Bharti ve Babu, 2017).

Frank ve arkadaşları tarafından geliştirilen KEA sisteminde eğitim belgelerinden Bayes teoremine dayalı bir sınıflandırıcı oluşturulur ve daha sonra yeni belgelerin anahtar sözcüklerini çıkarmak için kullanılır. KEA, giriş dokümanını ortografik sınırlar üzerinde analiz eder ve aday kelimeleri bulmak için noktalama işaretleri, yeni satırlar gibi özelikler kullanır. (Medelyan ve Witten, 2006).

Song ve arkadaşları (2003) yılında, KPSpotter adlı bir sistem önermiştir. Bu sistemde Bilgi Kazanımı (Information Gain), Terimin İlk Oluşumu ve Konuşmanın Bir Parçası gibi birkaç doğal dil işleme tekniği ile birleştirilmiştir. Çıkarımların doğruluğunu artırmak için WordNet KPSpotter' a dahil edilmiştir (Siddiqi ve Sharan, 2015).

Tang ve arkadaşları (2004)'te, anahtar kelime çıkarma için Bayes karar kuramını uygulamış olup kelimeler arası bağlantı bilgilerini kullanmışlardır (Siddiqi ve Sharan, 2015).

Uzun (2005)'te, TF-IDF skoru, metnin başlangıcından itibaren sözcüğün uzaklığı, paragraf ve metindeki anahtar kelimeleri tanımlamak için kullanılan cümle gibi özellikleri kullanarak Naive Bayes sınıflandırıcı kullanmıştır. Araştırmada anahtar kelime özelliklerinin normal olarak dağıtıldığı ve bağımsız olduğu varsayılmıştır (Siddiqi ve Sharan, 2015).

K. Zhang ve diğ. bir anahtar kelime çıkarımını sınıflandırma problemi olarak ele almıştır. Burada bir belgede yer alan kelimeler veya ifadeler üç gruba ayrılmıştır: 'iyi anahtar kelime', 'kayıtsız anahtar kelime' ve 'kötü anahtar kelime'. Anahtar kelime çıkarımı, önceden eğitilmiş bir SVM sınıflandırma modeli ile gerçekleştirilmiştir (Siddiqi ve Sharan, 2015).

Medelyan ve Witten (2006)'de, bir alana özgü eş anlamlılar sözlüğünden çıkarılan terimler ve terimlerle ilgili anlamsal bilgileri kullanarak otomatik anahtar sözcük çıkarımını geliştiren KEA ++'yı gerçeklemişlerdir (Medelyan ve Witten, 2006).

Nguyen ve Kan (2007)'de, bilimsel anahtar sözcüklerde bulunan belirgin morfolojik fenomenleri yakalayan özellikleri kullanarak bilimsel makalelerden anahtar kelime çıkarmayı gerçekleştirmişlerdir.

Zhang ve arkadaşları (2008)'de, anahtar kelimeleri çıkarmak için CRF (Koşullu Rastgele Alan) modelini kullanmıştır. CRF, dizi verilerini bölümlemek ve etiketlemek için yeni bir olasılık modeli ve belirli özellikler kümesiyle koşullu olasılık dağılımını kodlayan, doğrulanmamış bir grafik modelidir (Siddiqi ve Sharan, 2015).

Jiajia Feng ve arkadaşları (2011)'de, kelime dizileri olarak temsil edilen bir belgede uygulanabilir olan sıralı modellere dayanan bir algoritma önermiştir. Bu çalışmada kelimeler arasındaki anlamsal ilişkiyi yansıtan önemli sıralı desenler çıkarılmıştır.

Anahtar kelime çıkarma modeli oluşturmak için kelimelerdeki istatistiksel özelliklerin yanı sıra desen özellikleri kullanılmıştır (Siddiqi ve Sharan, 2015).

2.3 Denetimsiz

Steier ve Belew (1993)'te, iki kelimeli anahtar kelime öbeklerini keşfetmek için karşılıklı bilgi istatistiklerini kullanmıştır (Siddiqi ve Sharan, 2015).

Krulwich ve Burkey (1996)'da, bir belgeden anahtar sözcükleri çıkarmak için sezgisel yöntemler kullanmıştır (Siddiqi ve Sharan, 2015).

Munoz (1996)'da, iki kelimeli anahtar sözcükleri keşfetmek için Uyarlamalı Rezonans Teorisi temelli bir algoritma önermiştir (Siddiqi ve Sharan, 2015).

Barker ve Cornacchia (2000)'de bir belgeden anahtar sözcük olarak isim sözcük gruplarının seçmenin basit bir sistemini önermiştir. Bir isim ifadesi, uzunluğuna ve sıklığına bağlı olarak seçilir (Barker ve Cornacchia, 2000).

Mihalcea ve arkadaşları (2004)'te, metinler arasından çıkarılan grafikler ve kelimeler arasındaki ortak oluşum bağlantılarına dayalı olarak anahtar kelimeleri sıralamak için bir grafik tabanlı sıralama modeli olan TextRank'ı önermiştir. Anahtar sözcükleri çıkarmak için kelimeler arasındaki puanlama kavramı kullanılmıştır (Siddiqi ve Sharan, 2015).

Bracewell ve arkadaşları (2005)'te, bir belgeden isim cümlelerini çıkarmış ve ardından aynı isim terimine sahip terimleri kümelemiştir. Daha sonra terim ve isim cümlelerini frekanslarına göre sıralamıştır. Üst sıradaki kümeler, belgenin anahtar kelimeler olarak seçilmiştir (Siddiqi ve Sharan, 2015).

Liu ve arkadaşları (2009)'da, belgenin anahtar sözcüklerle semantik olarak örtülmesini sağlayan küme teknikleri kullanarak anahtar sözcük öbekleri çıkarmayı önermiştir (Siddiqi ve Sharan, 2015).

Stuart Rose ve arkadaşları (2010)'de, tek tek dokümanlardan anahtar kelimeleri ayıklamak için dilden bağımsız bir yöntem olan Hızlı Otomatik Anahtar Kelime Çıkarımı (RAKE) önermiştir (Rose vd., 2010).

Luit Gazendam ve arkadaşları (2010)'da, sıralama amacıyla bir sözlüğün yardımıyla sınırlı bir kelime dağarcığıyla anahtar kelimelerin çıkarılmasını ve sıralanmasını açıklamaktadır. Sıralama sözcükleri için, hem terim sıklığını hem de belirli belgede bulunan eşanlamlı terimler arasında gerçekleştirilen eş anlamlılar sayısını kullanan TF-RR adlı bir ağırlıklandırma şeması kullanılmıştır. Bu yaklaşım herhangi bir eğitim dokümanına ihtiyaç duymamaktadır (Siddiqi ve Sharan, 2015).

Marina Litvake ve arkadaşları (2011)'de, grafik tabanlı, çapraz dilli bir anahtar sözcük çıkarıcı olan DegExt'i önermiştir. DegExt, belgenin basit grafik tabanlı sözdizimsel gösterimini temel alarak grafik gösterimini kullanılmış ve bazı yapısal belge özelliklerini dikkate alarak geleneksel vektör uzay modelini geliştirmiştir Siddiqi ve Sharan, 2015).

2.4 Yarı Denetimli

Decong Li ve arkadaşları (2010)'da, genel olarak kabul edilen bir belgenin başlığının her zaman belgenin içeriğini yansıtacak şekilde tasarlandığı ve bu nedenle anahtar sözcüklerin doğal olarak başlığa yakın semantik olduğu fikrini kullanan yarı-denetimli bir yaklaşım önermiştir (Siddiqi ve Sharan, 2015).

Anahtar kelime çıkarımı, semantik ağdaki öbek öneminin hesaplanmasıyla gerçekleştirilmiştir ve bu sayede başlık cümlelerinin etkisi diğer cümlelere tekrar tekrar ulaşmıştır.

3. METİN MADENCİLİĞİ

Veri madenciliğinin bir alt dalı olan metin madenciliği; bilgi edinme, veri toplama, makine öğrenimi, istatistik ve hesaplama dilbilimine dayanan bilimler arası branş alanıdır. Bilgilerin önemli bir bölümü; haberler, makaleler, kitaplar, dijital kütüphaneler, e-posta mesajları, bloglar ve web sayfaları gibi metin halinde saklanmaktadır. Bu nedenle, metin madenciliği alanında yapılan araştırmalar önem kazanmaktadır. Bunun en büyük nedenlerinden biri de metinden yüksek kaliteli bilgi elde etmektir. Bu genellikle istatistiksel desen öğrenme, konu modelleme ve istatistiksel dil modellemesi gibi yöntemlerle eğilimleri keşfederek yapılır (Han vd., 2012).

Metin incelemesi, genellikle giriş metninin yapılandırılmasını gerektirir. Bunu, yapılandırılmış verilerdeki kalıpları türetmek ve çıktının değerlendirilmesi ve yorumlanması takip eder. Metin madenciliğinde "yüksek kalite" genellikle uygunluk, yenilik kombinasyonunu ifade eder.

Tipik metin madenciliği görevleri arasında metin sınıflandırma, metin kümeleme, kavram veya varlık çıkarma, taksonomilerin üretimi, duygu analizi, belge özeti ve varlık-ilişki modellemesi bulunmaktadır. Diğer örnekler arasında çok dilli veri madenciliği, çok boyutlu metin analizi, metin verisinde güven analizi, çevrimiçi medya analizi ve analitik müşteri ilişkileri yönetimi gibi metin madenciliği uygulamaları yer almaktadır. Akademik kurumlarda, açık kaynak forumlarında ve endüstride, metin madenciliği için gerekli yazılım ve araçlar mevcuttur. Metin madenciliğinde, genellikle metin verilerinin anlaşılmasını ve metin içeresinden istenilen bilgilerin alınabilmesi için WordNet, Sematic Web, Wikipedia ve diğer bilgi kaynaklarını kullanır (Han vd., 2012).

3.1. Metin Ön İşleme

Veri madenciliği ve Metin madenciliği görevlerinde öncelikli çalışma, dil bilgisel ayırıcılar (",", "/", ";"), bozuk veya gereksiz verilerden temizlemek ve kabul edilebilir bir düzeye getirilmesi gerekmektedir. Özel karakterler ve sembollerin kaldırılması, büyük, küçük harfe duyarlı metnin hariç tutulması gibi metin içerisinde

ön işlemlerden geçer. Bu sayede, doküman, dergi gibi metinler içerisinde çıkarılması beklenen anahtar kelimelerin doğruluğunun artması amaçlanmaktadır. (Guan, 2016) Veri kümesini etkin bir şekilde işlemek için ham verilerin önceden işlenmesi gerekmektedir.

3.1.1 Biçim dönüşümü

Metinler işlenmek için dergi, kağıt, not gibi çeşitli formatlarda gelir ve bu sayede insanlar tarafından okunabilir. Dijital çağda metinler, çeşitli formatlarda elektronik ortamda saklanır. Ancak, bu formatlar farklıdır ve bu formatlara Microsoft Word, PDF (Portable Document Format) gibi farklı yöntemlerle erişilmesi gerekmektedir. Kullanıcılar bu formattaki verileri okuyabilmeleri için çeşitli yazılımlar kullanmaları ve verilerin kullanılacak algoritmalar tarafından İşlenip anlamlandırılabilmesi için uygun formatlara çevrilmesi gerekmektedir.

3.1.2 Tokenleştirme (Tokenization)

Temsili bir bileşen olarak metin, makineler tarafından okunabilmeli ve metindeki kelimeler bir dizi karakter olmalıdır. Bu süreç, cümleleri küçük kelimeler halinde parçalara ayırmayı gerektirir. Bu, tek kelime veya kelime benzeri ifadeler oluşturarak tokenizasyon olarak adlandırılır. Metnin boşluk, — , vb. istenen özelliklere göre parçalara ayrılması olarak örneklendirilebilir (Guan, 2016).

3.1.3 Etkisiz kelime (Stop Words)

Metin çözümlemesi için anlamsız olarak kabul edilen sözcüklere, durma sözcükleri veya etkisiz kelime adı verilir. Etkisiz sözcükler, İngilizce dilinde sözdizimsel yapılardır ve bu kelimeler ilgisiz kelimelerdir. Bunlar bir dizi yüksek frekans kelimesidir ve metin işlemeden önce bu sözcüklerin tipik olarak filtrelenebildiği gereksiz içeriklerdir. Durdurma sözcükleri, genellikle küçük bir sözcük içeriğine sahiptir ve bir metindeki varlıkları onu diğer metinlerden ayırt etmez (Bird vd., 2009). Yüksek frekanslı kelimeler “the”, “is”, “at”, “which”, “on”, “and” ve “to”

kelimeleri içerebilmektedir. Özelliklerine göre, bu zarflar, edatlar, bağlaçlar ve noktalama işaretleri içermektedir (Guan, 2016).

3.1.4 Kök bulma (Stemming)

Kök bulma işlemi, gramer formunun bağlamdaki etkisini kontrol eder. Metinde içeresinde geçen aynı sözcükten türemiş kelimelerin ayrı değerlendirerek tek bir kelimeye indirgemesi işlemi olup dilden dile farklılık göstermektedir. Temelinde Morphologic analysis ve n-gram yöntemlerini kullanmaktadır. Örneklemek gerekirse car, cars, car's, cars' kullanım aslında car ile gösterilmesidir (Stanford NLP Group, 2017).

Metin madenciliğinin bir dalı olan anahtar kelime çıkarımı, algoritmalarının çoğunda metin ön işleme süreçlerinde geçmektedir. Bu sayede metin içerisinden çıkarılmak istenilen kelime veya kelime gruplarının doğruluk oranını yükseltmekte olup algoritmaların metini işlemesinde kolaylık sağlamaktadır.

4. ANAHTAR KELİME ÇIKARIMI

Uluslararası Bilgi ve Kütüphane Bilimi Ansiklopedisi, anahtar kelimeyi bir belgede tartışılan konuyu, özneyi ya da konunun bir yönünü özlü ve doğru bir şekilde tanımlayan bir sözcük olarak tanımlar. Hem tek kelimeler (anahtar kelimeler) hem de sözcük öbekleri (anahtar sözcük öbekleri) anahtar terimler olarak adlandırılabilir (Feather ve Sturges, 1996). Manning ve Schutze, İstatistiki Doğal Dil İşleme Temelleri adlı kitabında yer alan ifadeler hakkında "Sözler, herhangi bir eski düzende gerçekleşmez. Dillerin kelime düzeni üzerinde kısıtlamaları vardır. Ama aynı zamanda bir cümledeki kelimelerin sadece bir kolyenin üzerindeki boncuklar gibi bir dizi konuşma parçası olarak bir araya getirilmemesi de söz konusudur" demiştir (Manning ve Schütze,1999). Bunun yerine, sözcükler, bir birim olarak kümelenen sözcük gruplamaları halinde ifade edilir. Temel bir fikir, belli sözcük gruplarının kurucu olarak davrandıklarıdır.

Anahtar kelimeler, büyük miktardaki çevrimiçi haber verilerinden anahtar kelimelerin belirlenmesi, haber makalelerinin kısa bir özetini oluşturabilmesi açısından çok yararlıdır. Çevrimiçi metinler internetin büyümesiyle birlikte hızla arttıkça, anahtar kelime çıkarımı, arama motoru, metin sınıflandırma, özetleme ve konu algılama gibi birçok metin madenciliği uygulamasının temeli haline gelmiştir.

El ile anahtar kelime çıkarımı, son derece zor ve zaman alıcı bir iştir; Aslında, hacimleri nedeniyle tek bir günde yayınlanan haber makalelerinde anahtar kelimeleri manuel olarak çıkarmak neredeyse imkansızdır. Akademik makalelerde yazar tarafından anahtar kelimeleri belirlenmekte olup içerik ile benzerlik taşımadığı durumlar gözlemlenmiştir. Yapılan bir çalışmada yazar tarafından atanan anahtar kelimelerin %19'unun makaleye dahil olmadığı göstermiştir (Kim vd., 2010). Bu nedenlerden dolayı, metin ve veri madenciliğinin bir dalı olan verilerin otomatik olarak çıkarılmasının önemi giderek önem kazanmakta ve otomatik anahtar kelime çıkarma algoritmaları kullanılabilmektedir.

4.1 Anahtar Kelime Çıkarımı Yaklaşımları

Anahtar kelime çıkarımı, özellikle çevrimiçi dokümanların ve Web, Vikipedi ve WordNet gibi kaynakların giderek yaygınlaşmasıyla, çok çeşitli doğal dil işleme görevlerine uygulanmasından dolayı aktif olarak incelenmiştir. Anahtar kelime çıkarımı yaklaşımları dilbilimsel (linguistic), istatistiksel, grafik tabanlı (graph-based) ve karma yaklaşımlara (hybrid) geniş bir şekilde ayrılabilir.

4.1.1 Dilbilimsel yaklaşım

Bu yaklaşım, metin belgelerinde anahtar kelime algılama ve çıkarma için kelimelerin dil özelliklerini kullanır. Yapılan ilk araştırmalarda, anahtar sözcükleri çıkarmak için sözcük türü, morfolojik, sözdizimsel veya semantik düzeydeki dilsel yapılar kullanılmaktadır.

Dilsel yaklaşımların avantajı, daha az bozuk data çıktısı üretebilmeleridir; dezavantajları ise, kurallar genellikle el ile oluşturulduğu veya belirli bir alan için oluşturulduğu için yeni bir alana kolayca uyarlanamayacaklarıdır. Buna ek olarak, web sayfaları genellikle belirli bir stili takip etmemektedir ve yapılar, iyi yazılmış ve kötü yapılandırılmış olarak geniş bir yelpazeye yayılmıştır ve bunlar dilsel yaklaşımları zorlaştırmaktadır.

Sonuç olarak, dilbilimsel yaklaşımlar kısıtlı alanlarda uygulanabilir ya da hibrid yaklaşımlar gibi diğer yaklaşımlarla birlikte kullanılabilmektedir.

4.1.2 İstatistiksel yaklaşım

Bu yaklaşım basittir ve hiçbir eğitim setine sahip olmadan çıktı üretebilmesi amaçlanmaktadır. Belgenin dil dışı özelliklerinden elde edilen istatistiklere, örneğin, belgenin içindeki bir sözcüğün konumuna, frekans terimine ve ters belge sıklığına odaklanılır ve bu bilgiler ile daha sonra bir anahtar kelime listesi oluşturmak için kullanılır.

İstatistiksel yaklaşımlar üçe ayrılabilir: basit istatistiksel, makine öğrenimi ve kümeleme yöntemleri. Basit istatistiksel yöntemler. terimler bulmak için terim frekansları (Luhn, H. P., 1957), TF-IDF eşdizimlilik veya bağımsızlık gibi n-gramların ölçümlerini kullanırlar. Bu tür yaklaşımların genel uygulamalar için kullanımı kolay ve iyidir. Bununla birlikte, hassasiyet genellikle düşüktür.

Ayrıca, Naive Bayes, Maksimum Entropi Modelleme ve SVM gibi makine öğrenme yaklaşımlarını kullanarak bir eğitim veri setinden alınan ifadeleri otomatik olarak çıkarılabilir (Zhang vd., 2006).

Kümeleme yaklaşımı, Bagging algoritması gibi, otomatik olarak sözcük öbeklerini çıkarmak için çoklu sınıflandırıcıların yapılmasını içerir. Kesinlik performansını artırabilir, ancak çok sınıflandırıcıların eğitiminin tamamlanması uzun zaman alabilir.

4.1.3 Grafik tabanlı yaklaşım

Grafik tabanlı yaklaşımlar, grafikte tekrarlanan toplanan genel bilgilere dayanan bir grafik içindeki kelimelerin önemine karar verir. Bu yaklaşım ilk olarak 2004 yılında Mihalcea ve Tarau tarafından yapılan anahtar kelime çıkarımında uygulanmıştır. Grafik tabanlı bir yaklaşımda, köşeler (veya düğümler) belirteçlerdir, (kelimeler veya öbekler) kenarları temsil etmektedir.

4.1.4 Hibrit yaklaşım

Bu yaklaşım, önceden belirttiğimiz yaklaşımların en iyi özellikleri almak için tasarlanmıştır. Anahtar kelime çıkarma görevinde, kelimelerin konumu, uzunluğu, düzen özelliği, sözcüklerin html etiketleri, vb. gibi bazı sezgisel bilgi kullanırlar.

4.2 Anlamsal (Semantic) Analiz Kullanarak Anahtar Kelime Çıkarımı

Anahtar kelime çıkarımı, metin içerisinde anlamlı kelimelerin belirlenmesinde yaygın olarak kullanılmaktadır. Kümeleme, sınıflandırma veya bilgi edinme gibi

metin süreçlerinde önemlidir ve belgenin temel içeriğini ifade etmelidir. H. Haggag tarafından, semantik benzerlik özelliklerini kullanan bir çıkarma modeli öne sürülmüştür (Haggag, 2013). Algoritma, tek tek kelimeler (kelime-kelime) arasındaki anlamsal ilişkileri değerlendirir ve buna göre genel bir benzerlik (kelimeden bütüne) puanlanır. Anahtar kelime çıkarma işlemi, içeriğe uyum sağlayan anahtar kelimeleri doğru bir şekilde tanımlamak için anlamsal analiz tekrarından geçer. Sonuçlar, geleneksel istatistiksel ve semantik yaklaşımlarla karşılaştırıldığında gelişmiş doğruluk ve geri çağırma sağlar.

4.3 ACM Taksonomi Ve Anahtar Kelime Seçimi

Anahtar kelime seçiminin okuyucular için öneminden önceki bölümlerde bahsedilmiştir. ACM tarafından belirlenen taksonominin yapısı incelenmiştir vee makaleyi belirleyen anahtar kelimelerin ortalama uzunluğunu hesaplanması için kullanılmıştır. Çünkü ACM Hesaplama Sınıflandırma Sistemi, semantik web uygulamalarında kullanılabilecek bir hiyerarşik ontoloji olarak geliştirilmiştir. Bu sayede ACM bir makaleyi sınıflandırabilmek için bu taksonomi yapısını kullanmaktadır (ACM, 2012).

Hardware
- Printed circuit boards
  - Electromagnetic interference and compatibility
  - PCB design and layout
- Communication hardware, interfaces and storage
  - Signal processing systems
    - Digital signal processing
    - Beamforming
    - Noise reduction
  - Sensors and actuators
  - Buses and high-speed links
  - Displays and imagers
  - External storage
  - Networking hardware
  - Printers
  - Sensor applications and deployments
  - Sensor devices and platforms
  - Sound-based input / output
  - Tactile and hand-based interfaces
    - Touch screens
    - Haptic devices

Şekil 3.1 Örnek bir ACM taksonomisi

Şekil 3.1'de Hardware kelimesi 0. derece olarak belirlenir ve alt kırımlılarına inildiğinde derecesi ve konu ile ilgili detayı artmaktadır.

	Ortalama Kelime Uzunluğu	**Toplam İlgili Derece Bulunan**
0.Derce	2	104
1.Derce	3	87
2.Derce	3	828
3.Derce	3	1087
4.Derce	2	334
5.Derce	2	25

Çizelge 3.1 ACM taksonomi çıktıları

Çizelge 3.1'deki sonuç incelendiğinde ACM taksonomi verilerini kullanarak ortalama makale sınıflandırmasının yapılabilmesi için üç kelimeli bir anahtar kelimeye ihtiyaç duyulduğu görülmektedir.

5. OTOMATİK ANAHTAR KELİME ÇIKARMA SİSTEMLERİ

Çoğu belge; analiz, dizin oluşturma ve erişim konularındaki yararlarına rağmen belirlenmiş anahtar kelimelere sahip değildir. Mevcut yaklaşımların çoğunda, anahtar kelimeler, sabit bir taksonomi kullanan profesyonel editörler tarafından el ile atanmakta veya yazarlardan bir temsilci liste sunmaları beklenmektedir. Bu nedenle araştırmaların odak konusunu, profesyonel bir dizin oluşturucuya anahtar kelimeler önermek veya belgelerin başka türlü erişimi mümkün olmayan özelliklerini özetlemek için belgelerden anahtar kelimeleri otomatik olarak çıkarmaya yönelik yöntemler oluşturmuştur.

5.1 TF-IDF

TF-IDF en çok bilinen ve en yaygın kullanılan terim çıkarma algoritmalarından biridir (Guan, J., 2016).

TF, belirli bir terimin belirli bir belgede görüntülenme sayısı olan terim frekansını ifade eder. Tek bir belge veya çoklu belgeler içinde bir terimin oluşumlarını tanımlayabilir. İstatistiksel olarak, frekans ne kadar yüksek olursa, potansiyel olarak önemli bir terim olduğu varsayılmaktadır (Adji vd., 2014). Aşağıdaki formüle göre hesaplanır:

$$\boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) = \frac{\boldsymbol{f}(\boldsymbol{t})}{\boldsymbol{n}} \tag{5.1}$$

Formül açıklanacak olursa;

- $\boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d})$ terim sıkılığını (frekansı) ifade eder.
- $\boldsymbol{n}$, verilen belgedeki terimlerin toplam sayısıdır.
- $\boldsymbol{f}(\boldsymbol{t})$**,** terim sıklığı (frekansı).
- $\boldsymbol{d}$**,** doküman.

Bununla beraber daha sık olan ancak anlamsız kelimler mevcuttur: “and”, ”the”, ”or”, ”only” gibi. Bunlar dokümanlarda çok olan ve okuyucu tarafından tek başına çok anlam ifade etmeyen bağlaç veya kelimelerdir. Genelde veri ön işleme sırasında

bu tip kelimeler kaldırılır. Aksi takdirde yüksek frekanslı çıkan bu kelimeler söz konusu olan belgede asıl belirtilmek istenilen bilgiyi veremeyecektir (anahtar kelime veya cümle) olacaktır. Bu sorunu çözmek için, araştırmacılar terimlerin ağırlığını azaltmak için ters belge frekansı (IDF) kullanırlar. IDF, belgelerindeki yüksek sıklıkta önemsiz ifadeleri dengelemenin bir yoludur (Jing vd., 2002).

Liu ve arkadaşları (2008)'de, düşük IDF'ye sahip bir kelimenin birçok belgede gerçekleştiğini ve konu olarak belirteç olmadığını ifade etmiştir. Hussey ve arkadaşları, verilen belgede yüksek bir TF'ye sahip olması ve korpustaki kalan belgelerin düşük oluşu nedeniyle yüksek ağırlığın önemli olduğunu gösterdiğini vurgulanmıştır. Bu nedenle IDF, kalan dokümanlarda yüksek frekanslı terimler için azaltılmış ağırlığın bir ölçümüdür (Adji vd., 2014). IDF formülü şöyledir:

$$\boldsymbol{idf(t) = \log \frac{|D|}{|\{d : t \in d\}s|}} \qquad (5.2)$$

Formül açıklanacak olursa;

- ***D,*** korpus'taki tüm belge sayısıdır.
- ***d, t*** terimini içeren belgelerin miktarıdır.
- ***idf(t), t*** terimi ile ilgili ters belge frekansıdır.

TF-IDF ağırlığı, bir terimin belgede önemli olup olmadığını ve korpusun nadir olup olmadığını belirlemek için kullanılan istatistiksel bir ölçümdür. İlk kez 1972'de yayınlanmıştır ve birçok yeni anahtar kelime çıkarımı uygulaması hala teoriye dayanmaktadır (Guan, 2016). Formül;

$$\boldsymbol{tfidf(t,d) = tf(t) * idf(t)} \qquad (5.3)$$

TF-IDF ağırlığı, bir kelimenin bir koleksiyondaki bir belgeye verdiği önemi değerlendirir ve daha yüksek TF-IDF puanları olan kelime, belgede önemlidir ve belgeyi özetleyebilir.

Örnek olarak aşağıdaki cümleleri incelersek;

- Today, car have 2 traction and 4 traction models. But the 2 traction models most popular models are on sale.
- I have 2 traction car
- car prices are increasing nowadays.

İlk cümle için **traction** hesaplanırsa;

- TF = 3/20 = 0,15
- IDF = log(3/2) = 0,176
- TF-IDF = 0,0264 bulunmaktadır.

Dokümanlar	**1.Anahtar Kelime**	**2.Anahtar Kelime**	**3.Anahtar Kelime**
Today, car have 2 traction and 4 traction models. But the 2 traction models most popular models are on sale	**"models",** 0.07157	**"traction",** 0.02641	**"and",** 0.02386
I have 2 traction car	**"I"**, 0.09542	**"traction"**, 0.03522	**"2"**, 0.03522
car prices are increasing nowadays	**"prices"**, 0.09542	**"increasing"**, 0.09542	**"nowadays"**, 0.09542

Çizelge 5.1 Örnek cümlelerin anahtar kelime sonuçları

Örnek cümlelerde çıkarılan anahtar kelimeler Çizelge 5.1 gösterilmiştir. Çıkarılan anahtar kelimeler büyükten küçüğe göre sıralanmış olup en yüksek puana sahip ilk 3 anahtar kelime gösterilmiştir. Çizelge 5.1 detaylı incelenirse, en başta bahsedilen stop words veya anlamsız kelimeler diye tabir edilen bağlaç, sayı veya içeriği doğru tanımlanamayan kelimeler gibi anahtar kelimeler olarak seçildiği görülmektedir. Bu da okuyucular tarafından istenilen bir sonuç olmayacaktır.

TF-IDF'nin dezavantajı, belirli bir sözcüğün sıklığını, anlamsal anlam bakımından benzer herhangi bir sözcük göz önüne almadan, saymasıdır. Ek olarak, belge kısa olduğu zaman, TF-IDF kelimenin önemini istenildiği gibi karşılamamaktadır (Guan, 2016).

5.2 RAKE

RAKE'in geliştirilmesindeki amaç, tekil belgeler üzerinde çalışan ve özellikle belirli bir dilbilgisi kurallarına uymasına gerek olmayan, birden fazla belge türü üzerinde iyi çalışan, verimli bir anahtar kelime çıkarım yöntemi geliştirilmesi olmuştur (Rose vd., 2010).

RAKE; anahtar kelimelerin sıklıkla birden fazla kelime içerdiğini, ancak standart noktalama işaretleri veya dar anlama sahip "ve, ile" gibi etkisiz sözcükleri nadiren içerdiğine ilişkin gözlemine dayanır.

RAKE, bir dokümandaki metnini kelime dizileri şeklinde aday anahtar kelimelere ayırmak için etkisiz kelimeler ve tümcecik sınırlayıcıları kullanır. Bu aday anahtar kelimeler içinde yer alan kelimelerin eşdizimi anlamlıdır ve gelişigüzel boyutta kayan pencere (sliding window) uygulamadan kelime eşdizimini tanımlanmasını sağlar. Böylece, kelime ilişkileri, metnin stiline ve içeriğine otomatik olarak uyum sağlayacak şekilde ölçülür ve aday anahtar kelimeleri belirlemek için kullanılacak kelime eşdizimlerinin uyarlanabilir ve ayrıntılı ölçümünü sağlar (Rose vd., 2010).

RAKE, bir belgedeki metni, bir dizi aday anahtar kelimeye ayrıştırarak anahtar kelime çıkarımına başlar. İlk olarak, belge metni, belirtilen kelime sınırlayıcılarla bir kelime sırasına bölünür. Ardından bu sıra, tümcecik sınırlandırıcı ve etkisiz kelimelerin bulunduğu yerlerden, bitişik kelime dizilerine ayrılır. Bir dizi içindeki kelimeler, metinde aynı yere atanır ve birlikte bir aday anahtar kelime olarak kabul edilir.

Keyword extraction is tasked with the automatic identification of terms that best describe the subject of a document.

Key phrases, key terms, key segments or just keywords are the terminology which is used for defining the terms that represent the most relevant information contained in the document. Although the terminology is different, function is the same: characterization of the topic discussed in a document. The task of keyword extraction is an important problem in Text Mining, Information Retrieval and Natural Language Processing

Şekil 5.1 Anahtar kelime çıkarımının tanımı

Şekil 5.1'de örnek olarak ele alınan belgede anahtar kelime çıkarımını Wikipedia tarafından tanımı yapılmıştır. Bu belge üzerinden adaya anahtar kelimeler belirlenecek olursak Şekil 5.2'deki gibi bir sonuç çıkacaktır.

['keyword extraction', 'tasked', 'automatic identification', 'terms', 'describe', 'subject', 'document', 'key phrases', 'key terms', 'key segments', 'keywords', 'terminology', 'defining', 'terms', 'represent', 'relevant information contained', 'document', 'terminology', 'function', 'characterization', 'topic discussed', 'document', 'task', 'keyword extraction', 'important problem', 'text mining', 'information retrieval', 'natural language processing']

Şekil 5.2 Örnek tanımdan çıkarılan aday anahtar kelimeler

Şekil 5.2'de görüldüğü gibi etkisiz kelimler (Stop Words) ve noktalama işaretleri göz önüne alınarak belirlenmiştir. Örnek olarak **'function'** kelimesi Şekil 5.1'de bakıldığında virgül ile başlayıp etkisiz kelime ile bitmektedir ve sonuç olarak Şekil 5,2'de aday anahtar kelime olarak belirlenmektedir.

RAKE, sonuçlarını elde etmek için doğal dil işleme tekniklerine dayanan yöntemlerin aksine, basit bir girdi parametresi kümesi alır, anahtar kelimeleri tek bir geçişle otomatik olarak çıkarır ve böylece geniş bir doküman ve koleksiyon

yelpazesine uygun hale getirir. Son olarak, RAKE' in basitliği ve verimliliği, anahtar kelimelerin kullanıldığı birçok uygulamada kullanımını mümkün kılar.

Her aday anahtar kelime tanımlandıktan sonra belirlenen aday anahtar kelimler için puanları (skor) hesaplanır. Kelime skorlarını hesaplamak frekans ve derece değerleri bulunur. (1) kelime frekansı(freq(w)), (2) kelime derecesi (deg(w)) ve (3) derecenin frekansa oranı (deg(w)/freq(w)) (Rose vd., 2010).

Freq: Frekansın hesaplanması, ilgili aday kelimenin çıkarılan diğer aday anahtar kelimeler listesinde kaç kez geçtiğidir.
Deg: Bir kelimenin derecesi, aday anahtar kelimelerdeki diğer kelimelerin birlikte oluşma sıklığını temsil eder.

Kelime	Frekans (Freq)	Derece (Deg)	Deg(w)\Freq(w)
represent	1.0	1.0	1.0
relevant	1.0	3.0	3.0
characterization	1.0	1.0	1.0
text	1.0	2.0	2.0
describe	1.0	1.0	1.0
topic	1.0	2.0	2.0
keywords	1.0	1.0	1.0
subject	1.0	1.0	1.0
information	2.0	5.0	2.5
segments	1.0	2.0	2.0
extraction	2.0	4.0	2.0
contained	1.0	3.0	3.0
identification	1.0	2.0	2.0
document	3.0	3.0	1.0
defining	1.0	1.0	1.0
function	1.0	1.0	1.0
mining	1.0	2.0	2.0
terms	3.0	4.0	1.333
processing	1.0	3.0	3.0
tasked	1.0	1.0	1.0
important	1.0	2.0	2.0
key	3.0	6.0	2.0
task	1.0	1.0	1.0
natural	1.0	3.0	3.0
language	1.0	3.0	3.0
keyword	2.0	4.0	2.0
phrases	1.0	2.0	2.0
discussed	1.0	2.0	2.0
problem	1.0	2.0	2.0

automatic	1.0	2.0	2.0
retrieval	1.0	2.0	2.0
terminology	2.0	2.0	1.0

Çizelge 5.2 Örnek metin içinde geçen her bir kelimenin çıktısı

Çizelge 5.2' te tek tek kelime puanlarını hesaplamak için deg(w)/freq(w) metriklerini kullanarak örnek özetteki her aday anahtar kelimeyi listelenmektedir.

Aday Anahtar Kelime	**Puanı**
keyword extraction	4.0
tasked	1.0
automatic identification	4.0
terms	1.33333
describe	1.0
subject	1.0
document	1.0
key phrases	4.0
key terms	3.333
key segments	4.0
keywords	1.0
terminology	1.0
defining	1.0
represent	1.0
relevant information contained	8.5
document	1.0
terminology	1.0
function	1.0
characterization	1.0
topic discussed	4.0
task	1.0
keyword extraction	4.0
important problem	4.0
text mining	4.0
information retrieval	4.5
natural language processing	9.0

Çizelge 5.3 Örnek metinden çıkarılan anahtar kelime skorları

Örnek özetteki her içerikli kelimeye ait metrik skorlar Çizelge 5.3'te listelenmiştir. Özet olarak, *deg(w), daha sık* ve daha uzun aday anahtar kelimelerde ortaya çıkan kelimeleri desteklemektedir. Daha uzun aday anahtar kelimelerde baskın olarak bulunan kelimeler, *deg(w)/freq(w)* tarafından tercih edilir.

*deg(*keyword*)/freq(*keyword) skorları *deg(*extraction*)/freq(*extraction*)* skorlarından daha yüksektir. Her aday anahtar kelimenin skoru, bu kümeye ait kelime skorlarının toplamı olarak hesaplanmaktadır.

Örnek metinde çıkarılan puanlar büyükten küçüğe sıralanırsa **('natural language processing', 9.0), ('relevant information contained', 8.5), ('information retrieval', 4.5), ('keyword extraction', 4.0)** şeklinde sıralanmakta olup en **'natural language processing** en yüksek skora sahip olan aday anahtar kelime belirlenmektedir.

Mevcut koleksiyonların çeşitliliği ve hacmi ile belgelerin oluşturulma ve toplanma oranına bağlı olarak, RAKE, diğer analitik yöntemler için avantajlar sağlar ve bilgi işlem kaynaklarını kısıtlamaz.

6. BRAKE

BRAKE algoritmasının geliştirilmesindeki en büyük amaç makale, dergide gibi yazılan metin içerisinde geçen terimlerin, tekrarlamalar ve benzerliklerin oluşmaması için yazar tarafından eş anlamlı kelimeler kullandığı görülmüştür. Bu da incelenen denetimsiz yaklaşıma uygun algoritmalarda (RAKE, TF-IDF vs.) aynı anlama gelen kelime veya cümleleri ayrı ayrı anahtar kelime seçmesine sebep olmakta olup, metinde geçen asıl çıkartılmak istenilen anahtar kelimeyi çıkaramamış olmaktadır. Bu tür sebeplerden dolayı BRAKE algoritması geliştirilmiştir.

BRAKE algoritmasındaki amaç, denetimsiz yaklaşıma uygun olup, herhangi bir eğitim dokümanına bağlı kalınmadan, dilden bağımsız tekil dokümanlarda çalışan bir anahtar kelime çıkarımı algoritması olmuştur. BRAKE algoritması bu doğrultuda RAKE algoritmasına temelinden çıkarılmış olmaktadır.

BRAKE algoritmasını, RAKE algoritmasından ayıran en büyük fark, eş anlamlı kelimeleri tek bir kelime üzerinde toplayarak entropi değeri düşürülmesi amaçlanarak geliştirilmiştir. Bunun için thesaurus dokümanlarından yararlanılmış olunup denetimsiz yaklaşıma uygun bir algoritma çıkarıldı. Bu sayede bir doküman içinde geçen anlamdaş kelimelerin tek bir çatı alıntında toplanarak entropi değeri düşürülmüş olur.

6.1 Algoritma

BRAKE algoritmasını Şekil 6.1 metin üzerinden adım adım anlatılmak istenilirse;

There are two types of car engine. One is petrol and the other one is electric car engine. The one costs more is electric automobile engine

Şekil 6.1 Örnek Metin

1. Aday anahtar kelimelerin belirlenmesi için metnin içindeki kelimeler parçalanarak bir listede toplanır. **[there, are, two, types, of, car, engine, one,**

is, petrol, and, the, other, one, is, electric, car, engine, the, one, costs, more, is, electric, automobile, engine]

2. İlgili liste içindeki önceden hazırlanmış olan stopword liste yardımı ile bunlar are, is, the gibi bağlaç veya kelimler liste içinden çıkarılır. **[types ,car, engine, petrol, electric, car, engine, costs, electric, automobile, engine]**
3. Oluşabilecek aday anahtar kelimeler belirlenir, soldan sağa doğru tekrardan okunarak liste tekrardan belirlenir. **[types, car engine, petrol, electric car engine, costs, electric automobile engine]**
4. Aday anahtar kelime listesindeki her bir kelimenin puanı hesaplanır. Bu hesaplama metrik kullanarak gerçekleşir (**degree(word)/frequency(word)).**
5. Frekansın hesaplanması, ilgili aday kelimenin çıkarılan diğer aday anahtar kelimeler listesinde kaç kez geçtiğidir. Örnek: **Freq(car) = 2, Freq(engine) = 3, Freq(costs) = 1**
6. Frekansı hesaplandıktan sonra derecesi bulunarak, kelimenin puanı çıkarılır.
7. Bir kelimenin derecesi, aday anahtar kelimelerdeki diğer kelimelerin birlikte oluşma sıklığını temsil eder. Örnek: **Deg(automobile) = 3, Deg (engine) = 8, Deg(costs) = 1**
8. word_score = **degree(word)/frequency(word)** hesabı yapılır.
9. Eş anlamlı bir thesaurus listin bulunduğu kütüphaneden (db) den okunarak aday anahtar kelimelerin eş anlamları bulunur.
10. Çıkarılan eş anlamlı kelimelerin diğer aday anahtar kelimelerin içinde geçip geçmediği kontrol edilir eğer geçiyor ise işleme alınır geçmiyorsa yok sayılır.
11. İşleme alınan eş anlamlı kelimeler anlamdaş olduğu kelime ile kendisinin çıkarılan kelime puanına bakılır. Bu sayede dokümanda hangi anlamdaş kelime puanı yüksek ise hepsinin tek anlamda toplanacağı kelime aynı olur.
12. Belirlenen kelime mevcuttaki anlamdaşın yerine geçer ve oluşan yeni metnin frekansı ve derecesi hesaplanarak puan bulunur. **[types , automobile, engine, ,petrol, electric, automobile, engine, costs, electric, automobile, engine]**
13. Son olarak hesaplanan puan: **word_score("automobile ") = 12/4 = 3, word_score("engine ") = 12/4 = 3**
14. Bütün aday anahtarların puanları hesaplanır ve büyükten küçüğe sıralanarak ağırlığı yüksek olan aday kelime anahtar kelime olarak belirlenir.

Kelime	Frekans (Freq)	Derece (Deg)	Deg(w)\Freq(w)
Electric	2.0	8.0	4.0
Engine	3.0	12.0	4.0
Petrol	1.0	1.0	1.0
Automobile	3.0	12.0	4.0
Costs	1.0	1.0	1.0
Types	1.0	1.0	1.0

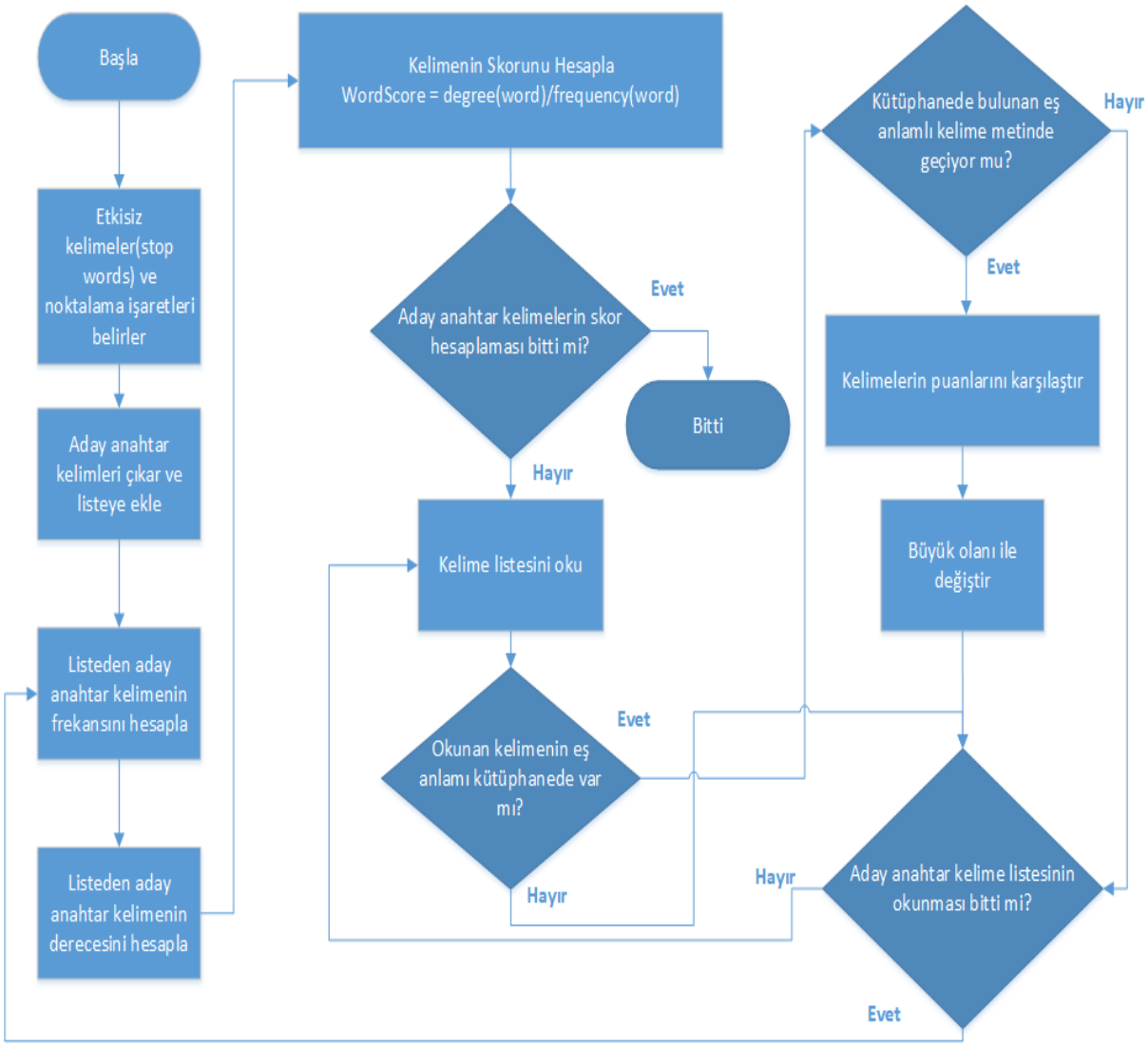
Çizelge 6.1 Örnek metin içinde geçen her bir kelimenin çıktısı

Aday Anahtar Kelime	Puan
Electric automobile engine	12
Automobile engine	8
Petrol	1
Costs	1
Types	1

Çizelge 6.2 Örnek metinden çıkarılan anahtar kelime skorları

Çizelge 6.1'te tek tek kelime puanlarını hesaplamak için deg(w)/freq(w) metriklerini kullanarak örnek özetteki her aday anahtar kelimeyi listelenmektedir. Örnek özetteki her içerikli kelimeye ait metrik skorlar Çizelge 6.2'te listelenmiştir.

BRAKE algoritmasının doğruluğu ve başarısı kullanılan eş anlamlılar (thesaurus list) listesine bağlıdır.



Şekil 6.2 BRAKE algoritmasına ait akış diyagramı

Şekil 6.2 görüldüğü üzere BRAKE algoritmasının çalışma prensibi anlatılmıştır.

6.1.1 Eş anlamlı kelimelerin belirlenmesi

BRAKE algoritmasında etkisiz kelimeler çıkarılıp aday anahtar kelimeler bulunmadan önce çıkarılan kelimelerin eş anlamı olup olmadığına bakılır. Eğer eş anlamı var ise, mevcut metin içerisinde eş anlamlarının geçip geçmediği kontrol edilir. Kontrol sonucu kelimenin metin içerisinde geçmesi durumunda ilk önce puanı bulunur ve bu sayede hangi eş anlamlı kelimenin tek çatı altında toplanacağı belirlenir.

Anahtar Kelime	Puan
Engine	2.6
Electric	3.0
Petrol	1.0
Automobile	3.0
Costs	1.0
Car	2.5
Types	1.0

Çizelge 6.3 Cümlede kelimelerin eş anlamları bulunmadan önce puanları

Şekil 6.1'deki örnek cümle ele alınarak gösterilirse, eş anlamları ile değerlendirilmeden önce hesaplanan puanlar Çizelge 6.3'de gösterilmiştir. Puanlama deg(w)/freq(w) hesabına göre yapılmaktadır.

Cümle İçinde Geçen	Cümle İçinde Geçen Kelime Karşılığı
engine	engine
electric	electric
car	**automobile**
petrol	petrol
automobile	automobile
costs	costs
types	types

Çizelge 6.4 Cümledeki eş anlamlı kelime ve değişimi

{auto, automobile, bucket, buggy, bus, clunker, compact, convertible, conveyance, coupe, gas guzzler, hardtop, hatchback, heap, jalopy, jeep, junker, limousine, machine, motor, motorcar, pickup, ride, roadster, sedan, station wagon, subcompact, touring car, truck, van, wagon, wheels, wreck}

Şekil 6.3 Car kelimesinin eş anlamları

Şekil 6.1'de araba motorları ile ilgili cümlede **'Car'** kelimesinin eş anlamı olduğu ve cümle içerisinde geçtiği belirlendikten sonra bu kelimenin **'Automobile'** olduğu Şekil 6.3'de gösterildiği gibi eş anlamlı listeden bakılarak çıkarılmıştır. Çizelge 6.3 çıkarılan puanlara göre karşılaştırılırsa '**Car' 2.5**, '**Automobile' 3.0** olduğu gözükmekte olup geliştirilen BRAKE algoritması bütün '**Car'** gözüken kelimeler '**Automobile'** çevrilmiştir. Çizelge 6.4 de bu değişim gösterilmektedir.

6.2 RAKE Ve BRAKE Algoritmasının Karşılaştırılması

Şekil 6.1 ve Şekil 5.1 kullanılan metinlerin üzerinden bu iki algoritmayı karşılaştıracak olursak;

SIRA	RAKE	BRAKE
1	electric automobile engine	electric automobile engine
2	**electric car engine**	**automobile engine**
3	car engine	petrol
4	petrol	costs
5	costs	types

Çizelge 6.5 Araba motorları ile ilgili metinden çıkarılanlar

SIRA	RAKE	BRAKE
1	natural language processing	relevant information contained
2	relevant information contained	natural language processing
3	**information retrieval**	**keyword extraction**
4	keyword extraction	information retrieval
5	important problem	keyword segments

Çizelge 6.6 Anahtar kelimenin tanımından çıkarılanlar

Cümle İçinde Geçen	Cümle İçinde Geçen Kelime Karşılığı
represent	**defining**
relevant	relevant
characterization	characterization
text	**topic**
describe	defining
topic	topic
keywords	keyword
subject	**topic**
information	information
segments	segments
extraction	extraction
contained	contained
identification	identification
document	**language**
defining	defining
function	**task**
mining	mining
terms	terms

processing	processing
tasked	tasked
important	**relevant**
key	**keyword**
task	task
natural	natural
keyword	keyword
language	language
phrases	phrases
discussed	discussed
problem	**topic**
automatic	automatic
retrieval	retrieval
terminology	**language**

Çizelge 6.7 Anahtar kelimenin tanımına ait eş anlamlı kelime ve değişimi

Şekil 6.1 deki cümlede anlatılmak istenilen araba motorları ve çeşitliği üzerine olup, RAKE ve BRAKE algoritmasının çıktıları görülmektedir. Tablo 6.5'te anahtar kelimelerin puanlarına göre büyükten küçüğe göre sırlanmışlardır. Tablo 6.5'de 3. sırada algoritmaların çıkardığı anahtar kelimelere bakıldığı zaman RAKE algoritmasının **electric car engine**, BRAKE algoritmasının da **automobile engine** anahtar kelimelerini çıkarıldığı görülür. Cümlede asıl anlatılmak istenilenin araba motorları olduğu düşünülürse BRAKE algoritmasının 2. sıradaki puanına göre daha başarılı olduğu söylenebilir. Bunun haricinde RAKE algoritmasın 1. ve 2. sıradaki anahtar kelimelere bakılırsa anlamca aynı kelimelerden oluştuğu görülür. Şekil 5.1'de anahtar kelime çıkarımı ile ilgili tanımını her iki algoritmada karşılaştırılıp 3. sıraya bakıldığında; RAKE algoritmasının **information retrieval**, BRAKE algoritmasının da **keyword extraction** anahtar kelimelerini çıkardığını görülmektedir.

BRAKE algoritması anahtar kelimeleri çıkarmadan önceki eş anlamlı ön işlemeye ait Çizelge 6.7 gösterilmektedir.

6.3 Algoritmanın Geliştirme Ortamı

Algoritma geliştirme ortam PyCharm IDE'si kullanılmış olup Python 2.7 ile geliştirilmiştir. Uygulama içinde eş anlamlılar SQLite'da tutulmuştur. Python,

dinamik semantik ile yorumlanmış, nesne yönelimli, yüksek seviyeli bir programlama dilidir. Girintilere dayalı basit sözdizimi, dilin öğrenilmesini ve akılda kalmasını kolaylaştırır. Bu da söz diziminin ayrıntıları ile zaman harcamadan geliştirmeye hızlıca başlanabilen bir dil olma özelliği kazandırır (Python, 2018).

SQLite, dünyada en çok dağıtılan ve tavsiye edilen açık kaynak kodu, tamamen C/C++ programlama dilleriyle geliştirilmiş sunucu yazılımı ve yapılandırma gereksinimi olmayan, işlemsel ve ilişkisel bir SQL veritabanı motorudur (SQLite, 2018).

Akademik makalelerin çoğu dijital ortamlarda bulunmakta olup PDF formatında barınmaktadır. Makaledeki metinlere erişilebilmesi ve algoritmalar tarafından okunabilmesi için TXT formatında olmasına ihtiyaç duyulmaktadır. Bu nedenlerden dolayı pdf2text ile PDF formatındaki yayınları TXT formatına çevrilmiş olup metin ön işleme sürecinden geçmiştir.

7. ÖLÇME YÖNTEMLERİ VE SONUÇ

Geliştirilen algoritmaların karşılaştırılarak başarı oranlarının değerlendirilmesi gerekmektedir. Çıkarılan anahtar kelimler ne kadar doğru, orijinal anahtar kelimelere benzerliği nedir gibi kullanıcı tarafında beklenen cevapların karşılanması gerekmektedir.

İki genel değerlendirme çerçevesi vardır. İlk yöntem insan açıklama veya insan hakemi tarafından uygulanır. Belirli bir bilgi birikimine özel bilgi gerektiren alan uzmanlığı veya uzmanlarına ihtiyaç duyar. İkinci yöntem, zamandan tasarruf edebilen ve daha az maliyetli olan otomatik bir değerlendirmedir. Araştırma projemiz, tüm deneylerin performanslarını değerlendirmek için ikinci yöntemi kullanmaktadır.

Sistemlerin çoğunluğu, elde ettikleri sonuçları "altın standart" ile karşılaştırır. 566 adet akademik makaledeki orijinal anahtar kelimler altın standart olarak kabul edilip her bir algoritmanın çıktısı ile karşılaştırılmıştır. Yani yapılan değerlendirmede yazar tarafından atanan anahtar kelimelerin doğru olduğunu varsayım yapılarak değerlendirme yapılmaktadır. Kullanılan model karışıklık matrisi (Confusion Matrix) ile ifade edilmiştir. (Turney, 2002)

	İnsanlar tarafından atanan anahtar kelimeler	İnsan tarafından atanan anahtar kelime olmayan
Sistem tarafından çıkarılan ilgili anahtar kelimeler	a	b
Sistem tarafından çıkarılan ilgisiz anahtar kelimeler	c	d

Çizelge 7.1 Karışıklık matrisi

Değerlendirmede, anahtar kelimelerin manuel olarak atanmasında, insanlar tarafından atanan anahtar kelimeler ve anahtar kelimeler olmayan iki tür kelime veya kelime öbeği vardır.

Çizelge 7.1 parametreler açıklanırsa;

- **a**, TP (True Pozitif)

- **b**, FP (False Pozitif)
- **c**, FN (False Negatif)
- **d**, TN (True Negatif)

Gerçek Pozitifler (TP) - Bunlar doğru tahmin edilen pozitif değerlerdir; gerçek sınıfın değerinin evet olduğu ve tahmin edilen sınıfın değerinin de evet olduğu anlamına gelir.

Gerçek Negatifler (TN) - Bunlar doğru tahmin edilen negatif değerlerdir; gerçek sınıfın değerinin hayır olduğu ve öngörülen sınıfın değerinin de hayır olduğu anlamına gelir

Yanlış Pozitifler (FP) - Gerçek sınıfın yok olduğu ve tahmin edilen sınıfın evet olduğu zaman.

Yanlış Negatifler (FN) - Gerçek sınıf evet olduğunda ama hayır olarak tahmin edilen sınıf.

Bilgi alma değerlendirmesinde kullanılan genel ölçüm yöntemi olan, hassasiyet veya geri çağrım (R), kesinlik (P) ve F1-Measure'a göre değerlendirmeler gerçekleştirilmiştir. Bu yöntem aynı zamanda arama motorunun performansını değerlendirilmesi içinde kullanılmaktadır.

7.1 Kesinlik (Precision) Ve Hatırlama (Recall)

Hassasiyet, doğru tahmin edilen pozitif gözlemlerin toplam tahmini pozitif gözlemlere oranıdır. Geri Çağırma, doğru tahmin edilen pozitif gözlemlerin gerçek sınıftaki tüm gözlemlere oranıdır. Basit bir ifadeyle kesinlik (Precision), bir algoritmanın önemsiz olanlardan önemli ölçüde daha alakalı sonuçlar verdiğini gösterirken, hatırlama (Recall), bir algoritmanın ilgili sonuçların çoğunu döndürdüğü anlamına gelir (Derczynski, 2016).

Kesinlik (Precision) ve Hatırlama (Recall) şu şekilde ifade edilir;

$$\textbf{percision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \quad (7.1)$$

$$\textbf{recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \quad (7.2)$$

Kesinlik (Precision) ve hatırlama (Recall) birbirlerini aralarında ters bir ilişki vardır. Örneğin, yapılan ölçümde daha yüksek bir hassasiyet elde ettiğinde, ilgili hatırlama (Recall) daha düşük olacaktır (Turney, 2002).

7.2 F-Ölçüsü (F1-Metrik)

Bilgi erişim sistemleri (IR) gibi uygulamalarda, genellikle hassasiyet ve geri çağırma hesaplamaları ile işlem yapılır. Ancak bazen, NLP uygulamalarında çok fazla anlam ifade etmediği durumlar gözlenmektedir. Bu nedenle, kesinliği ve hatırlamayı genel performansın tek bir ölçüsü olarak birleştirmek uygun olmaktadır. Bunu yapmanın bir yolu, Van Rijsbergen' in ortaya koyduğu bir varyant olan F ölçümüdür (Manning ve Schütze, 1999).

F ölçüsü şu şekilde tanımlanır;

$$\text{F-Ölçütü} = \frac{2 \cdot P \cdot R}{P+R} \quad (7.3)$$

- **P**, Kesinlik (Precision)
- **R**, Hatırlama veya Hassasiyet (Recall)

Şimdi F ölçüsünü ortalama olarak hesaplanırsa;

$$\text{F-Ölçütü} = \frac{2 * avgPrecision * avgRecall}{avgPrecision + avgRecall} \quad (7.4)$$

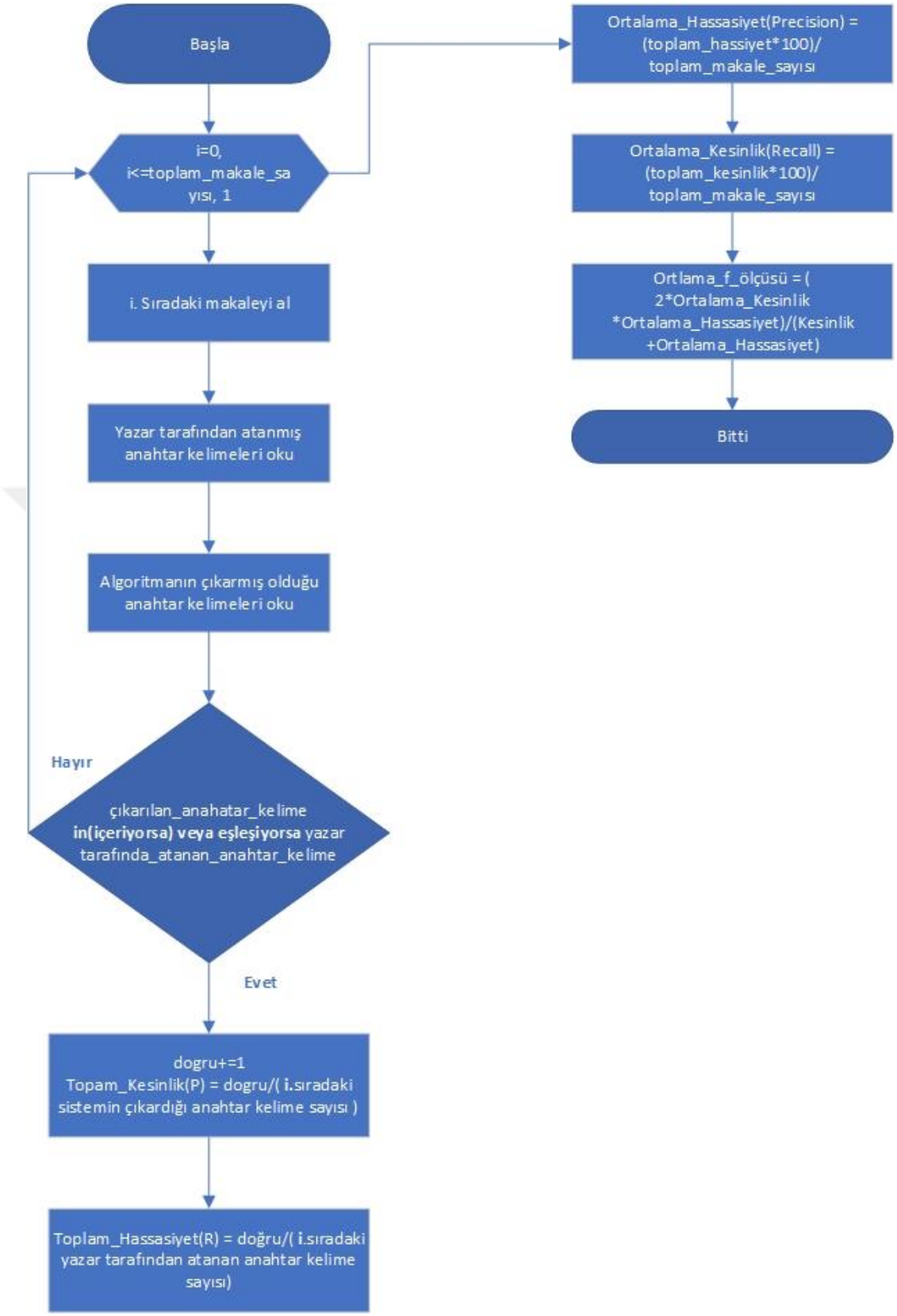
- **avgPrecision**, Ortalama Kesinlik (Precision) sayısı
- **avgRecall**, Ortalama Hatırlama veya Hassasiyet (Recall) sayısı

Algoritmaların karşılaştırılmasında, F ölçüsü kullanılarak anahtar sözcük çıkarma algoritmaların performansını değerlendirilmiştir.

7.3 Algoritmaların Değerlendirilmesi

TF-IDF, RAKE ve BRAKE algoritmaları, Kesinlik (Precision), Hassasiyet (Recall) ve F ölçüsü değerleri üzerinden karşılaştırılmasında 566 adet İngilizce akademik makaleler kullanılmıştır. Bu makaleler çeşitli mühendislik alanlarında hakemli olarak yayın yapan dergi ve konferanslardan PDF (Portable Document Format) formatında alınmıştır.

Makaleler pdf formatından pdf2text aracı ile ilk olarak metin biçimine dönüştürülmüştür. Ön işlemden geçen metinlerden, yazarların belirlediği anahtar kelimeler değerlendirme ölçütünde kullanılmak üzere Python programlama dili kullanılarak otomatik olarak metinlerden çıkarılmıştır. Yine Python programlama dili kullanılarak her bir algoritma gerçeklemiştir. Daha sonra, tüm makaleler bu algoritmalar ile işlenip bulunan kelimeler büyükten küçüğe puan sıralaması ile dizilmiştir. Yazar tarafından atanan anahtar kelimelerin sayısı beşten fazla ise ilk beş kelime, orijinal anahtar kelime sayısının beşten az olduğu durumlarda ise orijinal anahtar kelime sayısı kadar sözcük, makaleyi temsil eden anahtar kelime olarak belirlenmiştir.



Şekil 7.1 Otomatik değerlendirme sistemine ait akış diyagramı

Tüm işlemler yapıldıktan yazar tarafından belirlenmiş ve doğru olduğunu varsayılan (altın standart) anahtar kelimeler ile algoritmaların çıkarmış olduğu anahtar kelimeler Kesinlik (Precision), Hatırlama (Recall) ve F ölçüsü bakımdan karşılaştırılmıştır.

Ölçümlerin hesaplanmasında Python programlama dili ile geliştirme yapılarak sonuçlar belirlenmektedir. Şekil 7.1 de gösterildiği üzerine çıkarılan anahtar kelimeler geliştirilen yazılım üzerinde hesaplanarak çıkmıştır.

Bu değerlendirmede orijinal anahtar kelimenin, algoritmalar tarafından çıkarılan anahtar kelime içinde bire bir geçmesi veya aynı olması koşuluna göre hesaplanmıştır. İki durum için de eşlemenin tam olarak yapıldığı varsayılmıştır.

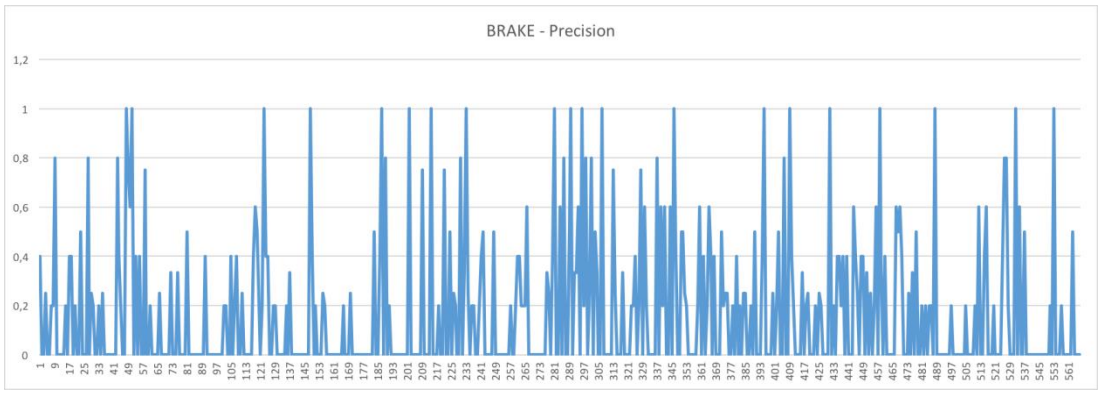
7.4 Değerlendirme Çıktısı

Yapılan değerlendirmeler sonucunda TF-IDF, RAKE ve BRAKE algoritmaları ortalama hassasiyet, ortalama kesinlik ve ortalama f ölçüsü çıktıları yüzdelik olarak Çizelge 7.2'de gösterilmiştir.

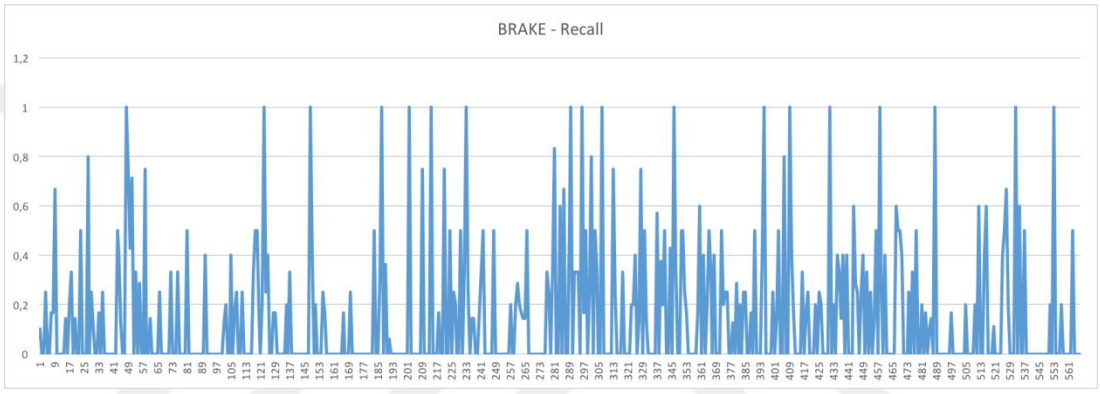
	Ortalama Kesinlik	**Ortalama Hatırlama**	**F Ölçüsü**
TF-IDF	1.8265	1.3965	1.5828
RAKE	9.9082	8.8035	9.3232
BRAKE	16.196	14.665	15.3925

Çizelge 7.2 Hassasiyet, Kesinlik ve F-Ölçüsü çıktıları

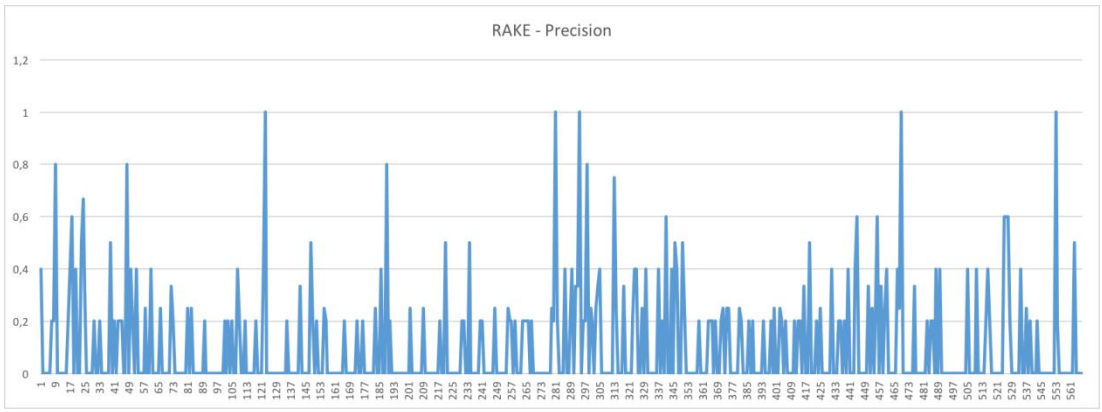
BRAKE algoritmasının, Çizelge 7.2'de gösterildiği gibi, kesinlik, hatırlama ve f-ölçüsüne göre diğer algoritmalardan daha başarılı olduğu hesaplanarak görülmüştür.



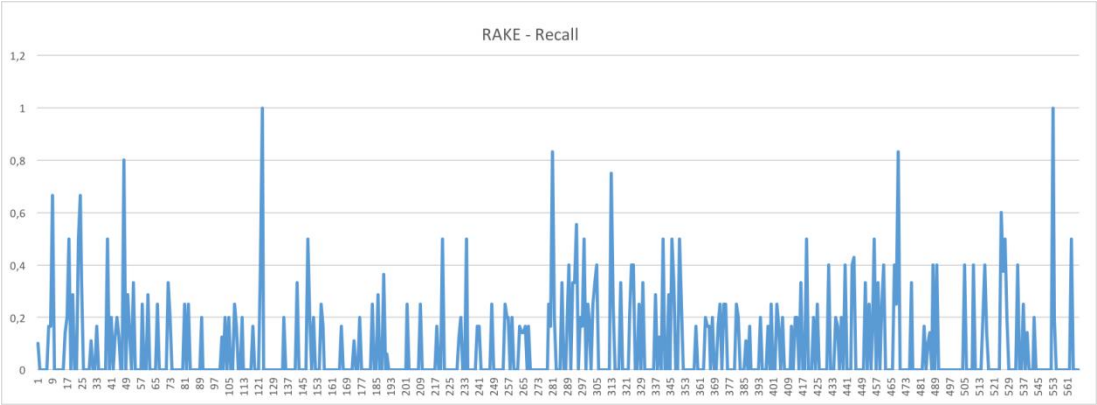
Şekil 7.2 BRAKE algoritmasına ait Kesinlik (Precision) değerleri



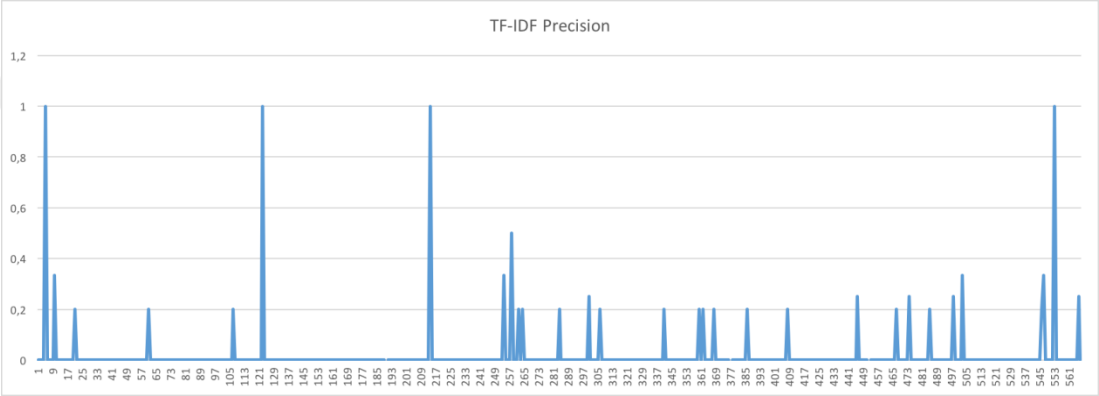
Şekil 7.3 BRAKE algoritmasına ait Hatırlama (Recall) değerleri



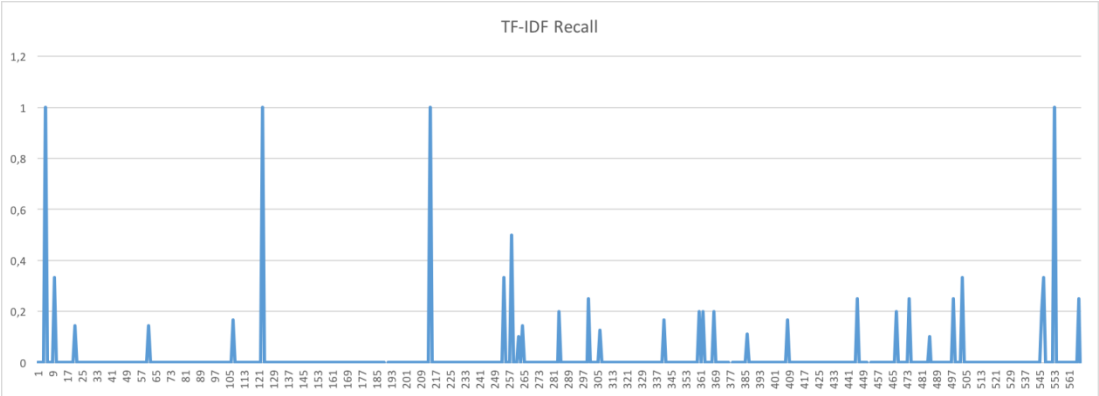
Şekil 7.4 RAKE algoritmasına ait Kesinlik (Precision) değerleri



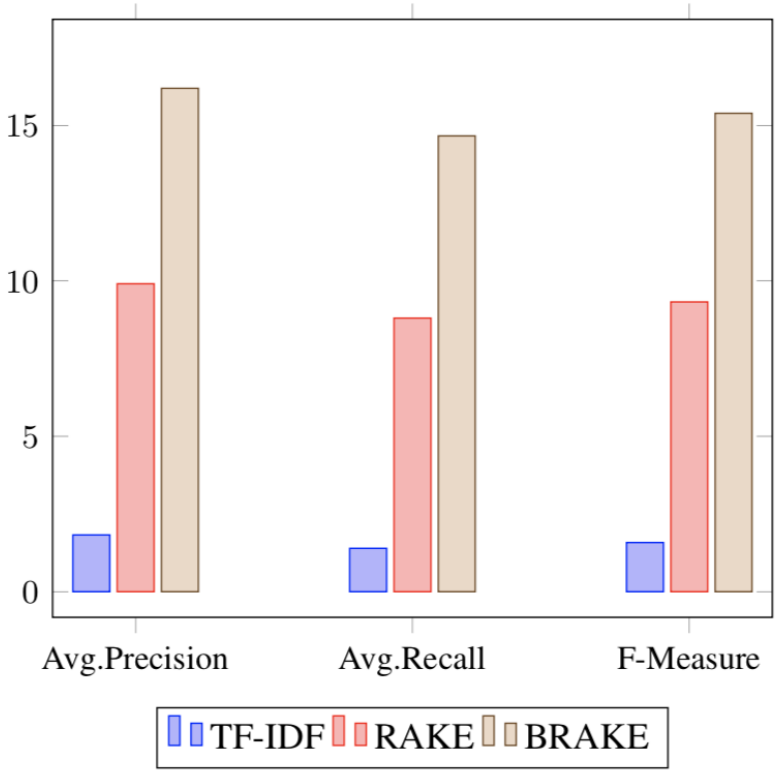
Şekil 7.5 RAKE algoritmasına ait Hatırlama (Recall) değerleri



Şekil 7.6 TF-IDF algoritmasına ait Kesinlik (Precision) değerleri



Şekil 7.7 TF-IDF algoritmasına ait Hatırlama (Recall) değerleri



Şekil 7.8 Algoritmalara ait değerlendirme sonuçları

Yapılan ölçümler sonucu Çizelge 7.2 ve Şekil 7.8'de gösterildiği üzere BRAKE algoritmasının başarılı olduğu görülmektedir.

Bunun beraber Şekil 7.2, Şekil 7.3, Şekil 7.4, Şekil 7.5, Şekil 7.6 ve Şekil 7.7' de her bir makale için hesaplana kesinlik (Precision) ve hatırlama (Recall) değerleri gösterilmiştir.

8. SONUÇ VE ÖNERİLER

Bu tezde, eş anlamlı temel denetimsiz yaklaşıma dayanan BRAKE algoritması sunulmuştur. BRAKE algoritması, RAKE algoritması üzerine, eş anlama gelen farklı kelimelerin tek bir anahtar kelime alına toplanmasına dayanan bir yaklaşım ile tasarlanmıştır. Akademik makaleler üzerine, yazarların kendi seçtikleri anahtar kelimeler esas alınarak yapılan değerlendirme sonuçlarına göre BRAKE algoritması kıyaslanan diğer algoritmalara göre daha yüksek oranla eşlemeler yakalamıştır.

Yayınlanan makalelerin yapılan ön işleme ile PDF formatından düz metne dönüştürmesi sırasında format ve şekillerden kaynaklı bozulmaları algoritmanın performansını olumsuz olarak etkilemektedir. Özellikle tablolar gibi anahtar kelimelerin sıklıkla geçtiği yapılardaki tekrarlar metin istatistiklerini değiştirdiği için farklı anahtar kelimelerin tespitine sebep olmuştur. Ön işlemdeki başarı algoritmanın performansına doğrudan etki etmektedir.

Eş anlamlı kelimelerin tespiti sırasında, metin içerisindeki kısaltmalar göz ardı edilmiştir. Ayrıca sözcüklerin sonlarına çeşitli ekler (‘s’, ’ing’ gibi) geldiği durumlarda genellikle eşanlamlılar listesinde bir eşleme yapılamamaktadır. BRAKE algoritmasının başarısı, eşanlamlı listenin başarısına ve doğruluğuna bağlıdır. Bu sebeplerden dolayı, eklerin ayrılması ve kısaltmaların gerçek kelimeler ile değiştirilmesi ile başarı oranının artırılması mümkündür.

Ayrıca, genellikle metin içerisinde tekrar eden başlık, özet ve sonuç gibi kısımlarda geçen kelimelerin farklı şekillerde puanlanması ile algoritmanın başarısı arttırılabilir.

KAYNAKLAR

ACM, 2012, The 2012 ACM Computing Classification System, Erişim Tarihi: 04.04.2018, https://www.acm.org/publications/class-2012

Adji, T. B., Abidin, Z., Nugroho, H. A., 2014 ,System of negative indonesian website detection using tf-idf and vector space model, in 2014 International Conference on Electrical Engineering and Computer Science (ICEECS), 174–178.

Bharti, S. K. and Babu, K. S, "Automatic keyword extraction for text summarization: A survey," CoRR, vol.abs/1704.03242, 2017. http://arxiv.org/abs/1704.03242

Barker, K., Cornacchia, N., 2000, Using Noun Phrase Heads to Extract Document Keyphrases, Springer-Verlag, AI '00, 40-52

Beliga S., 2014 Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, Rijeka.

Luhn, H. P., 1957 A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, 309 317.

Derczynski, L.,2016, Complementarity, F-score, and NLP Evaluation. LREC .

Feather, J. and Sturges. P., 1996, International encyclopedia of information and library science. London & New York: Routledge

Guan, J., 2016, A Study of the Use of Keyword and Keyphrase Extraction Techniques for Answering Biomedical Questions, A thesis submitted to Macquarie University for the degree of Master of Research Department of Computing,93,Sydney

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann Publishers. 740, Massachusetts

Haggag, M. H., 2013. Keyword Extraction using Semantic Analysis, International Journal of Computer Applications (0975 – 8887)

Jing L.-P., Huang H.-K., Shi H.-B., 2002, Improved feature selection approach tfidf in text mining, in Proceedings. International Conference on Machine Learning and Cybernetics, 944–94.

Leung. A, 2016, Evaluating automatic keyword extraction for internet reviews, Universıty Of Lorraine Realself Inc, M.Sc. Thesis,46, Lüksemburg

Manning, Christopher D. and Schütze, Hinrich., 1999. Foundations of Statistical Natural Language Processing, MIT Press,704,Cambridge, MA, USA

Medelyan, O., Witten, I. H., 2006, Thesaurus based automatic keyphrase indexing, in Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, ser. JCDL ’06. New York, NY, USA: ACM, 296–297.

Rose, S., Engel D., Cramer N., Cowley W., 2010 Automatic keyword extraction from individual documents, Text Mining: Applications and Theory, 1–20.

Python, 2018, What is Python Executive Summary, Erişim Tarihi: 14.04.2018, https://www.python.org/doc/essays/blurb/

Siddiqi, S., Sharan, A., 2015, Keyword and keyphrase extraction techniques: A literature review, International Journal of Computer Applications, sf. 1–6.

SQLite, 2018, About SQLite, Erişim Tarihi: 14.04.2018, https://www.sqlite.org/about.html

Stanford NLP Group, 2017, Stemming and lemmatization, Erişim Tarihi: 04.04.2018, https://nlp.stanford.edu/IRbook/html/htmledition/stemming-and-lemmatization-1.html

Kim, S. N., Medelyan, O., Kan, M.Y., Baldwin, T., 2010, Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles, Proceedings of the 5th International Workshop on Semantic Evaluation, 21-26

Turney, P. D., 2002, Extraction of keyphrases from text: Evaluation of four algorithms, CoRR, National Research Council of Canada

Zhang K., Xu H., Tang J., Li J.-Z., 2006 Keyword extraction using support vector machine, Advances in Web-Age Information Managemen, Springer Berlin Heidelberg, 85–96, 01

ÖZGEÇMİŞ

Adı Soyadı : Ahmet Sina BİRDEVRİM

Doğum Yeri ve Yılı : İstanbul, 20/02/1990

Medeni Hali : Bekar

Yabancı Dili : İngilizce

E-posta : ahmetsinabirdevrim@gmail.com

Eğitim Durumu

Lise :İhsan Kurşunoğlu Lisesi, 2008

Ön Lisans :Doğuş Üniversitesi, 2010, Meslek Yüksek Okulu, Bilgisayar Programcılığı

Lisans :İstanbul Ticaret Üniversitesi, 2014, Mühendislik Ve Mimarlık Fakültesi, Bilgisayar Mühendisliği

Mesleki Deneyim

Felece Teknoloji	2015-2016
KoçSistem	2016-2018
Türkiye Finans Katılım Bankası	2018- Devam

Yayınları

1 – Birdevrim S., 2018. İyileştirilmiş Otomatik Anahtar Kelime Çıkarımı (BRAKE). Teknoloji ve Uygulamalı Bilimler Dergisi, Basımda