©2018 Published in 6th International Symposium on Innovative Technologies in Engineering and Science 09-11 November 2018 (ISITES2018 Alanya – Antalya - Turkey)

Büyük Veri Setlerinde Varlık Tanıma: En Sık Geçen E-Posta, Web Adreslerinin ve Emojilerin Tespit Edilmesi

*[1]Ahmet Arslan, [1]Ahmet Alkılınç, [2]Bekir Taner Dinçer
[1]Bilgisayar Mühendisliği Bölümü, Eskişehir Teknik Üniversitesi, Türkiye
[2]Bilgisayar Mühendisliği Bölümü, Muğla Sıktı Koçman Üniversitesi, Türkiye

Özet

İnternetin ve sosyal web sitelerinin ortaya çıkmasıyla birlikte, dijital verilerin hacmi her geçen gün artmaktadır. Bu büyük miktardaki verilerden anlamlı bilgi elde etmek ve işlemek o kadar kolay değildir. Geleneksel yöntemleri ve araçları kullanarak bu büyük veriyi işlemek oldukça külfetli ve zaman alıcıdır. Bu gibi durumlarda, büyük veri işleme araçları bir çözüm olarak devreye girmektedir. Bu çalışmada büyük veri indeksleme ve arama yazılımı olan Apache Lucene kullanılarak, yarım milyar Web sayfası içinde en sık geçen e-posta, Web adresleri ve emojilerin nasıl tespit edildiği anlatılmaktadır.

Anahtar Kelimeler: veri indeksleme, Apache Lucene, e-posta adresi, Web adresi, emoji

Abstract

With the advent of the Internet and its social websites, the volume of digital data is increasing day by day. It is not so easy to process and to extract meaningful information from this massive amounts of data. It is quite cumbersome and time consuming to process data this big using conventional methods and tools. Big data processing tools come into play as a remedy in such cases. This paper describes how Apache Lucene, which is an open-source software for indexing and searching big data, can be used to extract the top-*n* most frequent e-mails, URLs and emojis from a half billion Web pages.

Key words: data indexing, Apache Lucene, e-mail, URL, emoji

1. Giriş

ClueWeb09 veri seti, bilgi erişimi ve ilgili doğal dil teknolojileri araştırmalarını desteklemek için Carnegie Mellon Üniversitesi'nin Dil Teknolojileri Enstitüsü tarafından oluşturulmuştur [1]. Söz konusu veri seti, on dilde Ocak ve Şubat 2009'da toplanan 1 milyar Web sayfasından oluşmaktadır. İngilizce olan kısmı kategori A olarak da adlandırılır ve toplam 500 milyon Web sayfasından oluşmaktadır. Sıkıştırılmış halde diskte kapladığı alan 2.3 terabayttır. Bu çalışmada, bu denli büyük bir veri üzerinde e-posta ve Web adreslerinin (URL) yansıra emojilerin tanınması ve bu üç varlığın (e-posta, URL ve emoji) derlemde gözlenme sıklıklarının nasıl çıkarıldığı anlatılmaktadır. Bu çalışmada anlatılan yöntem kolayca büyük Twitter[1] verisi içinde en sık geçen @mention ve #hashtag varlıklarının tespiti için özelleştirebilir. İlk etapta, verilen bir metin içinde e-posta, URL ve emojilerin tespit edilmesi düzenli ifadeler vasıtasıyla kolayca gerçekleştirebilecek bir işlem gibi görünse de büyük veri (500 milyon sayfa 2.3 TB) üzerinde bu işlemi gerçekleştirmek ve varlıkları gözlenme frekanslarına göre sıralamak o kadar da basit bir iş değildir. Bunun için büyük veri araçları kullanmak gerekir.

[1] https://twitter.com

*Corresponding author: Address: Faculty of Engineering, Department of Computer Engineering Eskisehir Technical University, 26555, Eskisehir TURKEY. E-mail address: aarslan2@anadolu.edu.tr, Phone: +902223213550

Biz bu çalışmamızda yazılım olarak Apache Lucene [2] (sürüm 7.4.0) kullandık. Apache Lucene endüstride en çok kullanılan bilgi erişimi aracıdır. Ancak akademik çalışmalarda kullanımı çok yaygın değildir. Bunu değiştirmek için çeşitli adımlar atılmaktadır. Örneğin, geçen sene en prestijli bilgi erişimi konferansı olan "Special Interest Group on Information Retrieval" [3] (SIGIR) 2017'de bu amaca yönelik "Lucene for Information Access and Retrieval Research"[2] adında bir çalıştay düzenlenmiştir [4]. Bu çalıştayın amacı Lucene kullanarak standart bilgi erişimi derlemleri üzerinde akademik deney yapılmasını yaygınlaştırmaktır. Önümüzdeki yıllarda daha çok akademik araştırmacının bilgi erişimi çalışmalarında Lucene kullanacağını göreceğiz.

Bu bildiri¸ şu şekilde devam etmektedir. Lucene'nin hangi bileşenlerinin kullanıldığı, kısım 2'de anlatılmıştır. URL, e-posta ve emoji varlıklarının tanınmasında kullanılan yöntem kısım 3'te anlatılmıştır. En sık geçen varlıklar kısım 4'te listelenmiştir. Son kısımda ise sonuç ve ileri çalışma verilmiştir.

2. Apache Lucene

Apache Lucene [2] geliştirilmesi ilk Dough Cutting tarafından başlanan arama kütüphanesidir. Özelliği ise ilk açık kaynak kodlu bilgi erişimi kütüphanelerinden birisi olmasıdır. Zaman içerisinde olgunlaşmış ve endüstri standardı haline gelmiştir. Bugün en popüler ve en çok kullanılan açık kaynak kodlu yazılım olduğu söylenebilir. ElasticSearch[3], Apache Solr[4] gibi Lucene üzerine inşa edilmiş açık kaynak koldu birçok proje vardır.

Lucene, gerçek hayat problemlerini çözmek için tasarlandığı için oldukça hızlı çalışmaktadır. Dünya üzerine dağılmış yetkin geliştiriciler zaman içerisinde Lucene üzerinde iyileştirmeler yaparak onu daha da hızlandırdılar. Canlı sistemlerde kullanılıyor olması, Lucene içindeki hataların kısa zamanda gerçek kullanıcılar tarafından tespit edilip bildirilmesine sebep oldu. Ve böylece geliştirici camia bildirilen bu hataları düzelterek kodu iyileştirdi. Bir yazılımdaki hataların tespit edilmesi o yazılımı ne kadar çok kişinin kullandığına bağladır.

Her ne kadar büyük veri indeksleme ve arama aracı olsa da Lucene NoSQL, büyük veri analizi, veri madenciliği gibi daha başka alanlarda da kullanılmaktadır. Örneğin bu çalışmada varlık ismi (named entity) olarak nitelendirebileceğimiz e-posta, URL ve emojiler tespit edilecektir. Lucene kütüphanesi oldukça modülerdir ve dolayısıyla var olmayan özelliklerin eklenmesi kolaydır.

Analiz bileşeni Lucene'nin kalbidir [5]. Analiz kısmında serbest halde bulunan düzensiz metin, Tokenizer tarafından önce kelimelere ayrıştırılır. Bu adım için ilk bakışta boşluklara göre ayırmanın yeterli olacağı düşünülse bile, farklı alanlardan gelen düzensiz metin için daha karmaşık yöntemler gereklidir. Bu duruma, tire ile ayrılmış ürün isimleri (wi-fi); hem harf hem de rakamlardan oluşan marka isimleri (SD500); büyük küçük harf içeren ürün adları (iPhone); Twitter verisi içindeki özel biçimlemeler (#hashtag, @mention); programlama dilleri (C++, C#), noktalama işaretleri, IP adresleri (193.140.21.123), kısaltmalar (T.C.) gibi örnekler verilebilir.

[2] https://liarr2017.github.io
[3] https://www.elastic.co
[4] http://lucene.apache.org/solr/

Tokenizer'ın çıktısı olan bu kelimeler daha sonra TokenFilter tarafından çeşitli dönüşümlere tabi tutulur. Bir TokenFilter girdisi olan kelimeyi yok edebilir, ikinci bir kelime yerleştirebilir ya da kelimeyi değiştirebilir. Örneğin eş anlamı kelimeleri ekleme işi SynonymFilter, kelimeleri küçük harfe çevirme işi LowerCaseFilter, çok sık kullanılan kelimelerin (ve, veya, ile) çıkarılma işlemi StopFilter tarafından yapılır.

İşte bu, bir Tokenizer ve sıfır ya da fazla TokenFilter uç uca zincirleme¸ seklinde eklendiğinde bir Analyzer elde edilmiş olur. Lucene kendi içinde arasından seçim yapmak üzere birçok Tokenizer ve TokenFilter implementasyonlarını barındırır. Örneğin bu çalışmada emojileri tanıyan ICUTokenizer; e-posta ve Web adreslerini (URL) tek bir kelime olarak tanıyan UAX29URLEmailTokenizer kullanılmıştır.

Aslında Tokenizer'dan önce çalışan üçüncü bir bileşen daha vardır. Bu bileşenin ismi CharFilter olup, HTML etiketlerinin ayıklanması, metnin ASCII'leştirilmesi gibi tokenizer adımı öncesi gerçekleştirilmek istenilen işlemler içindir.

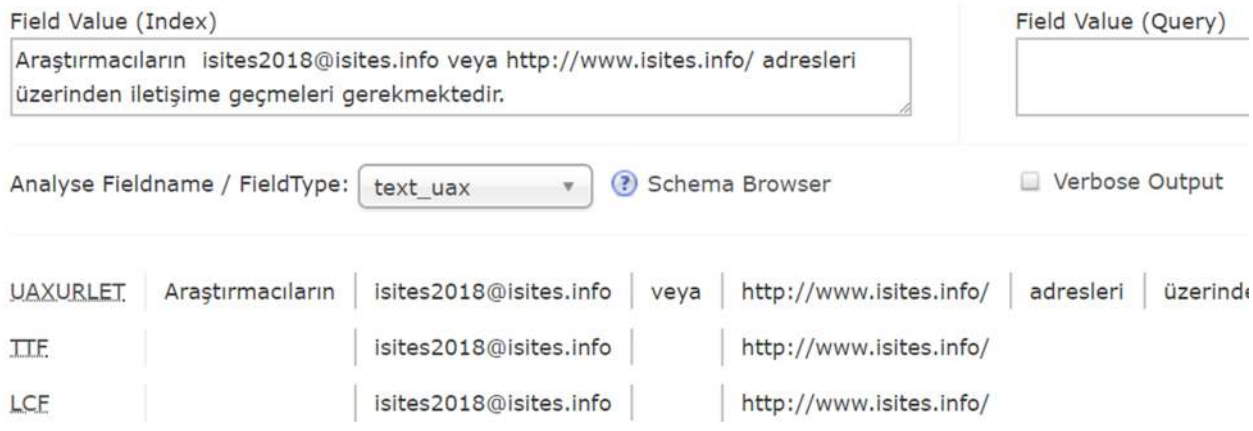
Lucene'nin analiz modülü dekoratör tasarım deseni kullanılarak yazılmıştır. Yukarda bahsedilen bileşenler bir zincir misali sırası ile çalışır. Yeni bir özellik eklemek istediğimizde bu özelliği yeni bir bilesen yazarak yaparız. Var olan bileşenleri değiştirmeyiz.

3. Yöntem

Bu kısımda e-posta, URL ve emoji varlıklarının nasıl tanındığı anlatılmaktadır.

3.1 E-posta ve URL

Kısım 2'de anlatılan analiz bileşenlerinin farklı farklı kombinasyonları ile değişik analyzer'lar elde edilebilir. Bu çalışmada 3 bileşenden oluşan bir analyzer kullanılmış olup, örnek bir cümle üzerinde nasıl çalıştığı Sekil 1'de gösterilmiştir.



Şekil 1. Lucene Örnek Cümle Analizi

Örnek üzerinden gidilirse, UAXURLET tokenizer'ın isminin baş harfleridir. Örnek cümleyi kelimelere bölmesinin yanı sıra, kelimelerin tipini de tayin etmektedir. Tanıdığı 9 tip sınıfı vardır ve <EMAIL> ve <URL> tipleri bunların arasında yer almaktadır. Tokenizer'dan sonra gelen ilk filtre tiplere göre eleme ya da tutma yapmaya yarayan TypeTokenFilter (TTF)'dir. Çalışmada ilk olarak sadece <EMAIL> ve <URL> tipinde terimler indekslenip, geri kalan her şey görmezden gelinmiştir. Sonuncu filtre ise kelimeleri küçük harfe dönüştüren LowerCaseFilter (LCF)'dir. Örnek cümle analiz zincirinden geçtiğinde geriye sadece iki tane kurtulan terim kalmıştır. Bu 3 bileşeni kullanarak bir derlemi indekslendiğimizde, elimizde sadece e-posta ve Web adreslerinden oluşan bir ters indeks olacaktır. Lucene bir indeks içinde en sık geçen terimleri listelemek için HighFreqTerms isimli bir API sunmaktadır. Terimleri iki kritere göre sıralamak mümkündür: doküman frekansı ve derlem frekansı. Doküman frekansı varlığın kaç adet belgede gözlemlendiğidir. Derlem frekansı ise varlığın tüm derlemde toplamda kaç defa gözlemlendiğidir.

3.2 Emoji

Çalışmanın ikinci aşamasında veri setlerinde en çok karşılaşılan emojiler tespit edilmiştir. Emoji, tipik olarak renkli bir formda sunulan ve metinde satır içi kullanılan resimler veya sembollerdir [6]. Yüzler, hava durumu, araçlar ve binalar, yiyecek ve içecek, hayvanlar ve bitkiler veya duyguları, etkinlikleri ve simgeleri temsil ederler. Tablo 1'de listelendiği üzere; aile, kalp, gözlük, kedi, gül, üzüm, otomobil gibi farklı alanlarda binlerce emoji bulunmaktadır. Emojiler duyguları ifade eden renkli ve eğlenceli bir kavramdır. Yeni nesil tarafından kısa metinli yazışmalarda yoğun olarak kullanılan bir iletişim aracı haline gelmiştir. Kelimelerle ifade etmekte zorlandığımız birçok şeyi emojiler aracılığıyla daha kolay bir şekilde ifade etmeye başladık. "The Emoji Movie" [7] isimli bir emoji filmi bile çekilip 2017 yılında vizyona girmiştir. Dahası, her yılın 17 Temmuz günü "Dünya Emoji Günü" [8] olarak kutlanmaktadır.

Tablo 1. Çeşitli Emoji Örnekleri

							
Aile	Kalp	Gözlük	Kedi	Gül	Üzüm	Dondurma	Otomobil

Emojiler bilgi erişiminden çok duygu analizi çalışmalarında önem yer teşkil etmektedir [9, 10]. Örneğin, kızgın surat içeren bir tweet mesajı, mesajı atan kişinin kızgın olduğuna, gülen surat içeren bir mesajda kişinin mutlu olduğuna işarettir.

Emojiler noktalama işaretleri ve parantezlerin kombinasyonlarından oluştuğu için bir metin içerisinde tespit edilmesi hiç de kolay bir işlem değildir. Bu çalışmada emojilerin tespiti için Lucene'nin icu-analiz modülünün bir alt bileşeni olan ICUTokenizer kullanılmıştır. ICU (International Components for Unicode - Unicode için Uluslararası Bileşenler) [11], temeli IBM firması tarafından atılmış, şu anda birçok firmanın destek verdiği C/C++ ve Java programlama dilleri ile yazılan uygulamalar için evrensel kod ve küreselleşme desteği sağlayan olgunlaşmış bir kütüphaneler projesidir. ICUTokenizer UAX #29: Unicode Metin Segmentasyon kurallarına göre metinleri kelimelere dönüştürür [12].

ICUTokenizer'ın emojileri tanıma özelliği Apache Lucene 7.4.0 versiyonu ile gelmiştir. Daha önceki versiyonlarda böyle bir özellik yoktur. ICUTokenizer sayesinde metin içerinde geçen bir emoji <EMOJI> tipinde gösterilmektedir. Bu özellik kullanılarak çalışmamızdaki veri setlerinde <EMOJI> tipinde terimler indekslenip, geri kalan her şey indeks dışında tutulmuştur. Bu sayede veri seti içerinde geçen emojiler tespit edilmiştir. Daha sonra bu emojilerin doküman frekansı ve derlem frekansı hesaplanarak en çok kullanılan 10 emoji tespit edilmiştir. Bir sonraki bölümde veri setlerinden çıkarılmış e-posta, URL ve emoji varlıkları listelenmiştir.

4. Deneysel Sonuçlar

4.1. En Sık Gözlemlenen E-posta ve URL Adresleri

ClueWeb09 ve ClueWeb12-B13 İngilizce derleminde en sık gözlemlenen 10 e-posta ve URL varlık isimleri Tablo 2 ve Tablo 3'te derlem frekanslarına göre büyükten küçüğe doğru sıralı şekilde listelenmiştir. ClueWeb12 [13] veri seti, ClueWeb09'ın halefidir. 733,019,372 adet İngilizce Web sayfasından oluşur ve Ocak-Mayıs 2012 tarihlerin arasında toplanmıştır. Yıllar arası karşılaştırma yapabilmek için anlatılan çıkarma yöntemi ClueWeb12 derlemi üzerinde de çalıştırılmıştır. Ancak elimizde tüm derlem olmadığı için, 50 milyon Web sayfasından oluşan ClueWeb12-B13 alt kümesi kullanılmıştır.

Tablo 2. ClueWeb09 ve ClueWeb12-B13 Web Adresleri

ClueWeb09			ClueWeb12-B13		
URL	Derlem Frekansı	Doküman Frekansı	URL	Derlem Frekansı	Doküman Frekansı
del.icio.us	13.190.210	10.575.050	astalavista.box.sk	1.560.974	1.560.694
amazon.com	8.536.716	5.198.552	crack.cd	1.559.920	1.559.65
myspace.com	5.638.749	1.972.189	lomalka.ru	1.559.396	1.559.396
alibaba.com	4.034.294	907.277	sta.sh	1.419.991	564.180
about.com	3.874.305	1.338.001	skyrock.com	1.209.964	865.424
asp.net	3.560.500	1.335.366	skyrock.com	1.082.025	1.081.990
wordpress.org	2.861.709	2.813.372	wordpress.com	868.049	559.784
wordpress.com	2.834.811	2.144.027	amazon.com	805.652	520.472
local.com	2.718.042	1.368.481	del.icio.us	726.745	440.941
wordpress.com.	2.650.914	2.636.796	wordpress.com	651.659	634.088

Bu Web adreslerinin Web sayfalarından hyperlink verilmiş sayfalar değil, metin içinde geçen adresler olduğunun altı çizilmesi gerekir. Tam da bu yüzden elde edilen Web adreslerinde http, https gibi protokol isimleri yoktur. ClueWeb09 veri kümesinde birinci sırada Web yer imlerini saklamaya, paylaşmaya ve keşfetmeye yarayan sosyal yer imi sitesi del.icio.us gelmektedir. İkinci sırada ise bir elektronik ticaret sitesi olan amazon.com gelmiştir. Bir sosyal ağ sitesi olan myspace.com üçüncü sıraya oturmuştur. Diğer bir e-ticaret sitesi alibaba.com derlem frekansına göre dördüncü olmuştur.

ClueWeb12-B13 veri kümesinde elde edilen sonuçlar incelendiğinde, güvenlik ile ilgili Web sitelerini arama motoru astalavista.box.sk birinci sırada gelmiştir. ClueWeb09'un ilk onuna giremeyen çevrimiçi radyo sitesi skyrock.fm, ClueWeb12'nin listesinde belirmiştir.

En sık geçen e-posta adresleri genelde Web yöneticisi adreslerinden oluşmaktadır. Dahası göreceli olarak e-posta adreslerinin derlem frekansları URL adreslerinin derlem frekanslarından daha azdır.

Tablo 3. ClueWeb09 ve ClueWeb12-B13 E-posta Adresleri

ClueWeb09			ClueWeb12-B13		
Eposta	Derlem Frekansı	Doküman Frekansı	Eposta	Derlem Frekansı	Doküman Frekansı
yellowpages@gatehousemedia.com	148.416	148.414	r@dio.mp	35.290	6.900
webmaster@owneriq.net	132.214	132.209	userhelp@guardian.co.uk	11.270	4.894
ubuntu-users@lists.ubuntu.com	131.222	556	admin@lemurproject.org	9.439	8.429
mhoy@cs.cmu.edu	117.129	81.676	commons@hellometro.com.	9.141	9.141
tech@allfind.us	89.193	89.193	letters@guardian.co.uk	8.770	4.539
admin@allfind.us	89.193	89.193	9fans@cse.psu.edu	8.362	15
webmaster@bitpipe.com	72.891	72.891	bugzilla-daemon@bugzilla.ximian.com	7.214	8
webmaster@techtarget.com	71.600	71.189	noreply@blogger.com	7.059	1.471
noreply@blogger.com	70.297	25.282	reader@guardian.co.uk	6.981	3.597
rss@youtube.com	66.619	22.473	janedoe@aol.com.	5.931	5.931

Eğer Tablo 2, ve 3'ün son satırları incelenirse, nokta ile biten URL ve e-posta adresleri görülür. Bu satırlar Lucene'nin hata yaptığını gösterir. Bir başka değişle UAX29URLEmailTokenizer.java kodunda hata vardır. Bu durum iki türlü çözülebilir: (*i*) Son karakteri nokta ise o karakteri silen bir TokenFilter uygulamak. (*ii*) Bu hatayı Lucene kullanıcı veya geliştirici e-posta listesine bildirip, hatanın geliştiriciler tarafından düzeltilmesini sağlamak. Açık kaynak kod akımına katkıda bulunmak adına hatayı raporlayan ikinci adımı takip etmeyi düşünüyoruz.

4.2. En Sık Gözlemlenen Emojiler

ClueWeb09B ve ClueWeb12-B13 de en sık gözlemlenen 10 emoji Tablo 4'te derlem frekanslarına göre büyükten küçüğe doğru sıralı şekilde listelenmiştir.

Tablodaki sonuçlara baktığımızda ClueWeb09B de birinci sırada siyah küçük kare gelmektedir. Onu takiben ise iskambil kartlarındaki şekiller (kupa, maça, sinek ve karo) gelmektedir. Barış sembolü ilk 10'a girenler arasındadır.

ClueWeb12B derlemine baktığımızda ClueWeb09B'daki gibi ilk dört sıralamada aynı emojilerin farklı sıralamalarda geldiği görülmektedir. Kadın ve erkek cinsiyet simgeleri ilk 10'a girmeyi başarmıştır. Her iki veri kümesinde en sık gözlemlenen emojiler iskambil kağıdı simgeleridir.

Tablo 4. ClueWeb09B ve ClueWeb12-B13 Emojiler

ClueWeb09				ClueWeb12-B13			
Emoji	İsim	Derlem Frekansı	Doküman Frekansı	Emoji	İsim	Derlem Frekansı	Doküman Frekansı
▪	(small black square)	400.448	32.678	♥	(heart suit)	47.350.38	1.147.599
♥	(heart suit)	350.504	53.585	▶	(play button)	477.070	40.16
♦	(diamond suit)	141.081	19.121	▪	(small black square)	418.207	44.101
▶	(play button)	79.693	13.627	♦	(diamond suit)	375.644	55.398
♣	(club suit)	73.680	4.888	✔	(heavy check mark)	261.401	65.017
♠	(spade suit)	31.425	5.444	♂	(male sign)	223.431	45.576
▫	(white small square)	27.503	1.996	♀	(female sign)	222.562	47.561
◀	(reverse button)	6.921	3.481	♣	(club suit)	207.155	28.132
☺	(smiling face)	6.882	2.606	♡	(red heart)	163.410	92.596
☮	(peace symbol)	3.212	2.907	✖	(heavy multiplication x)	69.914	20.518

5. Sonuç ve Değerlendirme

Büyük veri günümüz bilgi çağının en önemli konularından birisidir. Büyük veri işlemek için çoğunlukla Drill, Flume, Flink, Kafka, Hadoop, HBase, Hive, Pig, Spark, Lucene gibi Apache Software Foundation projeleri kullanılmaktadır. Üstelik Apache 2.0 lisansı kısıtlayıcı değil müsamahakârdır. Ülkemiz endüstrisinde de bu açık kaynak kodlu yazılımların kullanımı gittikçe artmaktadır. Dünya çapında ticari çözümlerden açık kaynak kodlu yazılımlara göç olduğu söylenebilir. Bu çalışmada hem en çok kullanılan Apache projesi olan Lucene kısaca tanıtılmış hem de Lucene kullanarak yarım milyar Web sayfası içinde en sık geçen e-posta ve Web adreslerinin yansıra emojiler de tespit edilmiştir. Bu çalışma bir demo (kavram ispatı) amacı gütmektedir. Eğer akademik bir çalışmaya doğru evirilmek istenirse, e-posta ve Web adreslerinin tek kelime olarak indekslemenin bilgi erişimi başarımı üzerindeki etkisi incelenebilir. (Eğer harf olmayan karakterlerde bölme işlemi yapılsaydı e-posta ve Web adresleri birden fazla parçaya bölünürdü.)

Teşekkür

Bu bildiri Anadolu Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından 1709F516 numarası ile desteklenen bilimsel araştırma projesi kapsamında hazırlanmıştır.

Kaynaklar

[1] J. Callan, M. Hoy, C. Yoo, and L. Zhao, "The ClueWeb09 dataset," 2009. [Online]. Available: http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf

[2] A. Białecki, R. Muir, and G. Ingersoll, "Apache Lucene 4," in Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval, Portland, Oregon, USA, Aug. 2012, pp. 17–24. [Online]. Available: http://opensearchlab.otago.ac.nz/paper_10.pdf

[3] SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan: ACM, 2017.

[4] L. Azzopardi, M. Crane, H. Fang, G. Ingersoll, J. Lin, Y. Moshfeghi, H. Scells, P. Yang, and G. Zuccon, "The Lucene for information access and retrieval research (LIARR) workshop at SIGIR 2017," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '17. Shinjuku, Tokyo, Japan: ACM, 2017, pp. 1429–1430. [Online]. Available: http://doi.acm.org/10.1145/3077136.3084374

[5] M. McCandless, E. Hatcher, and O. Gospodnetic, Lucene in Action, Second Edition: Covers Apache Lucene 3.0. Greenwich, CT, USA: Manning Publications Co., 2010.

[6] Erişim tarihi: 31 Ağustos 2018, http://unicode.org/emoji/

[7] Erişim tarihi: 31 Ağustos 2018, https://www.imdb.com/title/tt4877122

[8] Erişim tarihi: 31 Ağustos 2018, https://worldemojiday.com

[9] Gezici G., Yanıkoğlu B., "Sentiment Analysis in Turkish." In: Oflazer K., Saraçlar M. (eds) Turkish Natural Language Processing. Theory and Applications of Natural Language Processing. Springer, Cham, 2018.

[10] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, K. Uzay, "Exploiting emoticons in sentiment analysis," Proceedings of the 28th Annual ACM Symposium on Applied Computing, March 18-22, 2013, Coimbra, Portugal

[11] Erişim tarihi: 31 Ağustos 2018, http://site.icu-project.org/

[12] Erişim tarihi: 31 Ağustos 2018, http://www.unicode.org/reports/tr29/

[13] J. Callan, "The Lemur project and its ClueWeb12 dataset," 2012. [Online]. Available: http://opensearchlab.otago.ac.nz/SIGIR12-OSIR-callan.pdf