YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 5, Sayı 1, Mayıs 2018

Metin Madenciliğinde Yazar Tanıma (Author Recognition in Text Mining)

Deniz İrem ÜNAL[1], Şadi Evren ŞEKER[2]

1. Yıldız Teknik Üniversitesi,Matematik Mühendisliği Bölümü
2. İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü

Özet

Bu makalenin amacı, metin veri kümelerinde uygulanan yazar tanıma yöntemini açıklamaktır. Yazar tanıma kavramı belirtildikten sonra kısa bir tarihi daha sonra da kullanım alanları açıklanacaktır. İnternetin kullanımın artması ile birlikte anonim yazılar büyük bir artış göstermektedir.

Anahtar Kelimeler:*metin madenciliği,yazar tanıma,veri madenciliği*

Abstract

The purpose of this article is explain author recognition method which as built in text data sets. After the concept of author recognition is indicated, its short history and then its areas of usage is clarified. Anonymous texts show a great increase associated with the increment in the usage of internet.

Keywords:*text mining,author recognition,authorship attribution,data mining*

1. Giriş

Yazar tanıma, yazar bilgisi içeren doküman külliyatındaki (corpus) dokümanların yazarlarını tahmin eden bir çeşit sınıflandırma problemidir. Korsan veya çalıntı yazıların tespitinin yanı sıra internetin kullanımının artması ile birlikte büyük artış yaşayan anonim yazıların yazarlarının tespiti gibi ortamlarda uygulanabilmektedir. Metin sınıflandırmadan farklı fakat doküman içeriğine metin sınıflandırma uygulanan bir tekniktir. Farklı veri kümelerinde farklı sınıflandırma teknikleri ile bambaşka sonuçlar doğurabilme potansiyeli yazar tanıma tekniğini karmaşık ve zor hale getirebilmektedir.

Yazar tanıma üzerine yapılan çalışmalardan Mosteller[1] Bayesian analizi kullanarak yazar tanıma yapmıştır. Burrows[2] en fazla sıklıkta kullanılan kelimeleri, Brinegar[3] kelime ,Morton[4] cümle boyutunu, Brainerd[5] hece sayısını, Holmes[6] kelime sayısı ve belge uzunluğunu, Twedie[7] birbirinden farklı olan kelime sayılarının toplam kullanılan kelime sayısına oranını kullanarak yazar tanıma yöntemini uygulamışlardır. Fürnkranz[8] , Tan[9] ,Çatal[10] öznitelik çıkarımı esnasında n-gramları kullanarak sınıflandırma başarısını arttırmışlar ve bazı sistemler

geliştirmişlerdir. Diri [11] ise belgenin yazarı ve türünü belirleyen ve bu belirlemeleri kullanan bir sınıflandırma yöntemi oluşturmuştur.

Bu çalışmada enron mail veri kümesindeki mail hesaplarının gönderdiği mailler üzerinden yazar tanıma sistemi kurulmuştur.

2.Metodoloji

Bu projede uygulanan işlem adımları Şekil 1'de gösterilmiştir.

Şekil 1: Uygulama Adımları Akış Şeması

- Verilerin Toplanması: Bu aşamada veri kümesi incelenmiştir. Gönderilen mailleri içeren bir excel dosyası oluşturulmuştur.
- Verilerin Önişlenmesi: Bu aşamada veri kümesinde gönderilen maillerin önişleme adımları gerçekleştirilmiştir.
- Öznitelik Çıkarımı: Öznitelik çıkarımı aşamasında önişlemesi tamamlanmış veri kümesi üzerinde öznitelik çıkarımı (feature extraction) yöntemlerinden Tf-idf hesaplanarak normalize edilmiştir.
- Veri Kümesinin Bölünmesi: Veri kümesi eğitim ve test kümesi olarak bölünmüştür.
- Algoritma Seçimi: Projede başarılı sonuç çıkarabilecek algoritmalar üzerinde çeşitli testler ve ayarlamalar yapılmıştır.
- Sonuç: Başarılı olan algoritmaların başarı oranları bu makale ile birlikte sunulmuştur.

3.Enron Mail Veri Kümesi

Enron şirketi (ENE), enerji sektöründe faaliyet gösteren Teksas eyaletinde Houston şehrinde bulunan bir Amerika şirketiydi. İflas etmeden önce çok sayıda çalışana sahip olan dünyanın en büyük elektrik ve doğalgaz şirketlerindendi. 2001 yılında çeşitli muhasebe hileleri ile hisseleri olduğundan değerli ve karlı gösterildiği açığa çıkmış ve iflas etmiştir.

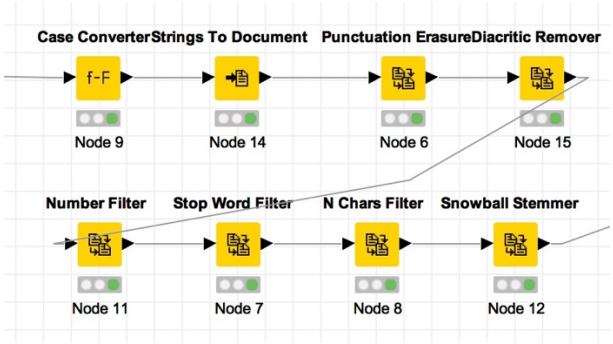
'Enron skandalı' olarak anılan bu olaydan sonra Enron şirketinin çalışanlarının mail veri kümesi yayınlanmıştır. Bu veri seti, var olan veri kümeleri içinde en büyük mail veri kümelerinden biridir. Veri kümesi [12], 150 tane klasör içermektedir. Bu klasörler, '@enron.com' uzantılı maile sahip olan enron çalışanlarının mail hesabından çekilen verilerdir.

3.1.Verilerin Toplanması

Bu çalışmada '_sent_mail' klasörüne sahip 78 tane kişinin bu klasörden çekilen mailleri üzerine yazar tanıma uygulaması yapılmıştır. '_sent_mail' klasöründeki veriler, Python programlama dili ile tarih, gönderen, alıcı, konu, mail kısımları çekilerek excel dosyasına yazdırılmıştır. Bu excel dosyasında her satır bir mail sayısına karşılık gelmektedir ve yazar bilgileri rastgele olarak seçilmiştir. Daha sonra proje entegreli olarak çalışan bir veri analiz platformu olan Knime üzerinde geliştirilmiştir. Bu excel dosyasını Knime'a aktarabilmek için "Excel Reader" düğümü kullanılmıştır.

3.2. Veri Önişleme

Geliştirilen yazar tanıma sisteminde, İlk olarak metinlerden "Punctuation Erasure" düğümü ile noktalama işaretleri kaldırılmıştır. Ardından mail içeriğindeki bütün ifadeler "Case Converter düğümü" yardımıyla küçük harfe çevrilmiştir. Bu işlemden sonra mailler içindeki sayılar "Number Filter" aracılığıyla filtrelenmiştir. İngilizce diline ait olup çok kullanılan kelimeler, bağlaçlar, içerikten bağımsız kelime grupları, kalıplaşmış kısaltmalar gibi çıkarılması gereken kelimeler (stop words)[13] "Stop Words Filter" ile, 4ten daha az sayıda karakter içeren kelimeler "N Chars Filter" düğümü ile dokümandan çıkartılmıştır. Bu işlemler Şekil 2 'de düğümler ile gösterilmiştir.



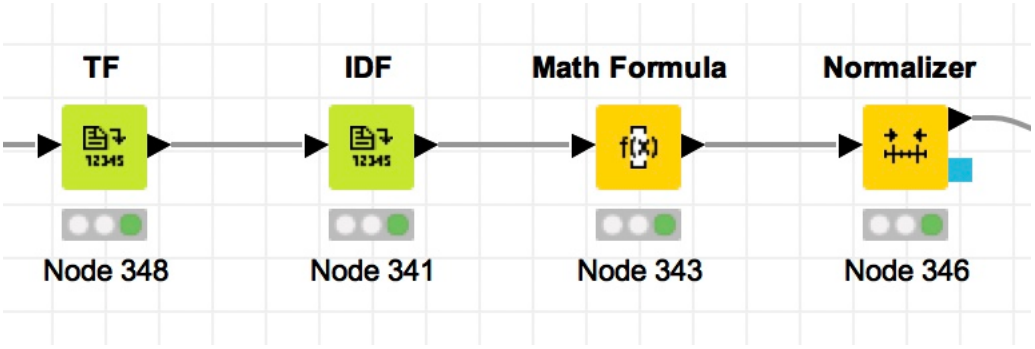
Şekil 2: Önişleme (preprocessing) aşaması

4.TF-IDF ile Öznitelik Çıkarımı (Feature Extraction)

TF (Term Frequency/Terim Sıklığı), bir metinin içinde geçen kelimelerin toplam terim sayısına bölünmesi hesabıdır. IDF (Inverse Document Frequency/Ters Doküman Sıklığı), hesap yapılan terimin hangi metinlerde geçtiğini belirtir. Toplam metin sayısının(x), terimi içeren metin sayısına(y) oranının logaritmik ifadesidir. IDF=log(x/y) olarak formüle edilebilir.

TF-IDF değeri ise bir terimin bir metine olan etkisini istatistiksel olarak hesaplayan ağırlık unsurudur. Formülü ise, TF-IDF=TF*IDF 'dir.

Enron veri kümesinde her bir terim için TF-IDF değeri hesabı Şekil3'de gösterilmiştir. TF değeri için "TF", IDF değeri için "IDF", TF-IDF değeri için "Math Formula" düğümleri kullanılmıştır. Ardından bu TF-IDF değerleri Min-max notasyonuna uygun olarak "Normalizer" düğümü yardımıyla 0-1 aralığında normalize edilmiştir. Burada amaç, en büyük TF-IDF değerini 1, en küçüğünü 0 alarak geri kalan değerleri bu aralığa uygun şekilde yayabilmektedir.



Şekil 3: Tf-idf ile öznitelik çıkarımı işlemleri

5.Veri Kümesinin Bölünmesi

Bütün bu işlemler uygulandıktan sonra veri kümesi, eğitim ve test kümelerine bölünmektedir. Bunun sebebi veri kümesinde etiket sınıfı(gönderen) bulunduğundan gözetimli (supervised) bir öğrenme uygulamasıdır. Eğitim kümesindeki veriler uygulanan algoritmaların öğrenmesine yardımcı olmaktadır. Test kümesi ise öğrenen algoritmanın daha önce görmediği veriler ile yazarını tahmin etmesine olanak sağlamaktadır. Bu adımda "Partitioning" düğümü ile veri kümesi %66 eğitim, %34 test kümesi olarak bölünmüştür.

6.Algoritma Seçimi ve Sınıflandırma (Classification) Algoritmalarının Uygulanması

Bu aşamada uygun algoritmalar seçilmiş, başarılı sonuçlar üretebilecek algoritmalarda ayarlamalar yapıldıktan sonra uygulanmıştır. Uygulanan Karar ağacı (decision tree), K en yakın komşuluk (K Nearest Neighborhood), Naive Bayes, Random Forest (Rassal Ağaç) algoritmalarının sonuçları Şekil 4'te gösterilmiştir.

	KARAR AĞACI	KNN	NAİVE BAYES	RANDOM FOREST
TF-IDF	*%78*	*%65*	*%54*	*%96*

Şekil 4: Uygulanan algoritmaların doğruluk (accuracy) oranları

7.Sonuçların karşılaştırılması

Naive Bayes algoritması, metin madenciliğinde daha iyi sonuç vermesine rağmen bu projede düşük çıkmasının sebebi mail içeriklerinin Naive Bayes algoritmasının bağımsızlık kabulüne uymadığını göstermektedir.

Kaynakça

[1] Mosteller, F., Wallace, D. L., Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Reading, MA:Addison-Wesley, 1984

[2] Burrows, J. F., "Not unless you ask nicely: the interpretative nexus between analysis and information", Literary Linguist Computing, Vol. 7, 1992, p 91-109

[3] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Automatic Text Categorization in Terms of Genre and Author", Computational Linguistics, p 471-495, 2000

[4] Morton, A. Q., "The Authorship of Greek Prose", Journal of the Royal Statistical Society, Series A, 128:169-233, 1965

[5] Brainerd, B., Weighting Evidence in Language and Literature: A Statistical Approach, University of Toronto Press, 1974

[6] Holmes, D. I, Authorship Attribution, Computers and The Humanities, Vol.28, 1994, p 87-106

[7] Tweedie, F., Baayen, H., How Variable may a Constant be Measures of Lexical Richness in Perspective, Computers and The Humanities, Vol. 32(5), 1998, p 323-352

[8] Fürnkranz, J., A Study using n-gram Features for Text Categorization, Austrian Research Institute for Artifical Intelligence, 1998

[9] Tan, C. M., Wang, Y. F., Lee, C. D., The Use of Bigrams to Enhance, Journal Information Processing and Management, Vol.30-4, 2002, p.529-546

[10] Çatal, Ç., Erbakırcı, K., Erenler, Y., "Computer-based Authorship Attribution for Turkish Documents", Turkish Symposium on Artificial Intelligence and Neural Networks, 2003

[11] Diri, B., Amasyalı, M. F., "Automatic Author Detection for Turkish Texts", Artificial Neural Networks and Neural Information Processing, 138-141, 2003
[12] https://www.cs.cmu.edu/~enron/

[13] http://xpo6.com/list-of-english-stop-words/