

Kemik Doğal Dil İşleme Grubu

Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü

Text2Arff

Türkçe Metinlerin Özelliklerinin Çıkarımı
Kullanım Kılavuzu

İçindekiler

1 Genel Bakış 3
1.1 Kullanım Amacı 3
1.4 Büyük Boyutlu Dosyalarla Çalışmak 4
1.5 Tanımlar 4
2 Metinlerin Yazılımla Çıkarılan Özellikleri 5
2.1 Sayılar 5
2.2 Kelime Kökleri 5
2.3 Kelime Tipleri (Türleri) 5
2.4 Ngramlar 5
2.5 Fonksiyonel Kelimeler 6
2.6 Kelime Ekleri 6
2.7 Kelime Kümeleme 6
2.8 Birlikte Geçme Matrisi 7
2.10 Saklı Anlam İndeksleme 7
2.11 Anlamsal Uzay 8
3 Sonuç Dosyalarının Formatı (.arff) 9

1 Genel Bakış

1.1 Kullanım Amacı

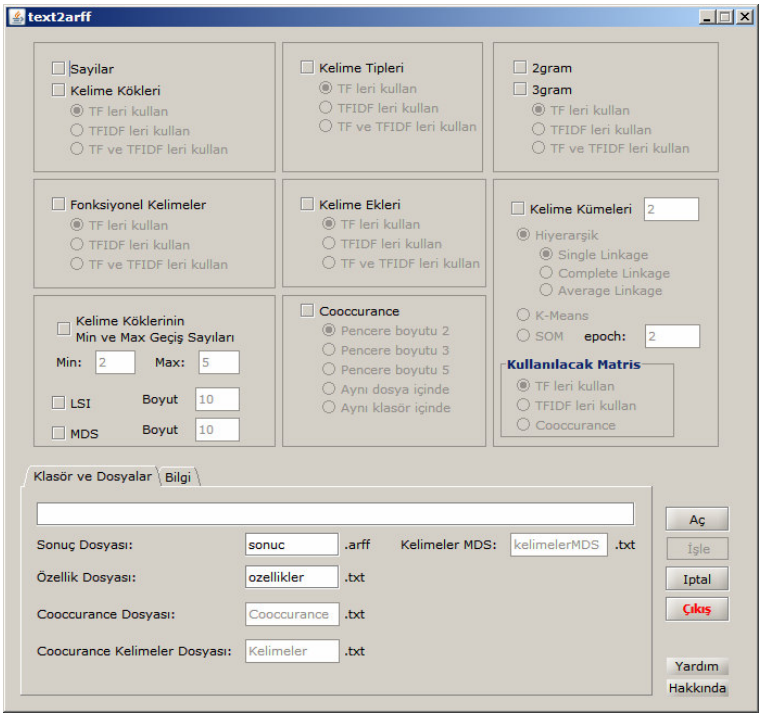
Yazılım temel olarak Türkçe metinlerin WEKA programının giriş dosyası olan ARFF formatına çevirmek için tasarlanmıştır. Program kullanıcısından giriş olarak işlenecek metinlerin bulunduğu klasörün yolunu ve metinlerin hangi özelliklerinin çıkarılmasını istediğini alır, işledikten sonra metinlerin ARFF formatındaki hallerini aynı klasöre kaydeder.

1.2 Metinlerini Formatı ve Klasör Yapısı

Yazılıma verilecek dosyaların uzantısı ".txt" olmalıdır.
Yazılımı verilecek klasörün içinde metinlerin sınıflarını gösteren alt klasörler, her bir alt klasörün içinde de o sınıftaki metinler yer almalıdır. Örneğin 9 sınıfa ait 10'ar adet metin varsa, her birinin içinde 10'ar metnin olduğu 9 adet alt klasörü içeren klasörün yolu sisteme verilmelidir. Bu yapıya uygun bir örnek olarak programla birlikte gelen "Ornek_metinler" klasörü verilebilir. Yazılım, sonuçları (.arff) bu klasörün içine kaydedecektir.

1.3 Programın Çalıştırılması

Program indirdiğiniz klasördeki "text2arff.jar"ın üzerine tıklanarak çalıştırılır. Programın ara yüzü Şekil -1'de verilmiştir.



Şekil 1: Programın Arayüzü

Kullanıcı öncelikle üstte bulunan özellikler bölümünden çıkarılmasını istediği metin özelliklerini seçmelidir. Daha sonra alt bölümdeki "Aç" butonuyla işlenecek dosyaların yolu seçmeli, sonuç ve özellik dosyasının kaydedilmesini istediği isimleri girmelidir. Bu işlemlerden sonra "İşle" butonuna basılarak metinlerin özelliklerinin çıkarılması başlatılabilir. İşlemlerin durumu yine alt bölümde bulunan "Bilgi" sekmesinden takip edilebilir. İşlem bittiğinde bu sekmeden "işlem bitti" mesajı verilecektir.

1.4 Büyük Boyutlu Dosyalarla Çalışmak

Eğer işleyecek dosyaların boyutu büyük ise ve / veya çok sayıda özellik çıkartılacak ise Java'nın heap size'ını arttırmak gerekmektedir. Bunun için programı console'dan aşağıdaki şekilde çalıştırınız:

java -jar -Xmx1024M text2arff.jar

1.5 Tanımlar

Bu bölümde yazılımda kullanılan tanımlar anlatılmıştır.

TF ve TFIDF: Kullanıcıya Kelime Kökleri, Kelime Türleri, Ngramlar, Fonksiyonel Kelimeler, Kelime Ekleri özellik grupları için TF ve TFIDF olmak üzere 2 farklı seçenek sunulmuştur. Aşağıdaki eşitlikte t özelliğinin d metnindeki TFIDF değeri verilmiştir.

$$TFIDF(t,d) = TF(t,d) \cdot \log \frac{|Tr|}{|Tr(t)|}$$

Eşitlikte TF(t,d), t özelliğinin, d metninde geçme sayısıdır. Tr, toplam metin sayısını, Tr(t) içinde en az bir kere t özelliği geçen metin sayısını göstermektedir.
Örnek olarak t özelliği "gr" 2gramı ise, TF(t,d), "gr" ifadesinin d metninde geçiş sayısı olacaktır. T özelliği "isim" kelime türü ise TF(t,d); d metninin içerdiği isim türündeki kelime sayısı olacaktır.

Aşağıdaki örnek matriste 4 metne ait, 7 kelimenin TF değerleri (geçme sıklıkları) verilmiştir.

	Kelime1	Kelime2	Kelime3	Kelime4	Kelime5	Kelime6	Kelime7
1.txt	0	2	1	0	5	7	0
2.txt	3	4	0	0	6	7	3
3.txt	8	9	2	9	0	4	2
4.txt	3	0	0	0	5	7	6

2 Metinlerin Yazılımla Çıkarılan Özellikleri

Yazılım kullanıcısına çıkarılacak metin özellikleri için çeşitli seçenekler sunmaktadır.

2.1 Sayılar

Metindeki kelime, cümle, harf, ek, cümle başına kelime, cümle başına harf, cümle başına ek, kelime başına harf, kelime başına ek, devrik cümle ve noktalama işaretleri sayısını bulduran 19 adet özellikten oluşan özellik grubudur.

2.2 Kelime Kökleri

Metinlerin ifade edilmesinde en yaygın kullanılan özellik grubudur. Metinler içerdikleri kelimelerle ifade edilirler. Metinleri gösteren vektörlerin boyutu, tüm metinlerde en az bir kere geçen farklı kelime sayısıdır. Kelimelerin kendilerini kullanmak Türkçe gibi eklemeli dillerde farklı kelime sayısını çok fazla arttırmaktadır. Ağaç, ağaçlı, ağaçlandırma, ağaçlandırılmış, ağaçlık kelimelerinin hepsinin farklı kelimeler ve dolayısıyla farklı özellikler olarak ele alınması buna örnek verilebilir.
Probleme çözüm olarak literatürde kelimelerin kendileri yerine sadece kelime köklerinin kullanılması önerilmektedir. Bu sayede hem aynı anlama işaret eden kelimelerin birleştirilmesi (böylelikle metinler arası benzerliğin daha iyi ifade edilmesi sağlanmaktadır), hem de özellik boyutunun azaltılmasıyla işlem karmaşıklığının ve gürültü veri olasılığının azaltılması sağlanmaktadır. Kelimelerin köklerinin bulunmasında Zemberek [1] kütüphanesinden faydalanılmıştır.

2.3 Kelime Tipleri (Türleri)

Metinlerdeki geçen kelimelerin türlerini bulduran özellik grubudur. Kelimelerin türlerinin bulunmasında da Zemberek kütüphanesinden faydalanılmıştır. Toplam 15 tane kelime türü (isim, sıfat, zamir, edat vb.) vardır.

2.4 Ngramlar

Bu özellik grubunda metinler içerdikleri ngramları ile ifade edilirler. Ngramlar N boyutlu karakter çerçeveleridir.

Örneğin “gitmek” metni için:

2 gramlar: gi – it – tm – me – ek

3 gramlar: git – itm – tme – mek

Yazılımla metinlere ait 2 ya da 3 gramlar ya da her ikisi birden çıkartılabilir. Daha büyük pencere boyutuna sahip ngramların kullanılmama nedeni çok seyrek frekans matrisleri üretiyor olmalarıdır.

Bu gruptaki özellikler arff dosyasına "2GRAM" ya da "3GRAM" ve frekans türü (TF ya da TFIDF) önekleri eklenerek yazılmaktadır.

2.5 Fonksiyonel Kelimeler

Programla aynı klasörde bulunun fonksiyonel kelimeler dosyasında "fWords.txt" yer alan 620 kelimeye göre metinlerin ifade edilmesinde kullanılan özelliktir. Fonksiyonel kelimeler esasen tek başlarına anlamları olmayan, ancak metinlerin yazarların üsluplarının belirlenmesinde önemli oldukları düşünülen kelimelerdir (ve, daha, gibi, de, için vb.).
Bununla birlikte kullanıcı fonksiyonel kelimeler dosyasını istediği gibi değiştirebilir. Bu sayede kullanıcı istediği özel bir kelime grubuna göre metinleri ifade edebilir. Bu gruptaki özellikler arff dosyasına "FONK" ve frekans türü (TF ya da TFIDF) önekleri eklenerek yazılmaktadır.

2.6 Kelime Ekleri

Türkçe eklemeli bir dil olduğundan anlamın oluşmasında eklerin önemi büyüktür. Bu sebeple yazılımda kelimelerin aldıkları eklerin türlerine çıkarılabilmektedir. Kelimelerin eklerinin belirlenmesinde yine Zemberek kütüphanesinden faydalanılmıştır. Zemberek toplam 126 farklı ek türü çıkarmaktadır. Bu gruptaki özellikler arff dosyasına "EK" ve frekans türü (TF ya da TFIDF) önekleri eklenerek yazılmaktadır.

2.7 Kelime Kümeleme

Birbirine yakın anlamlı kelimeleri birleştirip tek bir küme, tek bir özellik haline getirmek fikrine dayanmaktadır [2]. Kelimelerin kendileri yerine köklerinin kullanılması fikrinde olduğu gibi hem boyut sayısı azaltılmakta hem de benzer anlama sahip kelimeler tek bir kavram olarak kullanılmaktadır. Örneğin kelimeler hayvan kavramına ait kelimeler ayrı ayrı özellik değil, tek bir özellik olarak metinlerin ifadesinde kullanılacaktır. Küme "arff" dosyasında tek bir özellik olarak yazılır. Bir kelime kümesinin bir dokümandaki geçiş sayısı, küme içindeki her bir kelimenin o dokümandaki geçiş sayıları toplamıdır. Yazılım 3 kümeleme algoritmasını desteklemektedir.

1.Hiyerarşik: Başlangıçta her bir örneğin ayrı bir küme olduğu, her bir adımda birbirine en çok benzeyen iki kümenin birleştirildiği algoritmadır [3]. Kümelerin birbirine benzerliklerinin belirlenmesinde kullanıcıya 3 farklı seçenek sunulmuştur. A) En Yakın: İki küme birbirine en yakın elemanları kadar yakındır. B) Ortalama: Yakınlık, iki kümenin her bir elemanının diğer kümenin her bir elemanıyla arasındaki mesafelerin ortalamasıdır. C) En uzak: İki küme birbirine en uzak elemanları kadar yakındır.

2.K-means: Veri dağılımını en iyi temsil edebilecek küme merkezlerinin bulunması fikrine dayanır [4]. Algoritma rastgele atanmış küme merkezleri ile başlar. İterasyonlar ilerledikçe küme merkezlerinin kendi kümelerindeki elemanlara olan ortalama mesafesi azalır.

3.SOM: K-means ile aynı fikre dayanır ancak önerdiği çözüm biraz farklıdır. Algoritmada başlangıçta rastgele atanan küme merkezleri buna ek olarak birbirlerine rastgele bağlanmışlardır. Her bir küme merkezinin

hareketinde ona bağlı olan diğer küme merkezleri de hareket eder. Bu sayede hem verileri ifade edecek küme merkezleri hem de verinin topolojisi ortaya çıkarılmaktadır [5].
Kelimeler kümelemede, ya dokümanlarda geçme sayılarıyla (TF), ya TFIDFlerle ya da diğer kelimelerle birlikte geçme sayılarıyla (birlikte geçme matrisi) ifade edilebilmektedir. Kullanıcı bu 3 özellik türünden istediğin seçebilmektedir. Bununla birlikte kullanıcı, kelimelerin indirgenmesini istediği küme sayısını belirleyebilmektedir. K-means ve SOM algoritmaları için seçilen küme sayısı en fazla anlamındadır. Bunun sebebi başlangıçtaki rastgele atamadan kaynaklanan ya da daha sonradan oluşan boş kümelerin oluşmasıdır. Ayrıca yazılımda K-means ve SOM algoritmaları için kullanıcı iterasyon (epoch) sayısını da belirleyebilmektedir.
Kümeleme yapıldıktan sonra her bir kelime kümesinin içinde hangi kelimelerin olduğu otomatik oluşturulan "özellik.txt" dosyasına yazılmaktadır.

2.8 Birlikte Geçme Matrisi

Bu özellik seçildiğinde Birlikte Geçme (Cooccurance) matrisi oluşturulur ve kullanıcının ara yüzde belirttiği yere "coocurance.txt" yazılır. Aynı zamanda simetrik olan matrisin satır ve sütunlarını ifade eden kelimeler de yine kullanıcının yerini belirlediği başka bir dosyaya "kelimeler.txt" yazılır.
Birlikte geçme matrisinin i,j hücresinin değeri, i.kelime ile j. kelimenin metin kütüphanesi içinde çeşitli boyuttaki çerçevelerde birlikte bulunma sayılarını ifade eder. Kullanıcı seçebileceği pencere boyutları 2, 3, 5, aynı dosya içinde ve aynı klasör içinde olmak üzere 5 adettir. Pencere boyutunun bir sayı olması iki kelimenin tüm metinlerde o sayı büyüklüğündeki bir pencere içinde kaç kez birlikte bulunduğunu göstermektedir. Aynı metin ya da klasör içinde olması ise sırasıyla iki kelimenin birlikte bulundukları metin ya da klasör (uygulamamızda klasörler sınıfları temsil etmektedir) sayısını göstermektedir.

2.9 Kelime Filtreleme

Metinlerde çok fazla kullanılan kelimelerin ayır ediciliği azdır. Buna karşın çok az kullanılan kelimeler de gereksiz yere boyut sayısını arttırabilirler. Bu nedenlerle kullanıcı metinlerdeki tüm kelimeler yerine metinlerdeki geçiş sayısı belirli bir aralıkta olan kelimeleri kullanmak isterse bu filtre özelliğinden faydalanabilir. Kelimeler kullanıcının belirlediği minimum ve maksimum geçiş sayılarına göre filtrelenir. Bu özellik seçilirse filtreleme işlemi diğer tüm özellik gruplarındaki işlemlerden önce yapıldığı için kelimeleri kullanan tüm özellik gruplarının hesaplanmasına etki eder. Diğer bir deyişle, kelimeleri kullanan tüm özellik gruplarında sadece filtrelenen kelimeler kullanılır.

2.10 Saklı Anlam İndeksleme

Literatürde en yaygın kullanıma sahip özelliklerden biridir. Doküman matrisi (A) üzerinde Tekil Değer Ayrıştırma (Singular Value Decomposition) yaparak metinlerin ifadesinde kullanılan boyut sayısını

azaltır [6]. Bunun için A matrisi aşağıdaki eşitliğe göre özdeğer-özvektör çarpanlarına ayrılır.

$$A_{m\times n} = U_{m\times r}.S_{r\times r}.V^{T}_{n\times r}$$

Bu işlem sonunda S matrisinde, özvektörlerin özdeğerleri diyagonalde büyükten küçüğe sıralanır. Kullanıcının seçtiği boyut (*k*) kadarı işleme alınarak, A matrisi m*n'lik bir matristen (m→metin sayısı, n→ farklı kelime sayısı), m*k'lık (k→ metinlerin yeni boyut sayısı) bir matrise dönüşmüş olur.
Bu gruptaki özellikler *arff* dosyasına "LSI" öneki eklenerek yazılmaktadır. Tekil Değer Ayrıştırma için LingPipe [7] kullanılmıştır.

2.11 Anlamsal Uzay

Bu özellik grubunda metinlerin anlamsal bir uzaydaki koordinatları bulunur [8]. Metinlerin koordinatları içerdikleri kelimelerin koordinatlarının ortalaması alınarak bulunur. Harris [9], iki kelimenin birlikte geçtiği doküman / cümle / çerçeve sayısının iki kelimenin anlamca benzerliğiyle doğru orantılı olduğunu öne sürmüştür. Bu yaklaşım kullanılarak kelimelerin birbirlerine anlamca benzerliklerine göre konumlandırıldıkları anlamsal uzaydaki koordinatları Birlikte Geçme Matrisi kullanılarak bulunmaktadır. Benzerlikler matrisinden Çok Boyutlu Ölçekleme (Multi Dimensional Scaling) [10] ile koordinatlar elde edilmektedir. Kullanıcı, benzerliklerin ölçülmesinde 5 farklı şekilde oluşturulmuş Birlikte Geçme Matrisi kullanabilmektedir.
Metinlerin koordinatları arff dosyasına, kelimelerin koordinatları otomatik oluşturulan "kelimeMDS.txt" dosyasına yazılmaktadır. Bulunan kelime dolayısıyla metin koordinatlarının boyutu da, diğer bir deyişle kelimelerin ve metinlerin kaç boyutlu bir uzayda ifade edilecekleri, kullanıcı tarafından belirlenebilmektedir. Bu gruptaki özellikler arff dosyasına "MDS" öneki eklenerek yazılmaktadır. Çok Boyutlu Ölçekleme için MDSJ Kütüphanesi [11] kullanılmıştır.

3 Sonuç Dosyalarının Formatı (.arff)

Yazılım, özellik çıkarım işlemini sonunda kullanıcının seçtiği özelliklere göre farklı türde ve sayıda dosyalar üretmektedir. Üretilecek dosyaların yerleri ve adları ara yüzden belirlenebilmektedir. Ancak programın temel çıktısı arff formatındaki dosyadır. Metne ait kullanıcının seçtiği tüm özellikler bu dosyada yer almaktadır. Dosyanın her bir satırında her bir metne ait özellikler bulunmaktadır. Şekil 2'de örnek bir arff dosyası verilmiştir.

```
@ATTRIBUTE kelime_basina_harf_sayisi NUMERIC
@ATTRIBUTE cumle_basina_kelime_sayisi  NUMERIC
@ATTRIBUTE yazari {omer_seyfettin, peyami_sefa}
@DATA
5.1 3.5 omer_seyfettin
4.9 6.0 peyami_sefa
4.7 7.2 omer_seyfettin
4.6 4.1 peyami_sefa
5.0 9.6 peyami_sefa
```

Şekil 2: Örnek arff dosyası.

Şekil 2'deki örnek arff dosyasında 5 adet metne ait 2'şer özellik ve metinlerin ait oldukları kategoriler verilmiştir. Ömer Seyfettin'e ait 2 metne, Peyami Sefa'ya ait 3 metne ilişkin kelime ve harf sayılarıyla ilgili özellikler belirlenmiştir. Öncelikle metinlerin özelliklerinin tanıtımı yapılmış daha sonra @DATA satırından başlayarak metinlerin bu özelliklere ait değerleri verilmiştir. Her bir satırda bir metne ait özellikler verilmiştir. Metinlerin son özellikleri metnin yazarıdır (metnin ait olduğu sınıf/kategori). Metinlerin kategori bilgileri, içinde bulundukları klasörlerin isimleriyle belirlenmektedir. Programın çıktısı olan arff formatındaki dosyalarla kümeleme, sınıflandırma işlemleri yapabilmek için WEKA yazılımını kullanabilirsiniz.
WEKA yazılımını http://www.cs.waikato.ac.nz/ml/weka/ adresinden indirebilirsiniz.

Referanslar:

[1] http://code.google.com/p/zemberek/

[2] Bekkerman R., Ran El-Yaniv, Naftali T., Yoad W., "Distributional Word Clusters vs. Words for Text Categorization", *Journal of Machine Learning Research*, *pp.1-48, 2002.*

[3] Ethem Alpaydın, "*Introduction to Machine Learning*", The MIT Press, p. 146-148, 2004.

[4] Alpaydın, p. 135-139.

[5] Alpaydın, p. 282-283.

[6] B. Özel, "Küresel k-Ortalamalı Gruplama Yöntemi ile Metinlerin ve Terimlerin Saklı Anlam İndekslenmeleri", *ASYU, Istanbul, Turkey, p. 223-227, 2004.*

[7] Alias-i. 2008. LingPipe 3.8.1. http://alias-i.com/lingpipe
[8] M.Fatih Amasyalı, Aytunç Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", *SIU 2009.*
[9] Haris Zelig S., "*Mathematical structures of language*", Wiley, pp.12, 1968.
[10] Alpaydın, p. 121-124.
[11] Multidimensional Scaling for Java, University of Konstanz, Department of Computer & Information Science, Algorithmics Group, http://www.inf.uni-konstanz.de/algo/software/mdsj/