Sosyal Medya Verilerinin Ön İşlenmesi

Sinan KUL, Üstün ÖZEN
Atatürk Üniversitesi, Yönetim Bilişim Sistemleri Bölümü

Özet

Sosyal yazı dilinin doğasındaki çarpıklık, duygu analizi öncesinde yürütülen doğal dil işleme çalışmalarını zorunlu kılmaktadır. Bu çalışma ile sosyal medya verilerinin analiz öncesi ön işlenebilmesi için kullanılabilecek bir uygulama geliştirilmesi amaçlanmıştır. Kelime veritabanları bir araya getirildi.
Test verileri için Twitter verilerinin kullanıldığı bu çalışmada, metnin kelimelerine ve hecelerine ayrılması ve kelime hatalarının düzeltilmesi için algoritmalar geliştirildi.
Kelimelerin kök ve eklerine ayrıştırılması için de Zemberek kütüphanesinden faydalanıldı.

Giriş

Kelimenin doğru yazıldığının tespiti için sözlükte arattırılması yöntemi tek başına yeterli değil çünkü sondan eklemeli dil yapısındaki Türkçe kelimeler ekli halleriyle sözlüğe eklenmek istendiğinde sözlüğü çok büyütecektir. Örneğin "Oku" kelimesine eklenen sadece 3 yapım ekiyle 1461211 adet kelime türetilebilmektedir [1].
Kelimeler, bu yüzden kök halleriyle sözlükte arattırılabilir. Kelime köküne ulaşmak için de zemberek kütüphanesinden faydalanılabilmektedir [2].

Hatalı Kelime Durumları:

- Bitişik Kelime
- Kısaltma
- Bölgesel ağız farklılıkları
- Alıntı
- Eksik harf
- Fazla harf
- Yanlış harf sırası
- Yanlış harf (tombul parmak)

Hatalı Hece Durumları:

- 3 sessiz harf ardışık bulunan hece
- Kelime başında olmayan ve sesli harf ile başlayan hece
- Birden fazla sessiz harf ile başlayan hece
- 5 harfli hece
- İki aynı harf bulunan hece
- Sessiz harf ile biten heceden sonra sessiz harf ile başlayan bir hece gelirse ikinci hecede "ğ" bulunması
- Hece sonu iki sessiz harf ile bitiyorsa:
  - Sondan bir önceki harf, "l" iken son harf, "ç", "k", "p" veya "t" den biri değilse.
  - Sondan bir önceki harf, "r" iken son harf, "ç", "k", "p", "s" veya "t" den biri değilse.
  - Sondan bir önceki harf, "s" veya "ş" iken son harf, "t" değilse.

Materyal ve Yöntem

Kelime tablosu oluşturulurken, Zemberek veri tabanındaki 1140253 adet kelimeye, bir kaynaktan [3] 62784 adet kelime, Türk Dil Kurumundan [4] 72300 adet kelime, Wikipedia'dan [5] 63813 kelime alındı. Türkçede sıklıkla kullanılan ve duygu ifade etmeyen kelimeler, durak kelimeleri (stop words) [6], NVİ (Nüfus ve Vergi İşleri)'den il ve ilçe adları [7], 7119 adet özel isim [8] ve kelime tablosu birleştirilerek kelime veri tabanı oluşturuldu.
Tweet metni boşluk karakterlerinden ve noktalama işaretlerinden kelimelere ayrıştırıldı
Kelimeler Zemberek kütüphanesi vasıtasıyla ek ve köklerine ayrıldı.

```
Zemberek zemberek= new Zemberek(new TurkiyeTurkcesi());
kok = zemberek.kelimeCozumle(kelime)[0].kok().icerik();
```

Şekil1. Kelime veritabanının derlenmesi

Kelime, veri tabanıyla eşleştirilmeye çalışılırdı:

```
update tk set Duzeltilmis=s.Kelime, Sozlukte=1
from
    TweetKelimeler tk
        inner join sozluk s on s.kelime=tk.kelime
where
    sozlukte is null
```

Tombul parmak hatası için kelimedeki her bir harf için klavyedeki komşu tuşa basılma ihtimali ile sözlükte tarandı.
Aynı mesafede bulunan kelimeler içinden en sık kullanılan kelime kabul edildi.

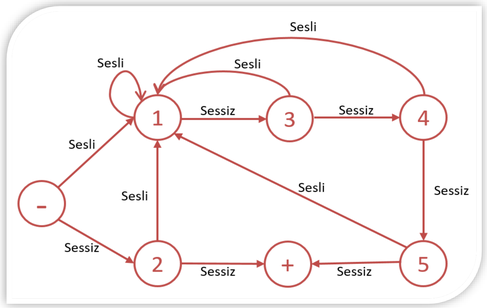
$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Sonuçlar

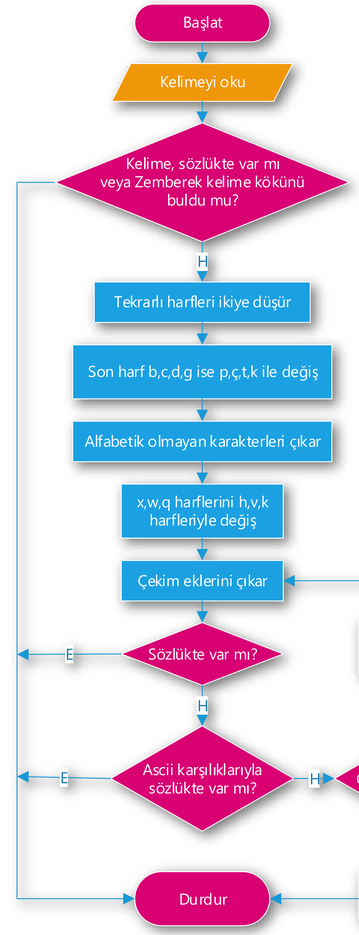
Türkçe'nin sondan eklemeli dil yapısı kelime veritabanı oluşturulabilmesini zorlaştırmaktadır. Yapılan bu çalışma ile yaygın kullanılan Türkçe sözlük tabloları birleştirilmiştir.
Ön işlem aşamasında işlem adımları için algoritma kurulmuş; TSQL ve Python dillerinde kodlanmış ve test edilmiştir.
Beklenenin aksine TSQL dili ilişkisel veri tabanı yönetim sisteminin sorgulama kabiliyetiyle daha başarılı sonuçlar vermiştir.
1591598 adet kelime hecelerine ayrılabilirken, 49561 adet kelime hecelerine ayrılamamıştır.
Çalışma kapsamında yalnızca Türkçe tweetler analiz edilmiştir. Kısaltma kullanımıyla ilgili otomatik düzeltme becerisi, yanlış yazılan kelimelerin düzeltilebilmesi yetersiz kaldığı için eşleştirme tablosunun oluşturulmasını sağlayan ve kullanıcı hatalarından öğrenen bir modül önerilmiş ve geliştirilmiştir.

Tablo 1. Türkçe Hece Yapıları

Hece Yapısı	Örnek
1 Sesli harf	A
1 Sesli 1 sessiz harf	Ar
1 Sesli 2 sessiz harf	Art
1 Sessiz 1 sesli harf	Ba
1 Sessiz 1 sesli 1 sessiz harf	Bay
1 Sessiz 1 sesli 2 sessiz harf	Kork



Şekil 2. Kelimenin hatalı olma durumu (SDM)



Şekil 3.kelime düzeltme algoritması

Gelecek Çalışmalar

Farklı ağız ve lehçeler için dönüşüm uygulamasının geliştirilmesi, diğer dillerdeki sözlükler dahil edilerek kelimenin alıntı olduğunun anlaşılması planlanmaktadır.

İletişim Bilgileri

Sinan KUL
Atatürk Üniversitesi
Yönetim Bilişim Sistemleri Bölümü
Email: sinan.kul@atauni.edu.tr

Kaynakça

1. Oflazer, K., & Bozşahin, H. C. (1994). "Türkçe ve Doğal Dil İşleme" [Bildiri]. Paper presented at the Proc. of Turkish Informatics Society TBD'94, Ankara.
2. Meral, M., & Diri, B. (2014). "Twitter Üzerinde Duygu Analizi" [Bildiri]. Paper presented at the IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kongresi, Trabzon.
3. Serdar, İ. (2013). TDK İmla Kılavuzu. Retrieved from http://www.excelgurusu.com/tdk-imla-kilavuzu/
4. TDK. (2017). Türk Dil Kurumu. Retrieved from http://www.tdk.gov.tr/
5. Vikisözlük. (2017). Viki Sözlük. Retrieved from https://tr.wiktionary.org/wiki/Anasayfa
6. Sarıgül, K. (2009). Türkçe Etkisiz Kelimeler (Stop Words) Listesi 1.1. Retrieved from http://www.turkceogretimi.com/Genel-Konular/article/541-turkce-etkisiz-kelimeler-stop-words-listesi-11/35#
7. Nvi. (2017). İl ve İlçe Listesi. Retrieved from www.nvi.gov.tr/Publishingimages/Pages/il-ilce-kod-tablosu/IL_ILCE_LISTESI.xls&usg=AFQjCNGhNZO1se7w2SJ6VRrthACW67NJSw
8. Dilipak, A., Dilipak, A., & Meriç, N. (2016). Türk Evim. Retrieved from www.turkevim.com/_ld/0/90_ANSIKLOPEDIK_IS.doc

İletişim Bilgileri

Üstün ÖZEN
Atatürk Üniversitesi
Yönetim Bilişim Sistemleri Bölümü
Email: uozen@atauni.edu.tr