BİTİRME PROJESİ

Haftalık Rapor – 01.04.2022

1 NİSAN 2022

KIRIKKALE ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ İÖ

AHMET MUNGAN – 160255081

İÇİNDEKİLER

ÖZET

Yemeksepeti’nden elde edilen veri seti, metin madenciliği süreçlerinden geçirilmiştir.

METİN MADENCİLİĞİ SÜRECİ

Metin Koleksiyonun Belirlenmesi

Metin madenciliği sürecinde Yemeksepeti'nin izin ve desteği sayesinde hedef metin veri seti seçilmiştir. Metin koleksiyonun seçilmesi aşaması geçen hafta elde edilenler ile gerçekleşmiştir. Elde edilen yorumlar kullanılarak büyük bir veri seti elde edilecek ve yorumlardan çıkarımlar, özetlemeler, analizler yapılacaktır.

Tekrarlayan Verilerin Temizlenmesi

Veri setinde tekrarlayan verilerin bulunması mümkün değildir. Çünkü iki farklı kişi aynı yorumu yapmış olsa bile tekrar eden bir veri olduğunu söylemek güçtür. Sonuçta iki farklı kişi aynı yorumu yapabilir ve restorana ait aynı duyguları besliyor olabilir. Ayrıca bir kişi aynı restorana sürekli yorum yapabiliyor ve aynı şeyleri yazıyor olabilir. Fakat ön incelemelere göre genelde önceden sipariş verenlerin yeni verdikleri siparişe kıyasla restoranı eskiye nazaran beğendiğini/beğenmediğini yorumlarından anlamak mümkündür. Çok düşük ihtimal de olsa tekrarlayan veriler bu veri seti özelinde büyük bir problem değildir.

Küçük Harf Dönüşümü

Veri setinde fark edileceği üzere büyük – küçük harf karışık bir durumdadır. Temel programlama dili python olduğundan aynı harfin büyük ve küçük halinin ASCII değeri farklı olacağı için birlik sağlanması gerekmektedir. Genellikle küçük harf dönüşümü yapılır.

Code 1

```
>>> yorumlar
['Hamburger ekmeğinin kenarları kuruydu.',
 'On numara beş yıldız.',
 'Sosları koymamissiniz. Burger güzel.',
... ]
>>> yorumlar_kucuk_harf = []
>>> for yorum in yorumlar:
>>>     yorum = yorum.lower()
>>>     yorumlar_kucuk_harf.append(yorum)
>>> yorumlar = yorumlar_kucuk_harf
>>> yorumlar
```

```
['hamburger ekmeğinin kenarları kuruydu.',
 'on numara beş yıldız.',
 'sosları koymamissiniz. burger güzel.',
... ]
```

Code 1'de yorumlar kısmının küçük harfe dönüştüğü görülmektedir.

Noktalama İşaretlerinin Silinmesi

Noktalama işaretlerinin de silinmesi gerekir. Yorumlarda noktalama işaretlerinin önemi yerine göre çok büyük olsa da hatalı noktalama işareti kullanımı söz konusudur. Hiddet, şaşırma, sinirlenme, mutluluk gibi anlamlar çıkarılabileceği için noktalama işaretleri önemli olabilir. Fakat unutulmamalıdır ki bu bir duygu analizi değildir.

Code 2

```
>>> import string
>>> string.punctuation
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
>>> yorumlarGecici = []
>>> for yorum in yorumlar:
>>>     yeniYorum = ""
>>>     for harf in yorum:
>>>         if harf not in string.punctuation:
>>>             yeniYorum += harf
>>>     yorumlarGecici.append(yeniYorum)
>>> yorumlar = yorumlarGecici
>>> yorumlar
['hamburger ekmeğinin kenarları kuruydu',
 'on numara beş yıldız',
 'sosları koymamissiniz burger güzel',
... ]
```

Durak Kelimelerinin Belirlenmesi

Tokenization aslında işlemi bu aşamada gereklidir.

Code 3

```
>>> !pip install nltk
Requirement already satisfied: nltk in c:\users\ahmet\anaconda3\lib\site
-packages (3.6.5)
...
Requirement already satisfied: colorama in c:\users\ahmet\anaconda3\lib\
site-packages (from click->nltk) (0.4.4)
>>> import nltk
>>> from nltk.corpus import stopwords
>>> nltk.download("stopwords")
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ahmet\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
>>> sw = stopwords.words("turkish")
>>> sw
['acaba','ama','aslında','az','bazı','belki','biri','birkaç','birşey',
 'biz','bu','çok','çünkü','da','daha','de','defa','diye','eğer','en',
 'gibi','hem','hep','hepsi','her','hiç','için','ile','ise','kez', 'ki',
 'kim','mı','mu','mü','nasıl','ne','neden','nerde','nerede','nereye',
 'niçin','niye','o','sanki','şey','siz','şu','tüm','ve','veya','ya',
 'yani']
>>> len(sw)
53
```

Durak kelime içeren yorumların içerisinden bu durak kelimelerin çıkarılması anlamsız kelimeler oluşumuna neden olacaktır. Ayrıca durak kelimeler silinmeden önce tokenization işlemi yapılmazsa algoritmanın maliyeti çok yüksek olacaktır. Bu aşamada sadece durak kelimelerin belirlenmesi olacaktır. Bunun bir örneği Code 3'te görülebilir. NLTK kütüphanesinde hazır türkçe için durak kelimeler Code 3'te görülmektedir. Fakat bu kelimeler aslında yetersizdir. Bir de bu iş özelinde kullanıcıların yaptığı yorumlarda türkçe karakter kullanmama durumları var. Örneğin hazır durak kelimelerde "aslında" kelimesini bir kullanıcı "aslinda" olarak yazabilir. Bu sebeple durak kelimelerin listesine bu tarz kelimelerin eklenmesi daha doğru sonuçlar verecektir.

Code 4

```
>>> sw_eklenecek = ["herşey","her şey","her sey","hersey","herkez","herk
es","hah","heh","ney","cok","bazi","maalesef","malesef","hic","nasil",
"aslinda","birkac","icin","nicin","sey","ıse","ıle","bırı","belkı",
"dıye","eger","tum","cunku","cünkü","çunku","orda","orada","oradan",
"ordan","burda","burdan","burada","buradan","şurada","şuradan","şurda",
"şurdan","surada","suradan","surda","surdan","şurdaki","şuradaki",
"hayhay","evet"]
>>> for stopword in sw_eklenecek:
>>>     sw.append(stopword)
>>> len(sw)
102
```

Code 4’te eklenecek liste belirlenmiştir. Bu listeye sayılar yazı olarak ve nümerik string olarak da eklenebilir. Ayrıca bu liste uzuyabilir ve farklı durak kelimeler bulunabilir.

REFERANS VE KAYNAKÇA

[1] NLTK Kütüphanesi. Link için tıklayınız. (Güvenlidir.)

[2] String Kütüphanesi. Link için tıklayınız. (Güvenlidir.)

[3] Yemeksepeti kişisel verilerin korunumu. Link için tıklayınız. (Güvenlidir.)

EKLER

Bitirme Projesi 2’ye ait doküman, program kodu, haftalık rapor ve ek bilgilerin paylaşıldığı github linki için tıklayınız. (Güvenlidir.)