Makine Öğrenmesi Destekli Anlamsal Ürün Arama Motoru

BIL 496
İkinci Buluşma

Simge Sarıçayır

Proje Danışmanı: Fatih Erdoğan Sevilgen
Aralık 2019

Proje Tanımı

- Proje nedir?

E-ticaret siteleri için makine öğrenmesi destekli anlamsal ürün arama motoru.

- Günümüzde yaygınlaşan e-ticaret sitelerinde doğru ürün listeleme ile satışın arttırılması.
- Satın alınmak istenen ürünün arama sonucunda doğru sırada bulunamaması.
- Yapılan listelemede istenen ürünün yer almaması ya da anlamsal olarak alakasız şeylerin listelenmesi

Önceki Sunumda Yapılanlar

- Eğitilen Word2Vec modeli kullanılarak arama yapılan kelimelere karşılık modelde en yakın kelimeler bulundu ve ürün listesinden bu kelimeleri içeren ürünler listelendi.

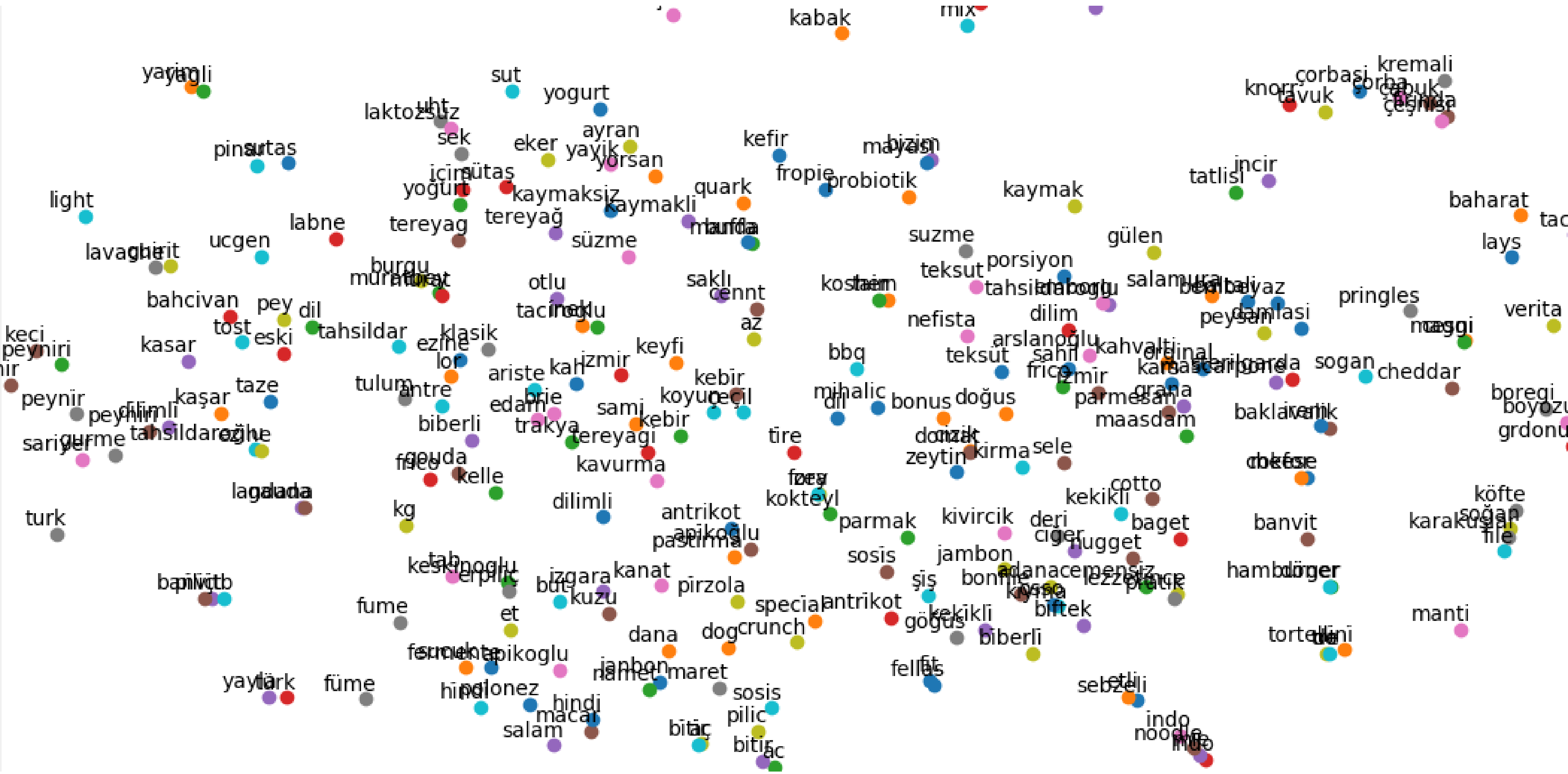
- Modelin iki kelime arasındaki benzerlik oranını vermesi ile ürün isimleri üzerinde bu benzerlik oranı kullanılarak sıralamada puanlama yapıldı.

- İki farklı e-ticaret sitesinden elde edilmiş olan ürün açıklama verileri ve arama motorunun gerçeklendiği ürün isimleri birleştirilerek başka bir Word2Vec modeli eğitildi.

- Oluşturulan diğer model ile aynı kelimeler üzerinde karşılaştırma yapıldığında ilk oluşturulan model daha iyi sonuç verdi.

Yapılanlar

- Ürün isimleri Google'da aratıldığında gelen sonuçlardan ilk 5 sitedeki bilgiler çekilerek daha geniş bir eğitim verisi elde edildi.
- Toplanan verilerin temizlenmesi ve ön işleme yapılmasından sonra bu veri kullanılarak Word2Vec Cbow modeli eğitildi.
- Model arama motoru geliştirilmesi için kullanıldı.
- Arama motoru için arama kelimesine karşılık ürün isimleri verisi bir sanal marketin log dosyasından çıkarıldı.
- Elde edilen veri hem test hem de makine öğrenmesi için kullanıldı.


kabak
mix
sut
yogurt
kremali
çorbasi
tavuk
knorr
laktozsuz
ayran
sek
eker
yayik
yorsan
kefir
probiotik
pinar
sutas
light
kaymaksiz
tereyağ
quark
fropie
kaymak
tatlisi
incir
baharat
labne
tereyag
kaymakli
suzme
gülen
lays
ucgen
süzme
burgu
otlu
sakli
teksut
porsiyon
pringles
verita
bahcivan
pey
dil
az
nefista
dilim
keci
kasar
tost
eski
tahsildar
klasik
ezine
keyfi
bbq
arslanoglu
sogan
cheddar
peyniri
lor
ariste
izmir
kebir
teksut
tulum
antre
taze
kaşar
mihalic
bonus
doğus
grana
brie
edam
sami
parmesan
peynir
biberli
trakya
tereyagi
maasdam
gurme
tire
sele
kirma
zeytin
gouda
kelle
kavurma
cotto
kekikli
köfte
turk
kg
dilimli
antrikot
kokteyl
parmak
kivircik
deri
baget
banvit
sosis
jambon
nugget
fume
kanat
izgara
pirzola
kuzu
special
antrikot
manti
et
dana
crunch
biberli
biftek
kekikli
füme
sosis
pilic
salam
hindi
macar
felles
indo

Yapılanlar

- Web sitelerinden veri çekerek eğitilen Word2Vec modeli kullanılarak arama motoru geliştirildi.
- Kelime vektörlerinin mesafesi azaldıkça anlamsal yakınlık oranı arttığı için anlamsal yakınlıkta ürünlerin listelenmesi kelime vektörleri ile sağlandı.
- Aranan kelime/kelimelerin vektördeki yeri ile ürün listesindeki her ürün arasındaki mesafeye bağlı olarak sıralama yapıldı.
- Aranan kelime/kelimelerin birebir eşleşme sağladığı ürünler üst sıraya taşındı. Bunun için kelimelerin kökleri de kontrol edildi.

Yapılanlar - Sonuçlar

Kullanıcıların arattığı kelimelerle satın almayı tercih ettikleri veriler kullanılarak arama motoru test edildi.
4134 tane ürün-arama kelimesi verisi kullanıldı.

Bulunma Sırası	Sayısı
1	1000
2	437
3	283
4	224
5	188
6	134
7	115
8	118
9	84
10	103

2687 tanesi ilk 10'da gelirken 1447 tanesi için ilk 10 sırada sonuç alınamadı.

Bu durumda 0.65 oranında başarı sağlandı.

Yapılanlar

- Arama motorunu test etmek için kullanılan veri ile makine öğrenmesi tekniklerinden sınıflandırma denendi.
- Sınıf olarak ürün isimleri , öğrenilecek veri için de arama kelimesinin Word2Vec modelinden elde edilen 300 boyutlu vektör kullanıldı.
- Verinin %20'si test , %80'i eğitim için kullanıldı.

Sınıflandırma Sonuçları

Bir sınıfın 10 veya 10'dan fazla üyesi varsa veri setine eklenerek eğitim gerçekleştirildi.

Algoritma	Doğruluk Oranı
Lojistik Regresyon(Logistic Regression)	0.88
En Yakın Komşu (K Nearest Neighbor)	0.83
Karar Ağaçları (Decision Trees)	0.87
Gaussian Naive Bayes	0.75
Destek Vektör Makineleri (Support Vector Machines)	0.79

Yapılması Planlananlar

- Kelime köklerini bulmak için daha iyi sonuçlar veren kütüphane ile test yapılacak.
- Yanlış yazım durumu için aratılan kelimeler Word2Vec modelinde yoksa Edit Distance algoritması ile modeldeki olası kelime için sonuçlar bulunacak.
- Sınıflandırma için veri toplanacak ve model daha çok veri ile tekrar eğitilecek.
- Sınıflandırıcı kullanılarak kullanıcı seçimleri sayesinde ürünlerin listelenmesinde daha iyi sıralama elde edilecek.

Başarı Kriterleri

- Arama kelimelerine karşılık listelenen ilk 10 ürünün test ve eğitim için kullanılan verilerdeki aynı arama kelimesine karşılık satın alınan ürünlerin en az 5'ini içermesi.(düzeltme : %90'ını içermesi)

- Kullanıcıdan toplanan verilere bakıldığında arama kelimesine karşılık satın alınan ürünlerin listelemedeki sıralarının toplamının ürünün sepete eklenme sayısına bölündüğünde 5'ten küçük bir değer elde edilmesi.

- Listelenen ilk 10 ürün adının içerdiği kelimelerden en az birinin arama kelimesi ile word2vec modelindeki benzerlik oranının %70'in üzerinde olması.

Kaynaklar

1. T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space", arxiv.org., September 2013.
2. A McCallum, K Nigam, J Rennine, K Seymore - IJCAI, 1999 – Citese