Makine Öğrenmesi Destekli Anlamsal Ürün Arama Motoru

BIL 496
Son Sunum

Simge Sarıçayır

Proje Danışmanı: Fatih Erdoğan Sevilgen
Ocak 2020

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

Proje Tanıtımı

- Bu projede kelime vektörleri ve makine öğrenmesi teknikleri kullanılarak anlamsal yakınlıkta ürünleri listeleyen bir sanal market arama motoru geliştirilmiştir.

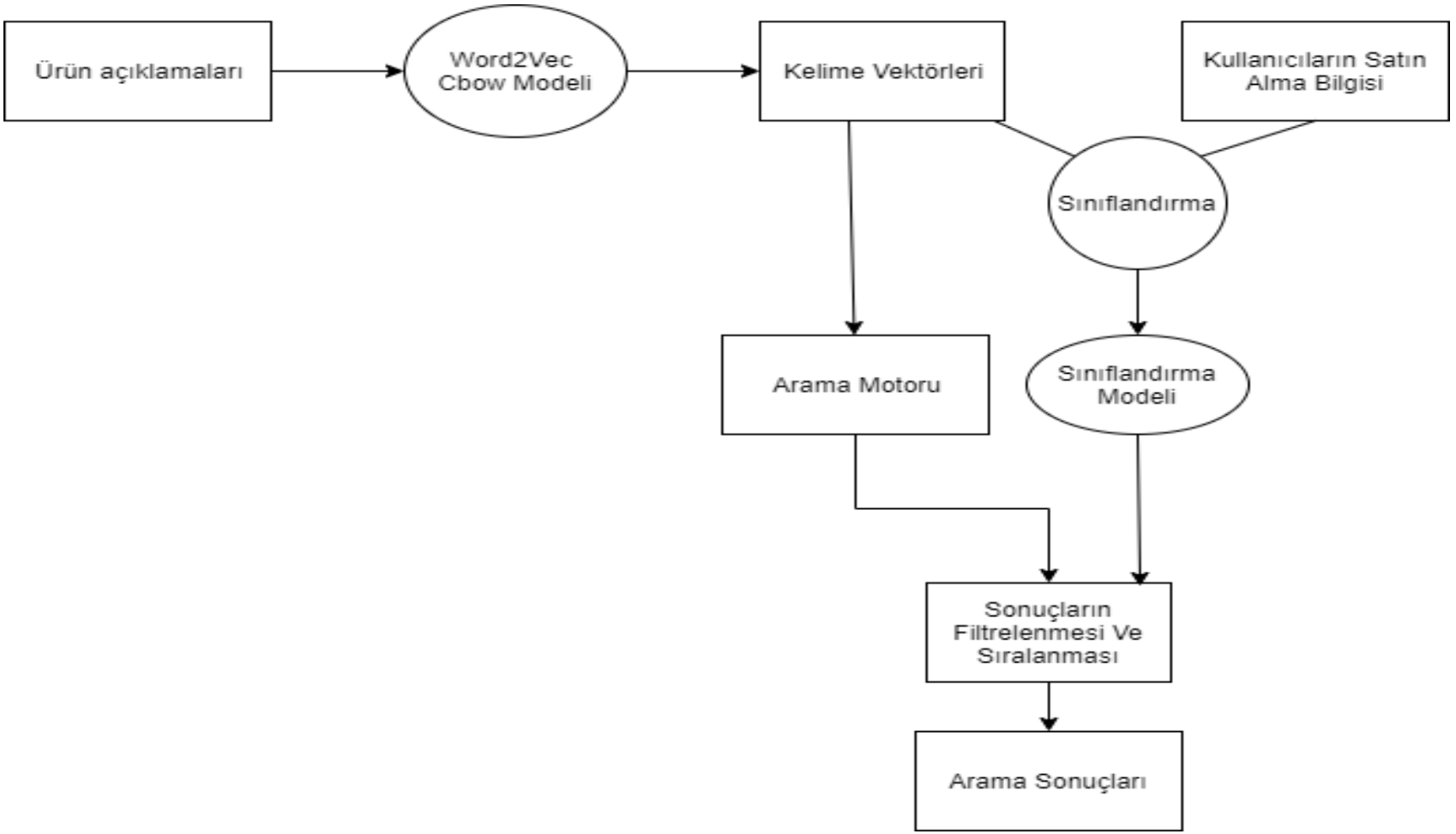
- Aranan kelime/kelimelere karşılık listelenen ürünlerin birbirleri ve arama için kullanılan kelimelerle anlamsal yakınlıkta olması projenin genel amacı içerisinde yer almaktadır

- Projede kullanıcıların bir ürünü satın almak için kullandıkları arama kelimelerinin makine öğrenmesi için kullanılması ile kullanıcı tercihlerine dayanarak ürün sıralaması yapılmıştır

Bu proje sayesinde:

Arama motorunu kullanan kişi arattığı kelimelerle bulmak istediği ürünü üst sıralarda görebilecek.
Satın almak istenen ürünün satışı olmaması durumunda arama yapılan kelimelerle anlamsal yakınlıktaki ürünlerin bulunması sağlanacak.
Kullanıcıların satın alma geçmişinden öğrenen bir sistem olması sebebiyle listelenen sonuçlar yalnızca anlamsal yakınlık içeren ürünlerden oluşmayacak, aynı zamanda benzer deneyimlere göre ürün listelemesi sağlanmış olacak

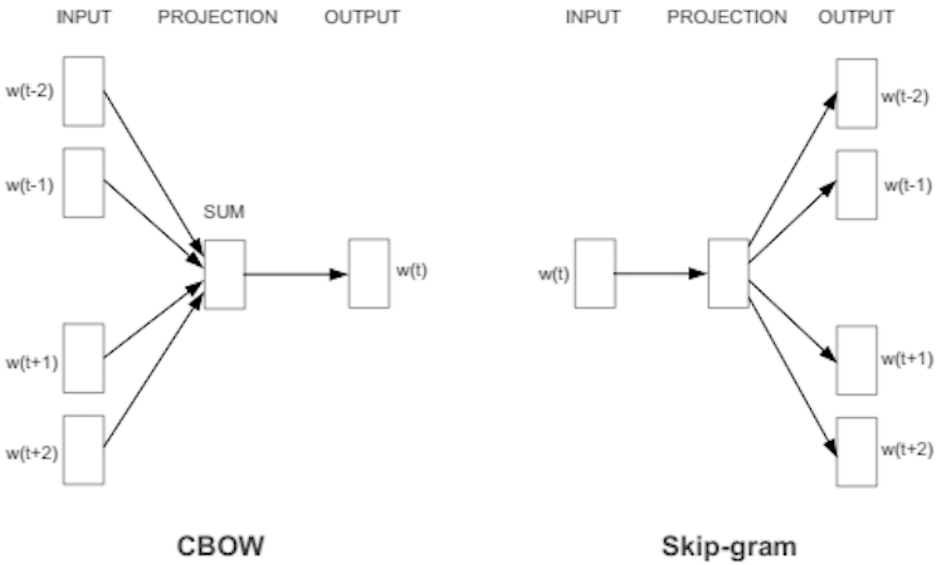
Sistem Mimarisi



Kelime Vektörleri

Kelime vektörleri yöntemi, kelimeleri n boyutlu bir uzayda birer vektör olarak temsil etmek ve bu yol ile kelimeler arası uzaklıkları hesaplayarak aralarındaki anlamsal benzerliği tespit etmek amacıyla kullanılmıştır.

Word2Vec ve Cbow



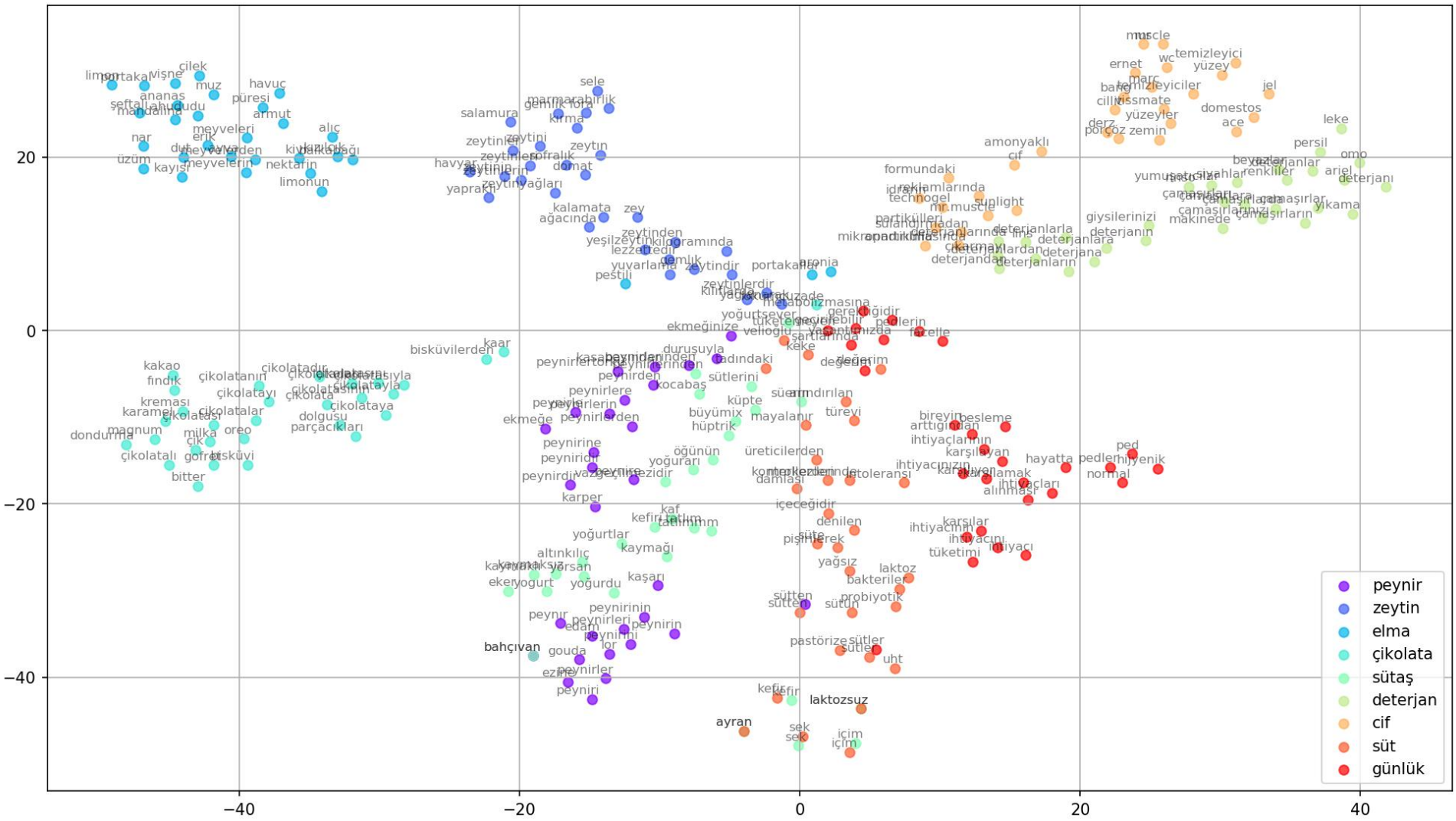
- Cbow bir kelimeyi tahmin etmek için kelimeyi çevreleyen bir bağlamı kullanırken, Skip-gram sabit bir pencere boyutuna sahip kelimeleri çevreleyerek kelimeyi tahmin etmeye çalışır.

- Proje geliştirilirken Cbow modeli kullanılmıştır. Bunun sebebi ise tahmin edilmek istenen kelimenin eğitim yapılan cümleler içerisindeki kelimeler ile ilişkilendirilmek istenmesidir.

Kullanılan Veri

- 6528 adet ürün adı.
- 25536 adet kullanıcıların arattığı kelimeye karşılık satın aldığı ürün bilgisi.
- Word2Vec modelini projenin amacına uygun şekilde eğitmek için proje geliştirilerken kullanılan ürünler ile alakalı içeriğe sahip veri kullanılması gerekmektedir.
- Bunun için googlesearch ve beautifulsoup kütüphanesi kullanılarak ürünlerin google'da aratılması ile gelen ilk 5 sayfadan içerik toplandı.

Eğitilmiş Word2Vec Modelinde Örnek Kümelenme




çilek
limon
portakal
vişne
muz
havuç
ananas
püresi
şeftali
kavun
mandalina
armut
meyveleri
alıç
erik
nar
dut
meyvelerden
kivi
üzüm
kayısı
meyvelerin
nektarin
limonun
biski
kakao
fındık
çikolatanın
çikolatadır
çikolatayı
çikolatayla
kreması
çikolata
karamel
çikolatalar
çikolataya
şikolatası
dolgusu
magnum
milka
oreo
parçacıkları
dondurma
çik
çikolatalı
gofret
bisküvi
bitter

Arama Kelimelerinin Kontrolü

- İlk önce fst yazılarak kelimelerin türkçe-ingilizce karakter uyuşmazlığı kontrol edildi.
- Karakter değişimi yapılan kelime de sözcük dağarcığında yoksa ED algoritması ile en az karakter değişimi ile uygun kelime arandı.
- Bu kontrol yapılırken birden fazla kelime olması durumunda yazım yanlışı düzeltilen kelimenin diğer kelime ya da kelimelerle benzerlik oranı kontrol edildi.
- Arama için kullanılan kelimelerin kökleri de kontrol edildi. Aynı kontrol ürün içerisindeki kelimenin kökünün arama kelimesi ile eşleşmesi durumu için de yapılarak kelime yapısından kaynaklı sorunların gidirelmesi için eklendi

benzerlik değeri büyükten küçüğe doğru ürünler sıralandı.

Kelime Vektörleri İle Arama Motoru

Aranan kelimelerin Word2Vec modelindeki vektörleri ile ürün içindeki kelimelerin vektörlerinin bütün olarak karşılaştırılması ile puanlama yapıldı.
Aranan kelimeler ürün içerisinde birebir eşleşiyorsa fazladan puan vererek sıralamada üst sıraya çıkması sağlandı

Makine Öğrenmesi Yöntemleri

- Makine öğrenmesi tekniklerinden sınıflandırma kullanıldı.
- Ürün isimleri sınıf olarak, özellik olarak ise bu ürünü aramak için kullanılan kelimenin Word2Vec modelinde temsil edilen vektörünün değerleri kullanılmıştır.

EKER YOGURT KAYMAKLI 500GR	yoğurt
LİPTON BARDAK POŞET ÇAY EARL GREY 25'Lİ 50GR	çay
COCA COLA 250MLX6 ŞİŞE	cola
LİPTON İCE TEA 330M ŞEFTALİ	tea
PİKNİK 110-P JAPON KÜRDAN 400'LÜ	Kürdan
PAREX STREÇ FİLM 100MT	streç
YUDUM AYÇİÇEK 1 LT	ayçiçek
TEREMYAĞ PAKET MARGARİN 250 GR	terem
TADIM SULTANİYE ÜZÜM 140GR	sultani
EKER YOĞURT SÜZME 400GR	yoğurt
EKER YOĞURT SÜZME 400GR	eker süzme
KOMİLİ SIZMA ZEYTİNYAĞI 2 LT	zeytinyağı

İndeks1	İndeks2	...	İndeks300	Sınıf
0.397183	0.435034	...	-0.245028	AKMİNA 200ML SODA SADE
-0.507915	0.245336	...	0.050537	YUDUM AYÇİÇEK 5 LT
0.206975	0.333119	...	0.054616	NAMET DANA MACAR SALAM 150GR
0.472770	0.255918	...	-0.139991	7DAYS KRUVASAN KAKAO KREMALI 300GR

Test Sonuçları

Bir sınıfa ait üye sayısı en az 10 ise eğitime katıldı.

ALGORİTMA	DOĞRULUK ORANI
LOJİSTİK REGRESYON	0.88
EN YAKIN KOMŞU	0.83
KARAR AĞAÇLARI	0.87
GAUSSİAN NAIVE BAYES	0.75
DESTEK VEKTÖR MAKİNELERİ	0.79

Tablo 2. İlk Sınıflandırma Sonuçları

Boyut Azaltma-PCA

Lojistik regresyon modeli ile 300 öznitelik kullanarak eğitilen modelin doğruluk oranı ve f1-puanı aynı veriseti kullanılarak PCA uygulanan modelin doğruluk oranı ve f1-puanı ile karşılaştırıldı.

Farklı boyutlarda yapılan testlerde en yakın sonuç 300 öznitelik sayısının 60'a indirilmesi ile elde edildi.

Verinin Çoğaltılması ve Sınıflandırma Modeli

Daha önce eğitilen 4200 sınıf- arama kelimesi verisi verisetinin temizlenmesi, veri çıkarımı ve çeşitli anket yöntemleri ile arttırılarak 25536'ya çıkarıldı.

Arama kelimeleri kelime vektörlerine dönüştürülerek sınıf ile ilişkilendiridi. Bu sırada bir ürün için birden fazla kelime varsa bu sınıf için kelimelerin her biri tek tek ilişkilendirildi.

Verinin Çoğaltılması ve Sınıflandırma Modeli

- Bu veriler filtrelendi ve eğer bir sınıfa ait üye sayısı en az 10 ise eğitime alındı.
- Bu sınıfların her birinden eşit miktarda (10) üye alındı.
- Burada amaç artan sınıf sayısı ile öğrenmenin zorlaşması yanında dengesiz veri problemini ortadan kaldırmaktır.
- Verinin uygun boyutta oluşturulmasından sonra PCA ile boyut azaltıldı, eğitim 1561 sınıf ve 26711 kelime-sınıf örneği ile yapıldı.

Makine Öğrenmesi Eklenerek Arama Motoru

	GNB	LR	LDA
P	0.43	0.67	0.60
R	0.48	0.70	0.62
A	0.57	0.81	0.77
F1	0.45	0.69	0.61

- Daha önce yalnızca Word2Vec kelime vektörlerinin anlamsal yakınlık bilgisi sağlaması ile geliştirilen arama motoru sonuçlarına ek olarak lojistik regresyon modelinden elde edilen sonuçlar eklendi.

- Daha sonra lojistik regresyon modelinin tahmin sonuçları sırayı bozmadan alındı. Birden fazla kelime aratılması durumunda tahmin sonuçlarındaki aynı sınıfa ait tahmin puanları toplanarak üst sıraya alınması sağlandı.

Başarı Kriterlerinin Değerlendirilmesi

1. 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.

soda {'SIRMA SODA 200CC SADE DOĞAL.',
'KIZILAY SODA 200 ML.',
'ÖZKAYNAK SODA 200 ML',
'BEYPAZARI SADE SODA 200ML',
'DAMLA SODA 200ML SADE'}

	WORD2VEC ILE	WORD2VEC VE SINIFLANDIRMA ILE
TÜM TEST VERISI	0.69	0.81
ÜYE SAYISI >1	0.57	0.83

Başarı Kriterlerinin Değerlendirilmesi

2. 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.

Birinci başarı kriterinin testlerine göre bulunanan ürünlerin sıralamadaki yerlerinin ortalaması alındığında aşağıdaki sonuçlar elde edilmiştir.

	BULUNMA SIRASI ORTALAMASI
TEST VERISININ TAMAMI	1.66
ÜYE SAYISI >1	2.00

Başarı Kriterlerinin Değerlendirilmesi

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

Bu kriter için 100 adet arama kelimesi ile yapılan arama sonucu test edildi. Her arama için listelenen ilk 10 ürünün her birinin içerdiği kelimelerin vektörleri ile arama kelimesinin vektörü arasındaki benzerlik oranı kontol edildi.

100 arama kelimesi için 77 tanesinin sonucunda istenilen başarı kriterine ulaşıldı. Ulaşılamayan 33 ürün ise listelenen 10 ürünün tamamı için değil ancak bu değere yakın değerler ile başarı kriterine yaklaşmıştır. %77 oranında kritere ulaşıldı

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
[3] Metin B., *Comparison of Modeling Time of Word Vector Methods*
[4] Mohamed Aly., *Survey on Multiclass Classification Methods,* November 2005
[5] Maja Pohar1 , Mateja Blas2 , and Sandra Turk, *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study,* Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161.