SuDer Türkçe Haber Derlemlerinin Doküman Sınıflandırması

Mehmet Umut Şen

Berrin Yanıkoğlu

VERİM | Veri Analitiği Araştırma ve Uygulama Merkezi

Sabancı Üniversitesi

VPA

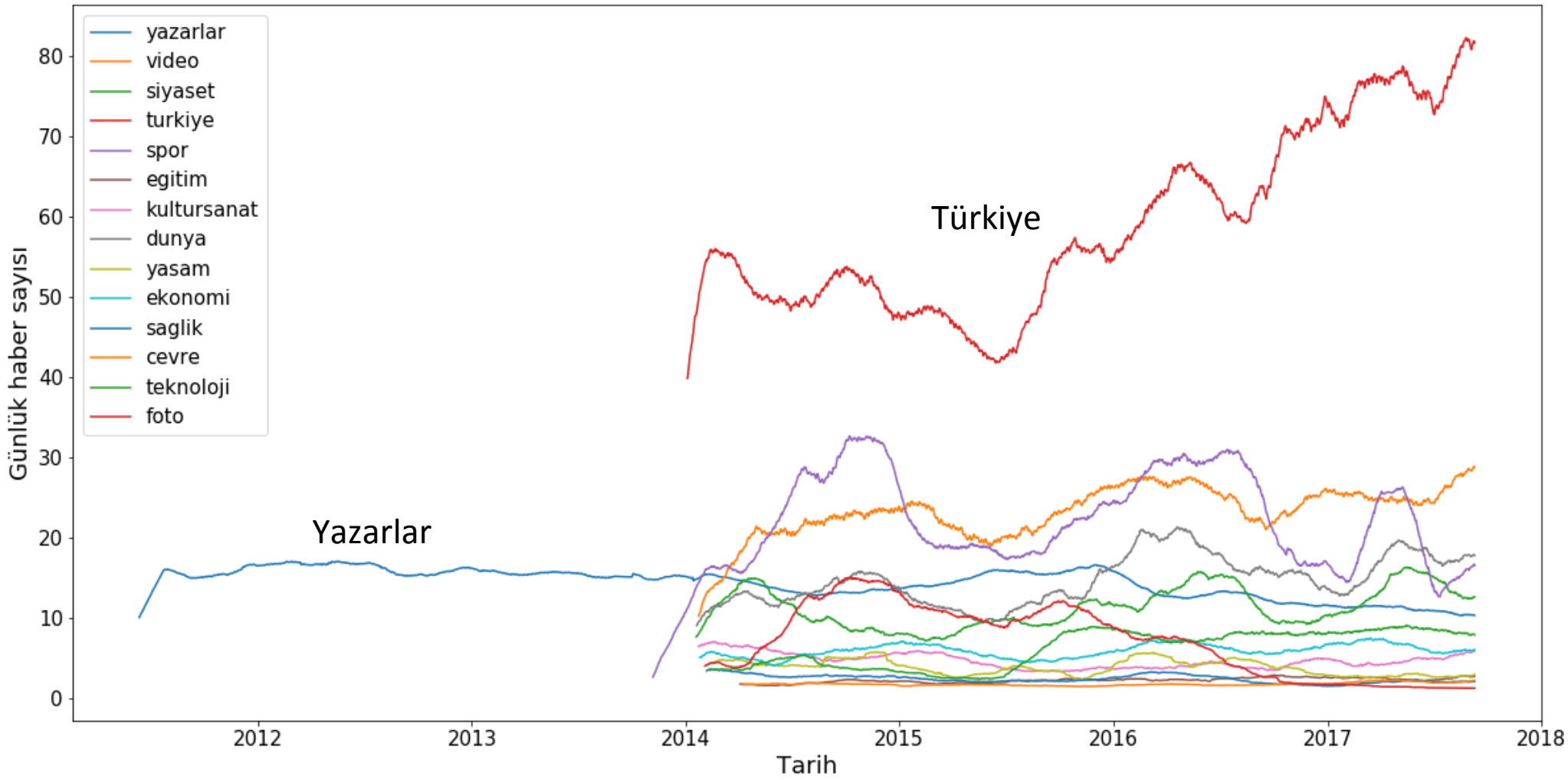
Motivasyon

- Doküman Sınıflandırma
- Kelime Temsilleri
- Büyük veri ihtiyacı

SuDer Haber Derlemleri

- www.cumhuriyet.com.tr ve www.sabah.com.tr
  - Haber linklerinin bulunması
  - Linklerden html sayfalarının indirilmesi
  - Sayfalardan metin, başlık, tarih ve etiket bilgilerinin çıkarılması
- Haberler, köşe yazıları, resim ve video galerileri
- Sabah → 2010 Ocak ile 2017 Temmuz arasında 426,000 web sayfası
- Cumhuriyet → 2017 Eylül tarihine kadar 463,000 sayfa
- 10 kelimeden az olan metinler ve az sayfalı bazı kategoriler elendi.

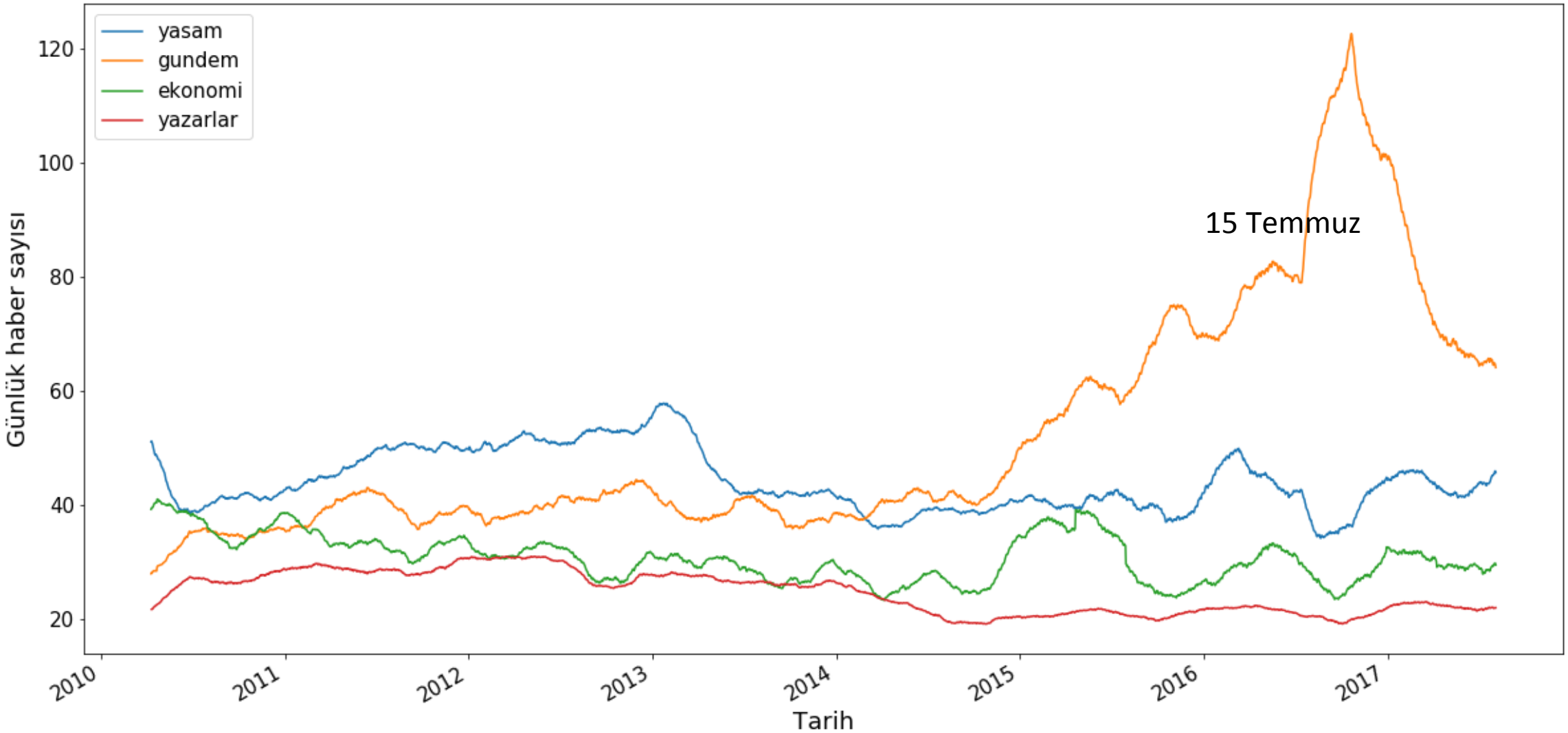
Kategoriler - Cumhuriyet



Kategoriler - Cumhuriyet

Kategori	Doküman Sayısı	Eğitim Dok. Sayısı	Test Dok. Sayısı	Toplam Kelime Say.	Ort. Kelime Sayısı
Türkiye	84,741	56,140	28,524	22,829,220	269.39
yazarlar	33,835	29,694	4,141	16,663,717	492.49
video	33,409	23,686	9,723	2,007,691	60.09
spor	31,396	24,627	6,730	7,240,974	230.63
dünya	21,005	14,684	6,152	4,416,708	210.26
siyaset	15,969	11,274	4,686	6,409,811	401.39
foto	14,302	9,729	110	248,871	17.4
ekonomi	8,187	5,811	2,356	2,520,473	307.86
teknoloji	7,913	5,089	2,810	1,734,268	219.16
kültür-sanat	6,506	4,680	1,806	2,664,020	409.47
yaşam	4,833	3,931	886	918,754	190.1
sağlık	2,573	2,047	514	863,208	335.48
eğitim	2,380	1,544	805	744,396	312.77
çevre	1,735	1,081	607	477,811	275.39
Toplam	**268,784**	194,017	69,850	**69,739,922**	259.46

Kategoriler - Sabah



Kategoriler - Sabah

Kategori	Dok. Sayısı	Eğitim Dok. Say.	Test Dok. Say.	Toplam Kelime Say.	Ort. Kel. Say
Gündem	143,842	117,019	26,823	35,749,880	248.54
Yaşam	123,086	108,202	14,884	22,878,732	180.86
Ekonomi	85,485	75,512	9,973	22,261,600	247.38
yazarlar	68,100	60,683	7,417	16,335,364	239.87
Toplam	**420,513**	361,416	59,097	**95,494,110**	227.09

TF-TDF öznitelikleri ve DVM sınıflandırıcısı

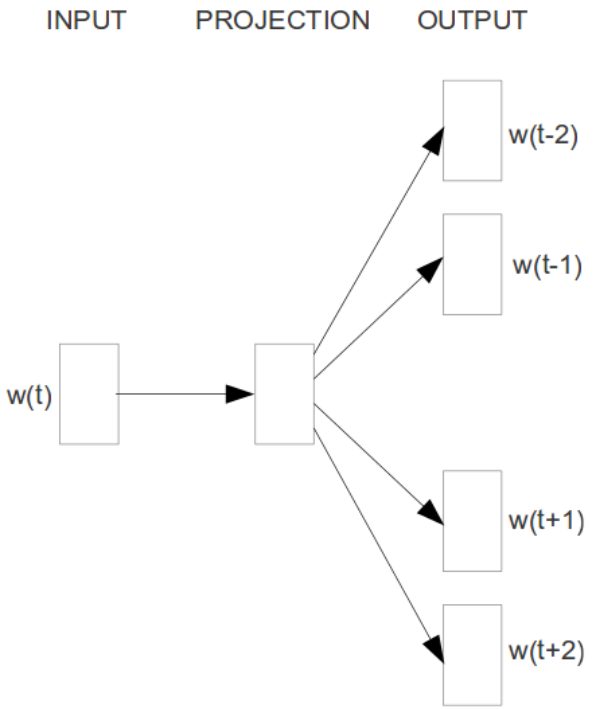
- Terim Frekansı – Ters Doküman Frekansı:
  - $tf(d,t)=c\downarrow dt\,/N\downarrow d$
  - $tdf(t)=\log(1+D/1+m\downarrow t\ )$
  - $tftdf(d,t)=tf(d,t)\ tdf(t)$
  - Gerçekleme: *Gensim* araç kutusu
  - Kelime hazne boyu için 1,000 ile 50,000 arasında değerler denenmiştir

$c\downarrow td$: t teriminin d dokümanındaki sayısı
$N\downarrow d$: d dokümanındaki toplam terim sayısı
D: toplam doküman sayısı
$m\downarrow t$: t teriminin geçtiği doküman sayısı

- Destek Vektör Makinası:
  - Doğrusal DVM
  - Birincil formülasyon
  - Penaltı (C) parametresi: varsayılan değer (1)
  - Çok sınıflı sınıflandırma: bire-hepsi yöntemi
  - Gerçekleme: *scikit-learn* araç kutusu

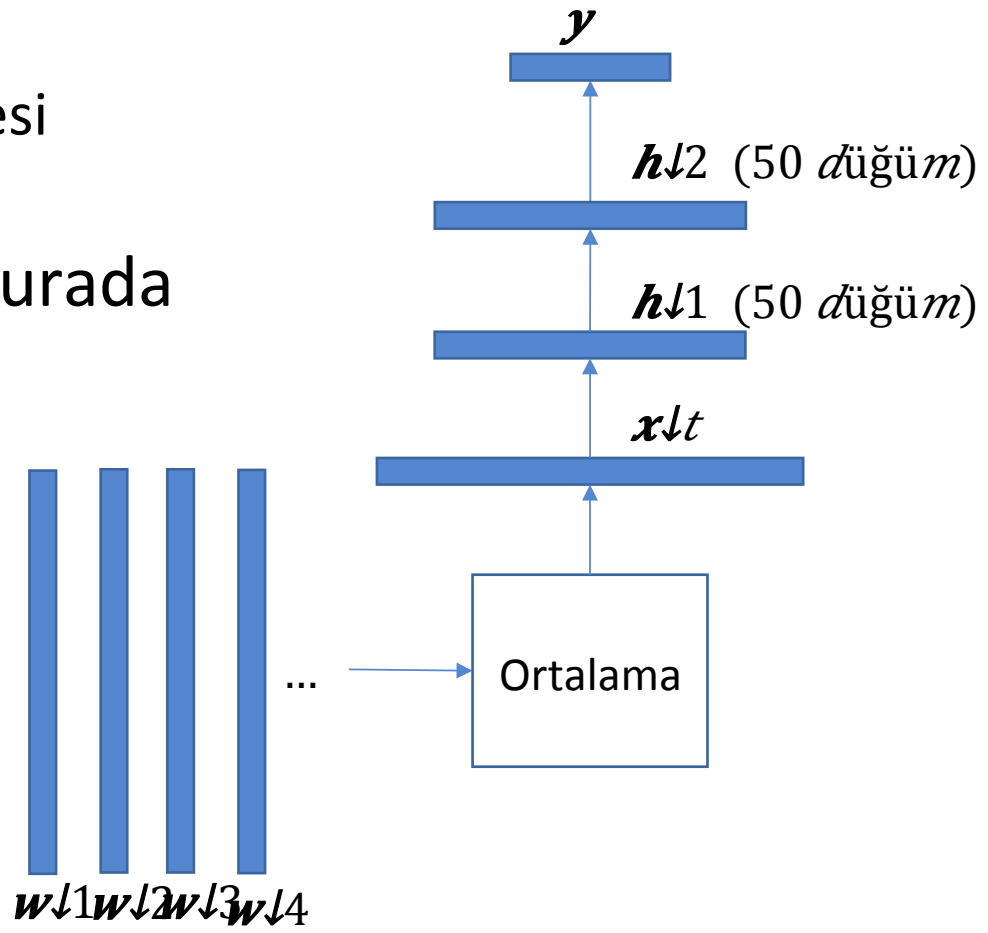
Kelime Temsilleri (KT) ve DVM sınıflandırıcısı

- Kelime → düşük boyutlu rasyonel vektör
- Gözetimsiz öğrenme: Atla-Gram modeli
- Eksi-örnekleme yakınlaştırması
- KT öğrenildikten sonra doküman temsilleri için KT'lerin ortalaması kullanılmıştır.
- DVM için TF-TDF ile aynı kurgu uygulanmıştır.
- Gerçekleme: *Gensim* araç kutusu
- Vektör boyutları için 100,200,400 ve 600 denenmiştir
- Derlemde 10'dan az geçen kelimeler elenmiştir. Kelime hazne boyları:
  - Cumhuriyet: 70,118
  - Sabah: 60,718



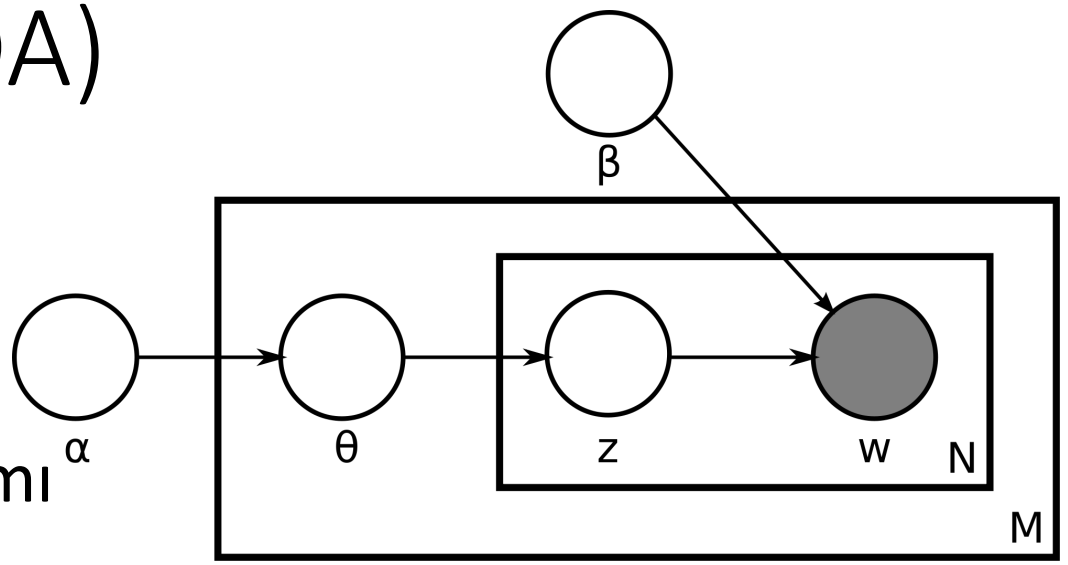
Yapay Sinir Ağı

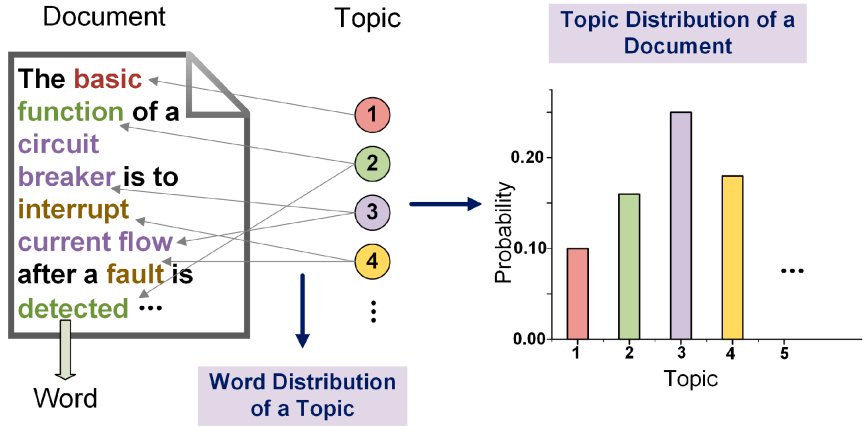
- Kelime temsillerinin ortalaması, YSA'na girdi olarak verildi
  - DVM'na avantajı: girdiye göre türev alınabilmesi
  - → KT'ler eğitim sırasında güncellendi.
- Atla-gram ile gözetimsiz öğrenilen KT'ler, burada ilklendirilme için kullanıldı.
- Hiper-parametreler:
  - Gizli katman aktivasyonları: ReLU
  - Çıktı aktivasyonu: S-biçim
  - Optimizasyon: RMSprop
  - Hedef fonk. : Ortalama Kareler Toplamı
- Gerçekleme: *PyTorch*



Saklı Dirichlet Ataması (SDA)

- Gözetimsiz bir yöntem
- K: konu sayısı, önceden belirleniyor
- Çıktı: dokümanın farklı konulara dağılımı
- Gözetimli probleme uyarlama:
  - Eğitim dokümanlarının konu dağılımları kullanılarak konu-sınıf eşleştirmesi
  - Her konu, ortalama skoru en yüksek sınıfa atanmıştır.
- Çevrimiçi SDA varyasyonu kullanılmıştır.





Sonuçlar – TFTDF: Hazne Boyu Etkisi

Derlem/Hazne Boyu	1K	5K	10K	20K	50K
Sabah	84,29	86,22	86,41	**86,52**	86.5
Cumhuriyet	69,12	71,71	**71,81**	71,72	71,69

Sonuçlar

Yöntem	Sabah	Cumhuriyet
SDA (K = 4 / K = 14)	65.41	47.94
SDA (K = 10 / K = 20)	67.60	43.31
SDA (K = 20 / K = 30)	72.08	45.37
TF-TDF (10K K. Haznesi) + DVM	86.41	71.81
KTV (d = 100) + DVM	85.47	70.34
KTV (d = 200) + DVM	86.16	71.55
KTV (d = 400) + DVM	86.72	72.24
KTV (d = 600) + DVM	86.89	72.50
KTV (d = 100) + YSA	**88.28**	**74.31**
KTV (d = 200) + YSA	87.93	73.64
KTV (d = 400) + YSA	87.94	72.29
KTV (d = 600) + YSA	87.53	72.97