YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ

DOĞAL DİL İŞLEME DERSİ DÖNEM PROJESİ

KONU MODELLEME (TOPIC MODELING)

18011613 – Mehmet Emre Gül

14011052 – Rahmi Cemre Ünal

Ocak, 2021

Konu Modelleme Nedir?

Konu modelleme, bir dizi belgeyi tarayabilen, içlerindeki sözcük ve kelime öbeği kalıplarını algılayabilen ve bir dizi belgeyi en iyi karakterize eden sözcük gruplarını ve benzer ifadeleri otomatik olarak kümelendirebilen, denetimsiz bir makine öğrenimi tekniğidir. Bir konu, genel temayı tanımlayan bir kelime koleksiyonudur. Örneğin haber makaleleri için konular politika, spor, ekonomi gibidir. Konu modellemesi size doğrudan konuların adlarını vermez, bunun yerine bir konuyu tanımlayabilecek en olası sözcükler dizisini verir. Sözcük kümesinin hangi konuyu ifade edebileceğini belirlemek bize kalmıştır.

Veri Kümesi

Projede kullanılan veri kümesi BBC'den alınan haberlerden oluşan 2225 tane doküman içermektedir. Bu haberler 5 farklı kategoride etiketlidir. Bunlar; business, entertainment, politics, sport and technology.

Uygulanan Adımlar

1) Veri önişleme
2) LDA modelinin oluşturulması
3) Sonuçlar

1) Veri Önişleme

Veri kümesindeki haber metinlerinde bulunan noktalama işaretleri atılır.

Metinler, kelime listelerine dönüştürülür.

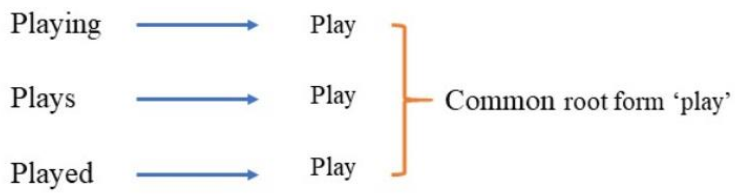
```
[['tate', 'lyle', 'boss', 'bags', 'top', 'award', 'tate', 'lyles', 'chief', 'executive', 'has', 'been', 'named', 'european', 'businessman', 'of', 'the', 'yea
', 'by', 'leading', 'business', 'magazine', 'iain', 'ferguson', 'was', 'awarded', 'the', 'title', 'by', 'us', 'publication', 'forbes', 'for', 'returning', 'o
', 'of', 'the', 'uks', 'venerable', 'manufacturers', 'to', 'the', 'countrys', 'top', 'companies', 'the', 'sugar', 'group', 'had', 'been', 'absent', 'from', '
e', 'ftse', 'for', 'seven', 'years', 'until', 'mr', 'ferguson', 'helped', 'it', 'return', 'to', 'growth', 'tates', 'shares', 'have', 'leapt', 'this', 'year',
boosted', 'by', 'firming', 'sugar', 'prices', 'and', 'sales', 'of', 'its', 'artificial', 'sweeteners', 'after', 'years', 'of', 'sagging', 'stock', 'price', '
d', 'seven', 'year', 'hiatus', 'from', 'the', 'ftse', 'one', 'of', 'britains', 'venerable', 'manufacturers', 'has', 'returned', 'to', 'the', 'vaunted', 'inde
', 'forbes', 'said', 'mr', 'ferguson', 'took', 'the', 'helm', 'at', 'the', 'company', 'in', 'after', 'spending', 'most', 'of', 'his', 'career', 'at', 'consum
', 'goods', 'giant', 'unilever', 'tate', 'lyle', 'which', 'was', 'an', 'original', 'member', 'of', 'the', 'historic', 'ft', 'index', 'in', 'operates', 'more'
'than', 'factories', 'and', 'more', 'additional', 'production', 'facilities', 'in', 'countries', 'previous', 'winners', 'of', 'the', 'forbes', 'award', 'incl
e', 'royal', 'bank', 'of', 'scotland', 'chief', 'executive', 'fred', 'goodwin', 'and', 'former', 'vodafone', 'boss', 'chris', 'gent']]
```

Şekil 1

Kelime listelerinden "stop words"'ler çıkartılır. 'Stop words' etkisiz kelimelerdir. Örneğin İngilizcede 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out' bunlara örnektir.

Bigram modeli oluşturulur. 2 gram (veya bigram), " please turn", "Ziraat Bankası" veya "Nasrettin Hoca" gibi iki kelimelik bir kelime dizisidir. Bu modelin oluşturulma amacı metinlerde bazı grupların sürekli tekrar etmesi ve birlikte anlam ifade etmesidir.

Daha sonra 'lemmatization' işlemi uygulanır. Lemmatizasyon, bir kelimenin çekimli biçimlerini bir araya getirme sürecidir, böylece tek bir öğe olarak analiz edilebilirler, sözcüğün lemması veya sözlük biçimi ile tanımlanabilirler.



Şekil 2

Sözlük ve corpus oluşturulur. Sözlükte kelimeler, id:string şeklinde tutulur. Corpus içinde ise id:frekans şeklinde tutulur.

Corpus : (id : frekans)

(0, 1), (1, 1), (2, 2), (3, 1), (4, 1),

(Sözlük[id], freakans):

('additional', 1), ('artificial', 1), ('award', 2), ('boost', 1), ('britain', 1)

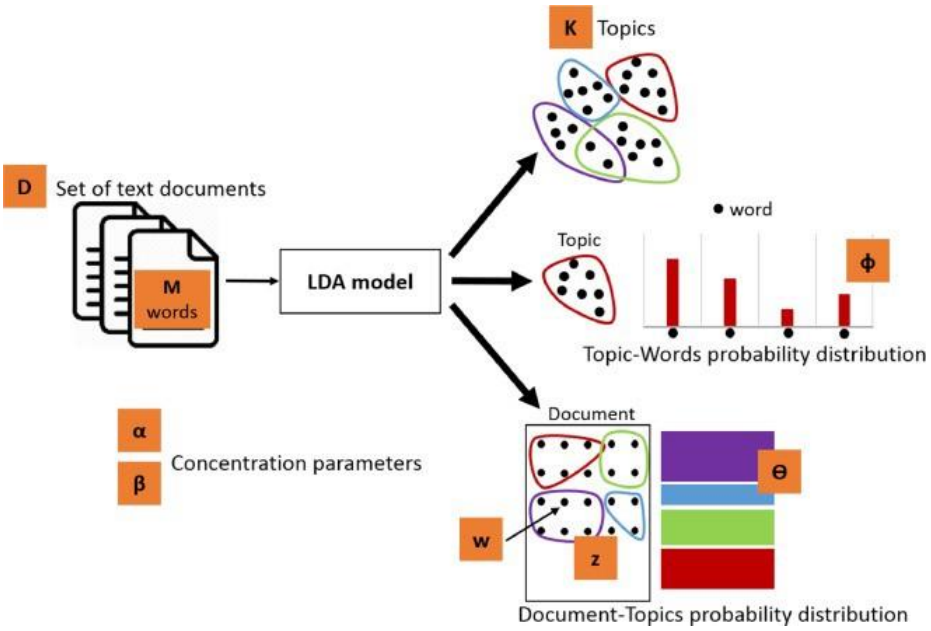
2) LDA Model

Latent Dirichlet Allocation (LDA), her belgenin bir konu koleksiyonu olarak kabul edildiği ve belgedeki her kelimenin konulardan birine karşılık geldiği bir topic modeling yöntemidir.

Dolayısıyla, bir belge verildiğinde LDA, belgeyi temel alarak her konu grubunu o grubu en iyi açıklayan bir dizi kelimenin olduğu konu gruplarına kümeler. LDA modeline verilen bir dökümanın çıktısı şu şekildedir.

```
Topic:business          Score: 0.63
Topic:tech              Score: 0.14
Topic:sport             Score: 0.13
Topic:politics          Score: 0.08
Topic:entertainment     Score: 0.01
```

Şekil 3



Şekil 4 *LDA Model*

3) Sonuçlar

Dökümanlar denetimsiz (unsupervised) olarak kümelenmiştir. Ancak, veri kümesinde her dökümanın etiket bilgisi mevcut olduğu için ilgili kümeler etiketler ile eşleştirildi. Buradan doğruluk oranı hesaplandı.

Veri kümesindeki başarı oranı %90.471'dir.

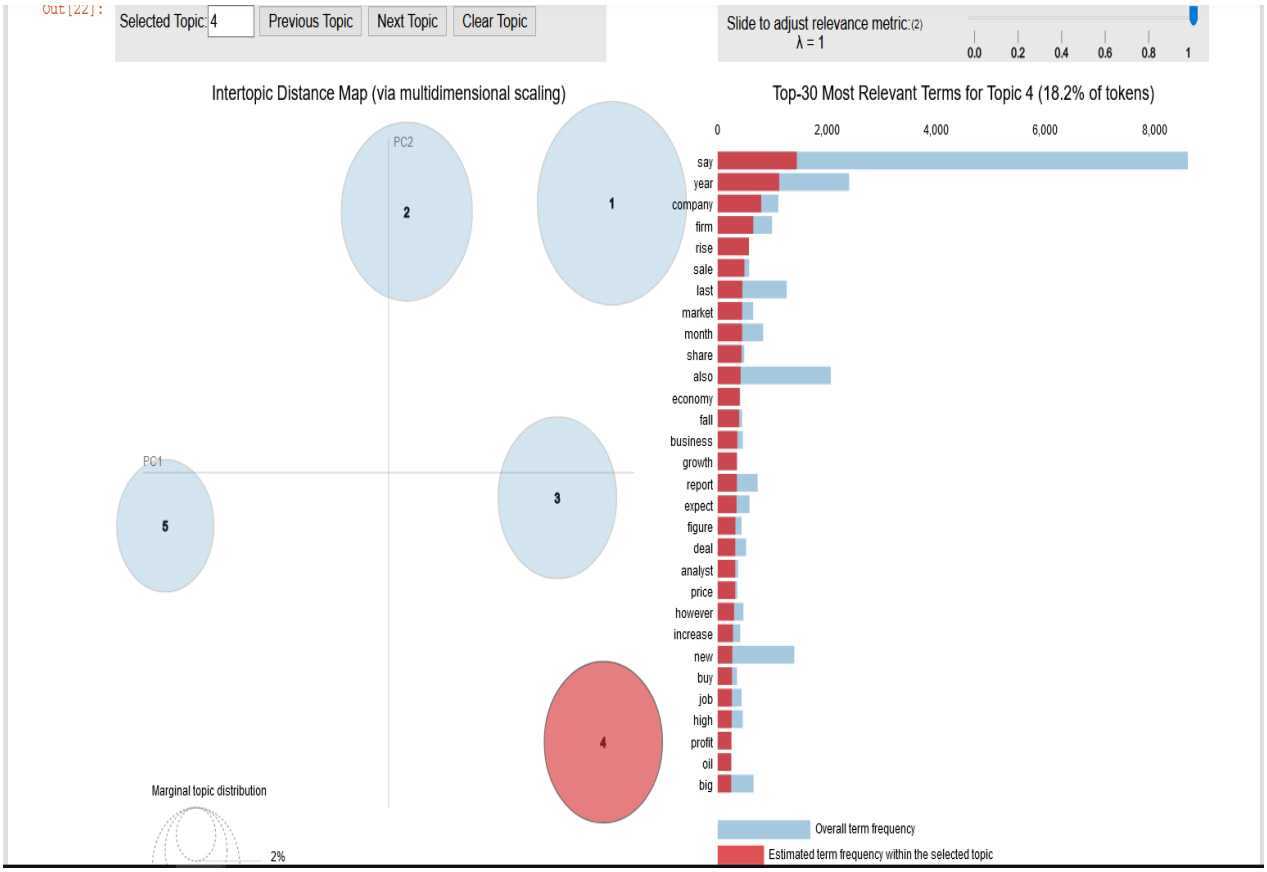
Eğitilen modelin çıkardığı kümeler ve her küme için bulduğu en belirleyici kelimeler aşağıdaki gibidir.

```
[(0,
  '0.023*"say" + 0.018*"year" + 0.013*"company" + 0.010*"firm" + 0.009*"rise" '
  '+ 0.008*"sale" + 0.007*"last" + 0.007*"market" + 0.007*"month" + '
  '0.007*"share"'),
 (1,
  '0.015*"say" + 0.011*"game" + 0.010*"go" + 0.010*"play" + 0.009*"time" + '
  '0.009*"take" + 0.008*"win" + 0.007*"first" + 0.007*"make" + 0.007*"player"'),
 (2,
  '0.046*"say" + 0.018*"would" + 0.010*"government" + 0.009*"people" + '
  '0.007*"could" + 0.006*"plan" + 0.006*"make" + 0.006*"election" + '
  '0.006*"tell" + 0.005*"also"'),
 (3,
  '0.024*"film" + 0.015*"good" + 0.013*"year" + 0.011*"include" + '
  '0.011*"music" + 0.011*"show" + 0.011*"also" + 0.009*"award" + 0.008*"say" + '
  '0.007*"take"'),
 (4,
  '0.018*"say" + 0.017*"use" + 0.014*"people" + 0.010*"technology" + '
  '0.008*"user" + 0.008*"site" + 0.008*"computer" + 0.007*"make" + 0.007*"new" '
  '+ 0.006*"net"')]
```

Şekil 5

Buna göre:

Topic 0: Business 1 : Sport 2 : Politics 3 : Entertainment 4 : Technology



Şekil 6 *Modelin görselleştirilmiş hali*

Modelin çalışmasının örnek bir haber üzerinden gösterilmesi

Bitcoin's value surged above $34,000 (£24,850) for the first time on Sunday as the leading cryptocurrency continued to soar.
It put the gain this year at almost $5,000, although by 17:00 GMT the price had drifted lower to about $33,000, according to the Coindesk website.
The rise was put down to interest from big investors seeking quick profits.
It comes after Bitcoin soared 300% last year, with the price of many other digital currencies also rising sharply.
Ethereum, the second biggest cryptocurrency, gained 465% in 2020
Some analysts think Bitcoin's value could rise even further as the US dollar drops further.
While the value of the US currency rose in March at the start of the coronavirus pandemic as investors sought safety amid the uncertainty, it has since dropped due to major stimulus from the US Federal Reserve. The currency ended last year with its biggest annual loss since 2017.
Covid worries help Bitcoin to three-year high
How do cryptocurrencies work?
Bitcoin is traded in much the same way as real currencies like the US dollar and pound sterling.
Recently it has won growing support as a form of payment online, with PayPal among the most recent adopters of digital currencies.
But the cryptocurrency has also proved to be a volatile investment.
The soaring price has raised concerns that Bitcoin is due for a dramatic correction, as happened three years ago when the value collapsed after a bull run.
During the rally in 2017 Bitcoin came close to breaking through the $20,000 level, only to hit extreme lows and fall below $3,300.
It passed $19,000 in November last year before dropping sharply again.
‘Robin Hood’ hackers giving stolen money to charity
'One day everyone will use China's digital currency'
In October, Bank of England Governor Andrew Bailey cautioned over Bitcoin's use as a payment method.
"I have to be honest, it is hard to see that Bitcoin has what we tend to call intrinsic value," he said. "It may have extrinsic value in the sense that people want it."
Mr Bailey added that he was "very nervous" about people using Bitcoin for payments pointing out that investors should realise its price is extremely volatile.

Şekil 7 *Örnek haber metni*

Lemmatizasyon sonrası:

```
['surge', 'first', 'time', 'lead', 'cryptocurrency', 'continue', 'soar',
'put', 'gain', 'year', 'almost', 'gmt', 'price', 'drift', 'lower',
'accord', 'website', 'rise', 'put', 'interest', 'big', 'investor', 'seek',
'quick', 'profit', 'come', 'bitcoin', 'soar', 'last', 'year', 'price',
'many', 'digital', 'currency', 'also', 'rise', 'sharply', 'ethereum',
'second', 'big', 'cryptocurrency', 'gain', 'analyst', 'think', 'bitcoin',
'value', 'could', 'rise', 'even', 'dollar', 'drop', 'value', 'currency',
'rise', 'start', 'investor', 'seek', 'safety', 'uncertainty', 'drop',
'due', 'major', 'stimulus', 'currency', 'end', 'last', 'year', 'big',
'annual', 'loss', 'covid', 'worry', 'help', 'year', 'high', 'work',
'bitcoin', 'trade', 'much', 'way', 'real', 'currency', 'dollar', 'pound',
'sterling', 'recently', 'grow', 'support', 'form', 'payment', 'online',
'paypal', 'recent', 'adopter', 'digital', 'currency', 'cryptocurrency',
'also', 'prove', 'volatile', 'investment', 'soar', 'price', 'raise',
'concern', 'bitcoin', 'due', 'dramatic', 'correction', 'happen', 'value',
'collapse', 'come', 'close', 'breaking', 'level', 'hit', 'extreme', 'low',
'fall', 'pass', 'last', 'year', 'drop', 'sharply', 'give', 'steal',
'money', 'charity', 'day', 'use', 'caution', 'payment', 'method',
'honest', 'hard', 'see', 'bitcoin', 'tend', 'call', 'intrinsic', 'value',
'say', 'extrinsic', 'value', 'sense', 'people', 'want', 'add', 'nervous',
'people', 'use', 'bitcoin', 'payment', 'point', 'investor', 'realise',
'price', 'extremely', 'volatile']
```

Model tahmini

Topic	Score	Topic Keywords
business	0.63	0.023*"say" + 0.018*"year" + 0.013*"company" + 0.010*"firm" + 0.009*"rise"
technology	0.14	0.018*"say" + 0.017*"use" + 0.014*"people" + 0.010*"technology" + 0.008*"user"
sport	0.13	0.015*"say" + 0.011*"game" + 0.010*"go" + 0.010*"play" + 0.009*"time"
politics	0.08	0.046*"say" + 0.018*"would" + 0.010*"government" + 0.009*"people" + 0.007*"could"
entertainment	0.01	0.024*"film" + 0.015*"good" + 0.013*"year" + 0.011*"include" + 0.011*"music"

KAYNAKLAR

- http://mlg.ucd.ie/datasets/bbc.html
- https://www.datacamp.com/community/tutorials/stemming-lemmatization-python
- https://en.wikipedia.org/wiki/Lemmatisation
- https://merttopuz.com/universite/veri-madenciligi/n-gram-nedir-neden-kullanilir
- https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation