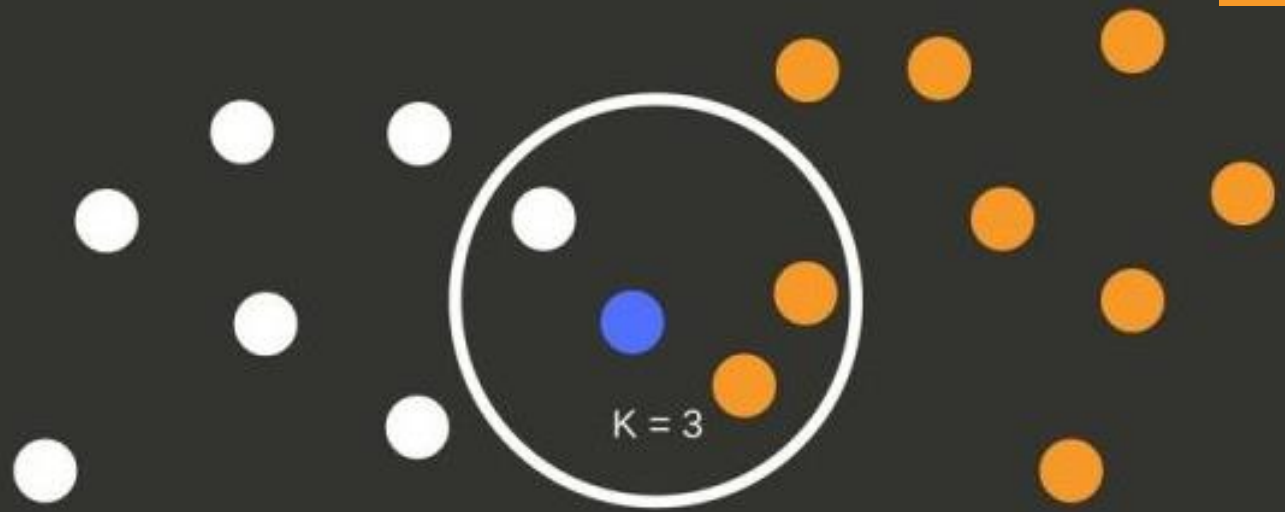
AI 01 김세진

KNN Recommendation Project



사용한 데이터

Kaggle

BBC News Summary

2004~2005년

2225개의 뉴스 요약

Kaggle

New York Times Articles

2020년

42420개의 뉴스 제목과 요약

프로젝트 배경

데이터 분석 목표

특정 기사와 비슷한 내용을 가지고 있는 기사를 찾기

⋮

NLP와 KNN 알고리즘을 사용하여 해결하기

전처리 과정

실제 텍스트 예시

Economy 'stronger than forecast'

The UK economy probably grew at a faster rate in the third quarter than the 0.4% reported, according to Bank of England deputy governor Rachel Lomax.

Private sector business surveys suggest a stronger economy than official estimates, Ms Lomax said. Other surveys collectively show a rapid slowdown in UK house price growth, she pointed out. This means that despite a strong economic growth, base rates will probably stay on hold at 4.75%. Official data comes from the Office for National Statistics (ONS). Though reliable, ONS data takes longer to publish, so now the BoE is calling for faster delivery of data so it can make more effective policy decisions. "Recent work by the Bank has shown that private sector surveys add value, even when preliminary ONS estimates are available," Ms Lomax said in a speech to the North Wales Business Club.

The ONS is due to publish its second estimate of third quarter growth on Friday. "The MPC judges that overall growth was a little higher in the third quarter than the official data currently indicate," Ms Lomax said. The Bank said successful monetary policy depends on having good information. Rachel Lomax cited the late 1980s as an example of a time when weak economic figures were published, but substantially revised upwards years later.

"The statistical fog surrounding the true state of the economy has proved a particularly potent breeding ground for policy errors in the past," she said. Improving the quality of national statistics is the single the best way of making sure the Monetary Policy Committee (MPC) makes the right decisions, she said. The Bank of England is working in tandem with the ONS to improve the quality and speed of delivery of data. Her remarks follow criticism from the House of Lords Economic Affairs Committee, which said the MPC had held interest rates too high given that inflation was way below the 2% target.

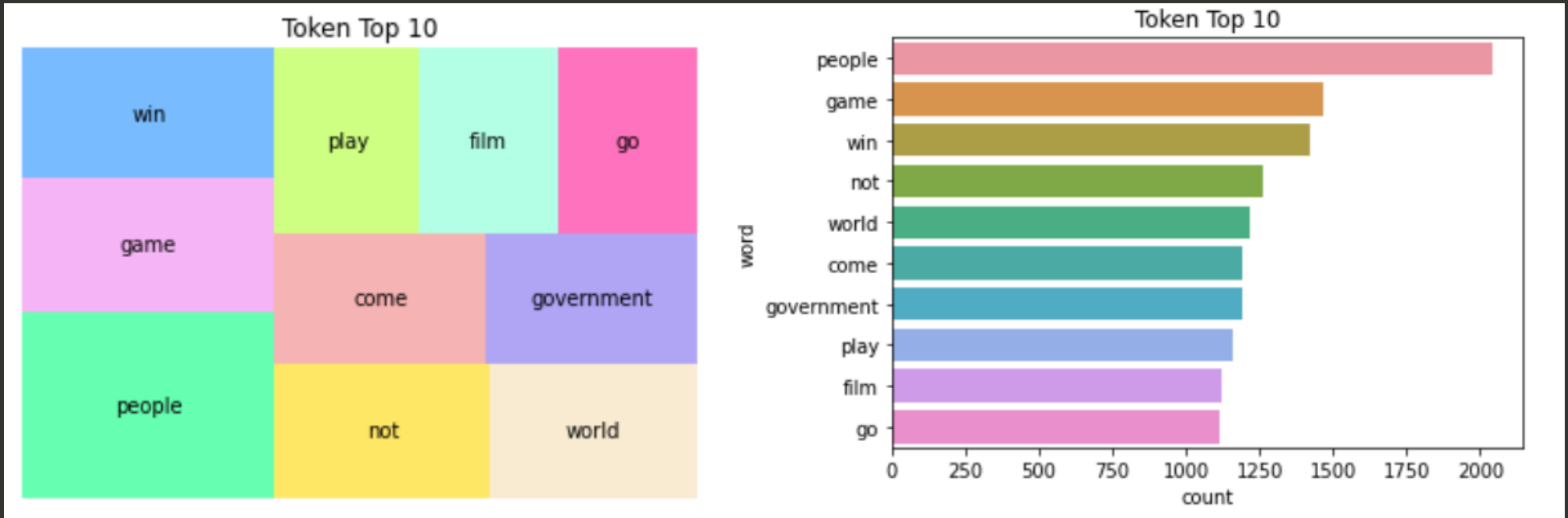
1. 분석에 필요 없는 데이터 정리
2. 중복된 데이터 정리
3. 특수문자 정리
4. 정규식을 사용하여 문자 정리(소문자와 숫자만 남겨 줌)
5. 기본 불용어 사전에 추가로 불용어를 추가하여 처리
6. 표제어 추출

처리된 텍스트 예시

```
'soar', 'oil', 'hit', 'world', 'economy', 'soar', 'cost', 'oil', 'hit',
'global', 'economic', 'growth', 'world', 'major', 'economy', 'weather',
    'storm', 'price', 'rise', 'accord', 'oecd', 'late', 'biannual',
   'report', 'oecd', 'cut', 'growth', 'prediction', 'world', 'main',
   'industrialise', 'region', 'growth', 'reach', 'fall', 'previous',
                        'estimate', 'oecd'.....
```

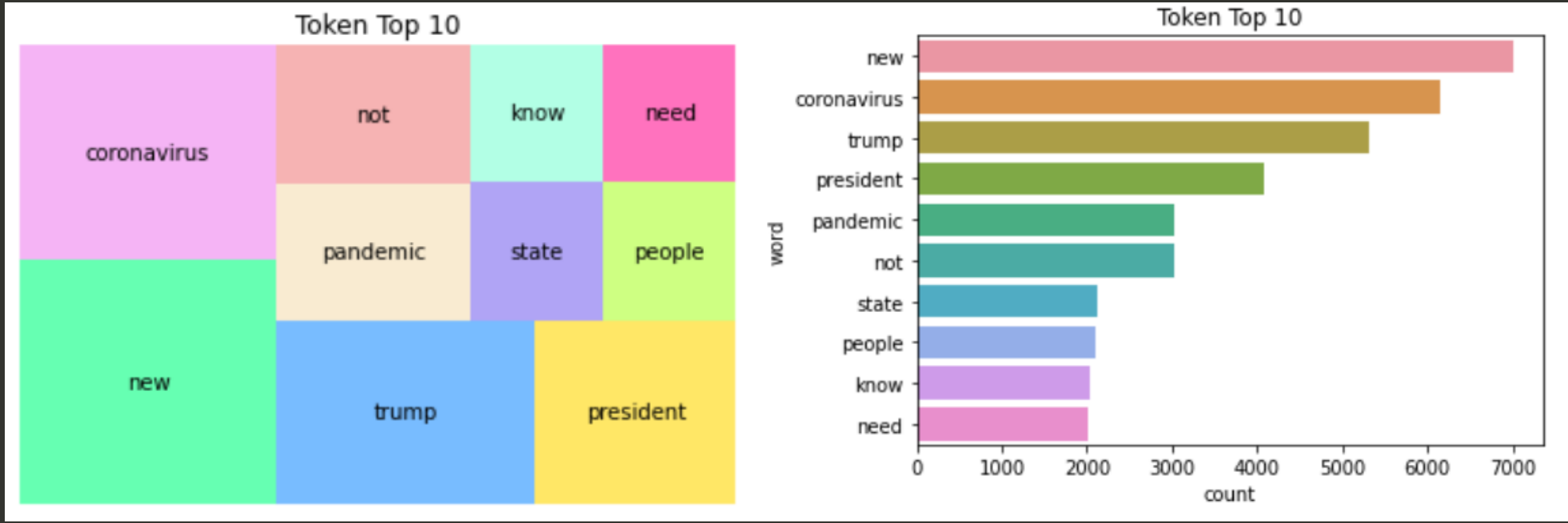
시각화

BBC Article



시각화

New York Times Article



→ 2020년에는 **신종 코로나 바이러스**가 유행하기 시작했고, **미국 대통령 선거**가 있었기 때문에 'new', 'coronavirus', 'trump', president'와 같은 단어들이 많이 쓰인 것을 알 수 있다.

분석 결과

BBC Article : **100**번째 기사와 유사한 기사들 번호 + 내용

120	us interest rates increased to 2% …
150	bank voted 8-1 for no rate change …
136	uk interest rates are set to remain on hold at 4.75% …
289	bank holds interest rate at 4.75% …
437	uk interest rates held at 4.75% …

→ **은행 금리**와 관련된 기사들이 추천됨

분석 결과

New York Times Article : **10000**번째 기사와 유사한 기사들 번호 + 내용

16469	Neighboring Iran, badly hit by the virus, continues to allow thousands of people to cross into Afghanistan daily despite requests to close the border ...
7717	What is this illustration saying? Maze ...
17301	No. Should I Still Be Going Out? ...
58484	Dani Raymon bamboozles us. Puns and Anagrams ...
26540	For this issue, a look at Afghanistan as the virus threatens to explode with few resources to contain it ...

→ **아프가니스탄 지역**과 관련된 기사가 추천됨

추가 분석 목표

특정 단어의 유무를 판단하는 분류모델 만들기

⋮

Random Forest 분류 모델로 해결하기

추가 분석을 위한 전처리

```
for row in df['sentence']:
  df['sentence'] = df['sentence'].str.lower()
```

```
corona = df['sentence'].str.contains('coronavirus')
df['label'] = corona
```

```
df['label'] = df['label'].astype(int)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 42420 entries, 0 to 69113
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   sentence  42420 non-null  object
 1   Tokens    42420 non-null  object
 2   label     42420 non-null  int64
dtypes: int64(1), object(2)
memory usage: 1.3+ MB
```

```
df['label'].value_counts()
```

```
0    36616
1    5804
Name: label, dtype: int64
```

→ 뉴욕 타임즈 기사 내용에 '**coronavirus**'가 포함되어 있으면 1을 표시하고, 포함되어 있지 않으면 0을 표시하여 라벨을 만들어 주는 작업

→ 총 42420 개의 기사 중에 이 단어가 표시된 기사는 **5804**개, 이 기사들을 찾아내는 모델을 만들어 보자!

추가 분석 베이스라인

TfidVectorizer

▼

TruncatedSVD

▼

LogisticRegression

	Precision	Recall	F1-Score
0	0.87	0.98	0.92
1	0.48	0.1	0.16
Accuracy			0.86
Macro Avg	0.68	0.54	0.54
Weighted Avg	0.82	0.86	0.82

추가 분석 모델 파이프라인

TfidVectorizer

↓

TruncatedSVD

↓

RandomForestClassifier

↓

RandomizedSearchCV

↓

Model Fit

	Precision	Recall	F1-Score
0	0.95	0.96	0.96
1	0.75	0.68	0.72
Accuracy			0.93
Macro Avg	0.85	0.82	0.84
Weighted Avg	0.92	0.93	0.92

Thank You