뉴스를 이용한 주식시장 예측(with Kaggle)

직방 부동산데이터팀 서범석

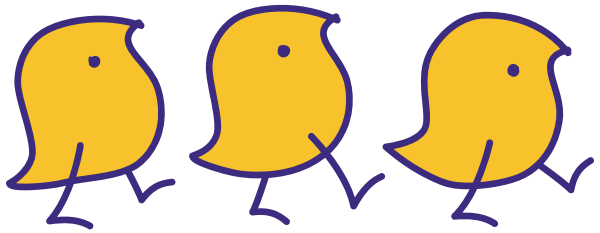
발표자 및 발표내용

발표자 : 직방 데이터노동자

발표 주제 : 텍스트 데이터 수치화 및 주식시장에 반영

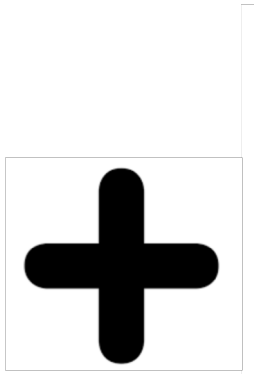
목차

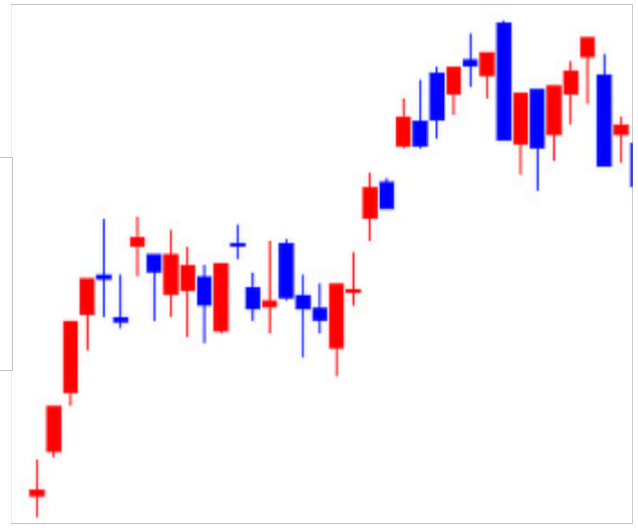
Language Conference 2019

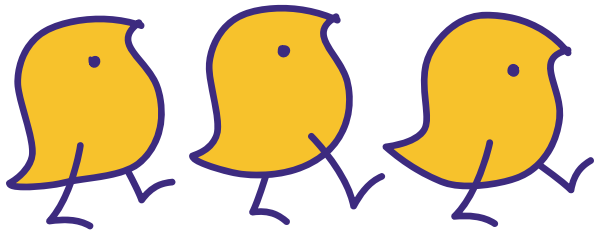


분석 배경

분석배경







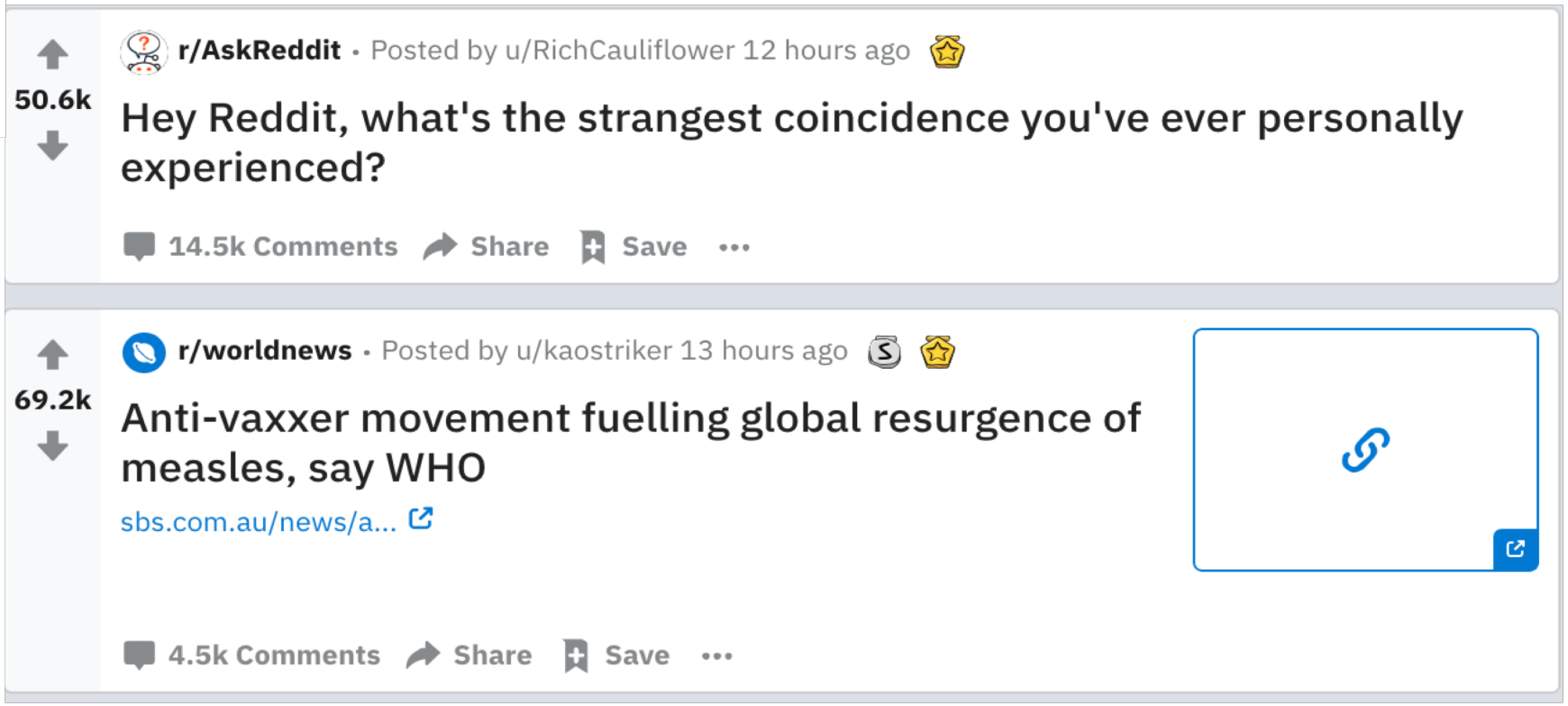
데이터 설명

데이터 설명 : 데이터 수집경로

- 데이터 제공 URL(https://www.kaggle.com/aaron7sun/stocknews)

데이터 설명 : 뉴스데이터(Reddit)

reddit



데이터 설명 : 주식데이터(다우 존스 산업 평균 지수(DJIA))

Coca-Cola

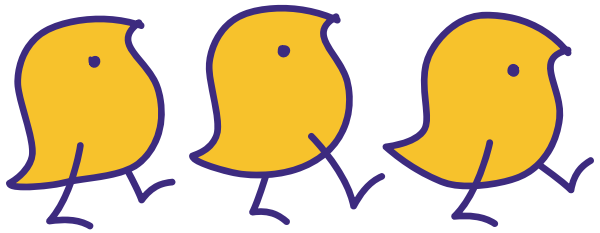
Microsoft

IBM

i'm lovin' it

VISA

intel

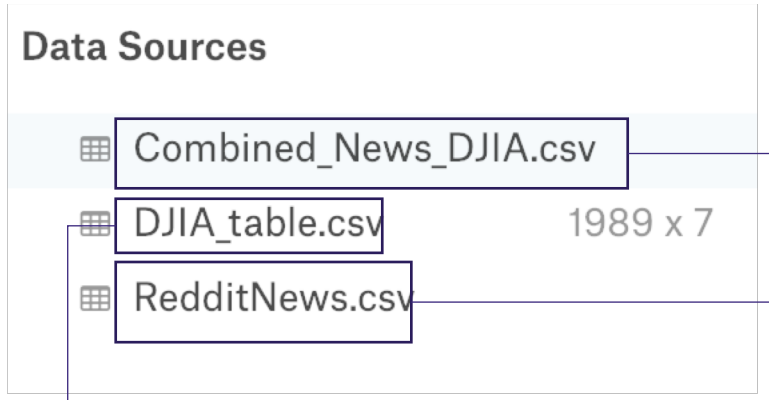


데이터 탐색

LangCon 2019

데이터 탐색 : 데이터 확인하기 (Viewing Data)

Language Conference 2019



News Data Shae : (1986, 27)

	Date	Label	Top1	Top2	Top3	Top4	
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b rap 'in

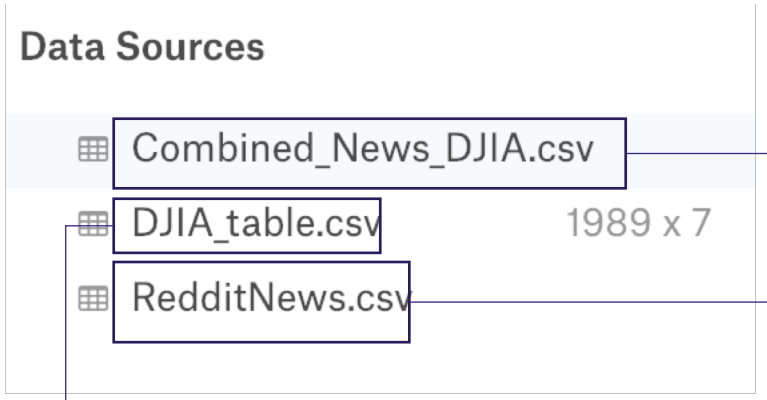
Reddit Data Shape : (73608, 2)

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host

Stock Data Shape : (1989, 7)

	Date	Open	High	Low	Close	Volume	Adj Close
0	2016-07-01	17924.240234	18002.380859	17916.910156	17949.369141	82160000	17949.369141
1	2016-06-30	17712.759766	17930.609375	17711.800781	17929.990234	133030000	17929.990234

데이터 탐색 : 결측치 확인 (Missing Data)




```
news_df[["Top23","Top24","Top25"]].isnull().sum()
```

```
Top23    1
Top24    3
Top25    3
dtype: int64
```

```
reddit_df.isnull().sum()
```

```
Date    0
News    0
dtype: int64
```

```
stock_df.isnull().sum()
```

```
Date         0
Open         0
High         0
Low          0
Close        0
Volume       0
Adj Close    0
dtype: int64
```

	Date	Label		Top23	Top24	Top25
277	2009-09-15	1		NaN	NaN	NaN
348	2009-12-24	1	b"Ayatollah Montazeri's Legacy: In death he m...	NaN	NaN	
681	2011-04-21	1	Prince Charles wins some kind of a record	NaN	NaN	

데이터 탐색 : 이상치 확인 (Outlier Detection)

Language Conference 2019

```
Reddit Data Max Date : 2016-07-01
Reddit Data Min Date : 2008-06-08
```

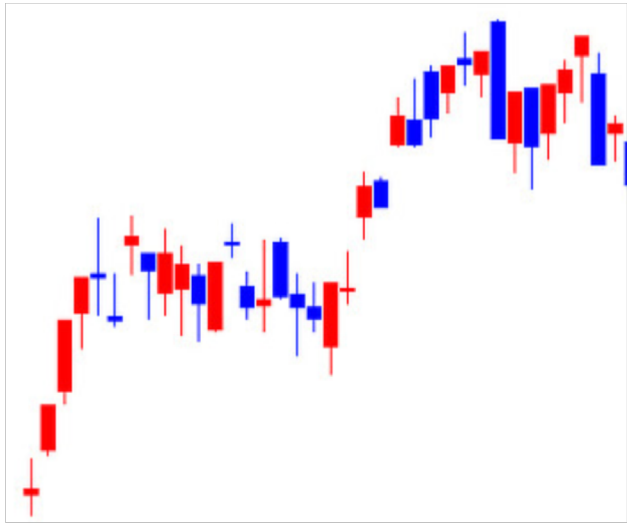
```
Stock Data Max Date : 2016-07-01
Stock Data Min Date : 2008-08-08
```


```
 &amp;amp; other imperiled species to be decid
creasing protection for iconic mammals &am
\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n  '
```

```
Max_News_Length :  318
```

```
'How Ironic. '
```

```
Min_News_Length :  12
```



```
stock_df["Close_M_Open"].describe()
```

```
count    1989.000000
mean        3.916206
std       141.227938
min      -774.169922
25%       -60.169922
50%         8.540039
75%        74.558594
max       925.190430
Name: Close_M_Open, dtype: float64
```

데이터 탐색 : 이상치 확인 (Outlier Detection)

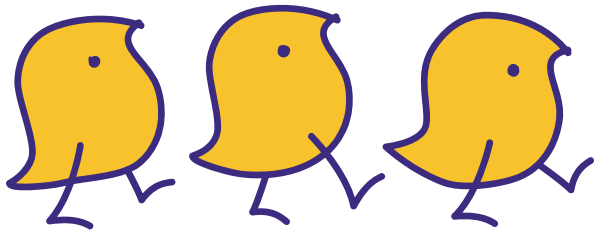
- 지수가 급등한 날의 뉴스

  - EU , 백열전구 금지
  - AFP : Paul Krugman, 노벨 경제상 수상
  - 크리스천들에게 힌두교의 위협 : 개종 또는 도망자
  - 유럽은 은행에 2 조 3 천억 달러를 투입
  - 노벨 경제상 폴 크루그먼 (Paul Krugman) 축하합니다!
  - 정부가 데이터를 잃어 버렸다! 최대 1.7 백만명의 사람들이 누락 된 데이터
  - 베네수엘라, 맥도날드 폐쇄

데이터 탐색 : 이상치 확인 (Outlier Detection)

- 지수가 급락한 날의 뉴스

  - 악몽의 힘 - BBC (파트 1)
  - 일본은 자동차를 만든다. 사우디 아라비아 펌프 오일. 중국은 양말과 평면 TV를 공급합니다. 미국의 1 위 수출은 무엇입니까? 빚
  - 거대한 유럽 은행과 iinsurance 거인은 3 명의 분리 된 정부에 112 억 유로 (164 억 달러)의 구제 금융을 제공하지 못한다
  - 일본 교통 장관, 제재를 피한다
  - 걸프 - 미국에 대한 월스트리트 저축 압력
  - 이라크 군인, 게이 인 이라크 지도자 암살



데이터 전처리

LangCon 2019

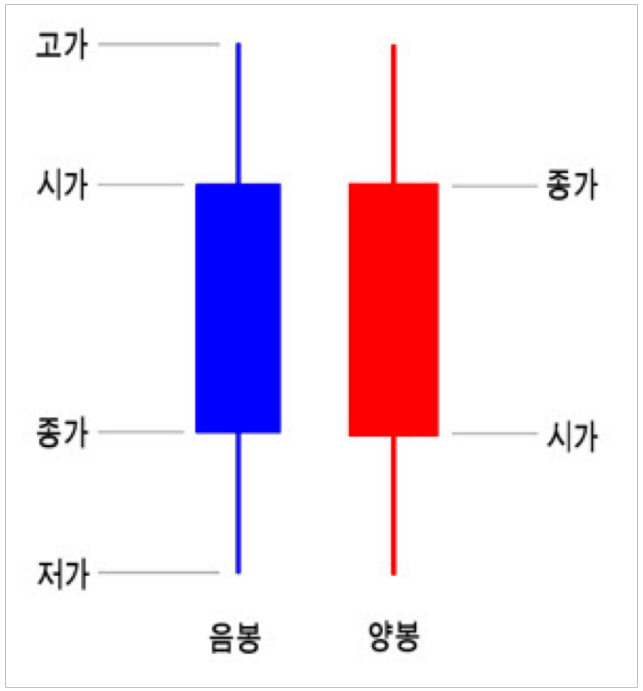
데이터 전처리 : 자연어

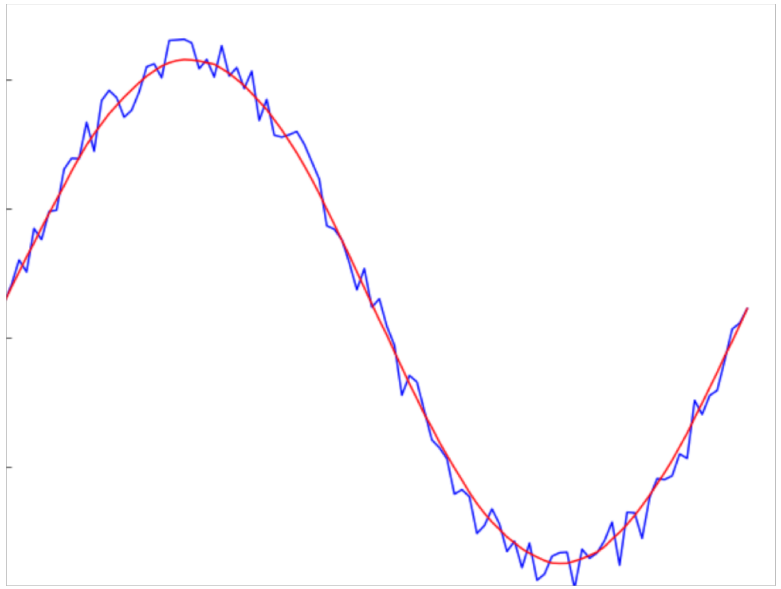
Language Conference 2019

Natural Language ToolKit

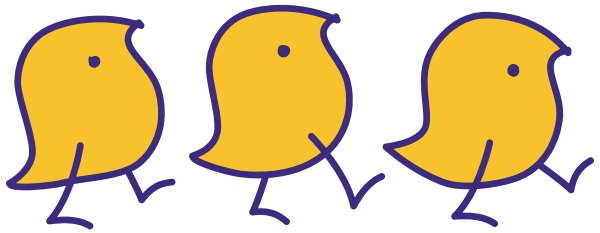
python™

데이터 전처리 : 시계열데이터





1	.	.
2	1	.
3	2	1
4	.	.
5	4	.
6	.	.
7	6	.

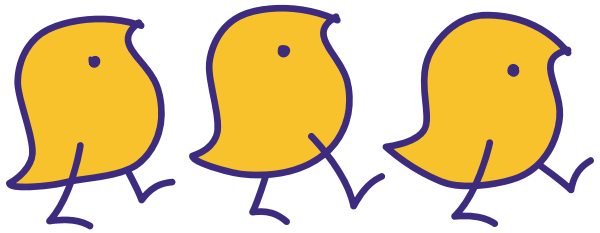


토픽 모델링

LangCon 2019

데이터 모델링 : 토픽 모델링(Topic Modeling)

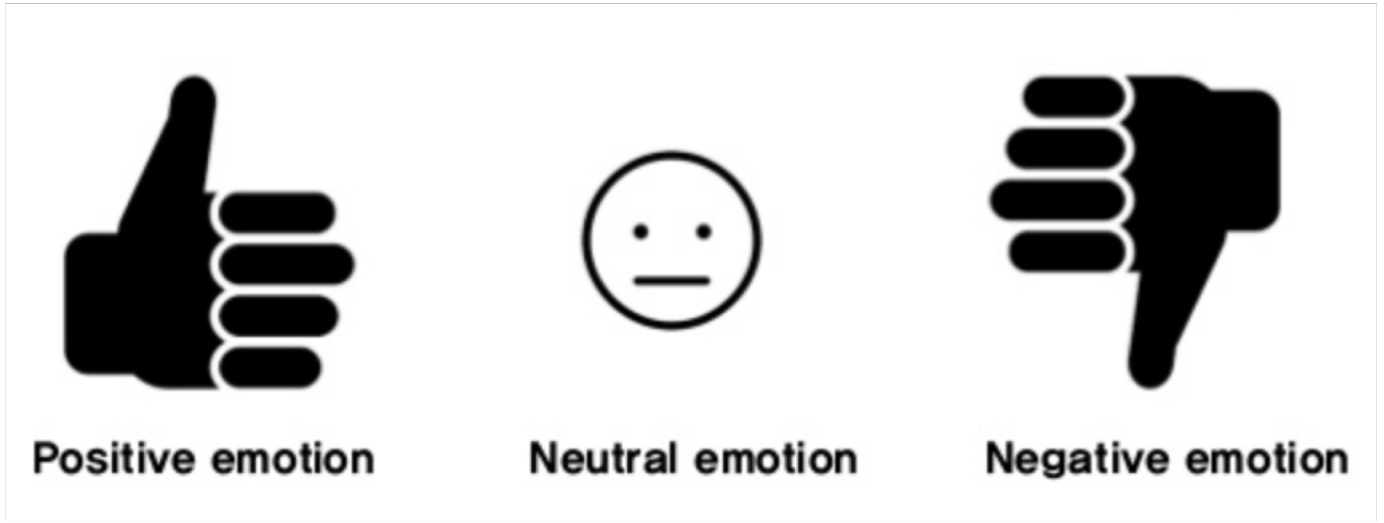
- LDA(Latent Dirichlet Allocation)를 이용한 토픽모델링



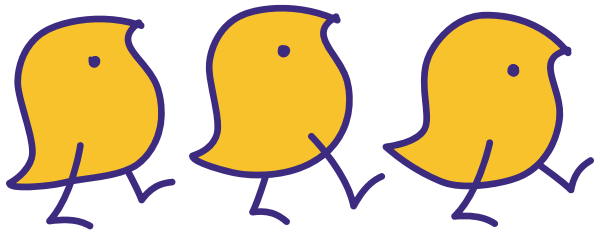
감성분석

LangCon 2019

감성분석(Semantic Analysis)



	text	compound	neg	neu	pos
0	georgia downs two russian warplanes as count...	-0.5994	0.262	0.738	0.0
1	breaking musharraf to be impeached	0.0000	0.000	1.000	0.0



결과 및 개선방향

결과 및 개선방향 : 개선방향

Microsoft

2σ TWO SIGMA