Unit 2

GloVe & NLP주가

12주차.
자연어처리와 주가예측

학습 내용

- GloVe
- NLP와 주가

학습 목표

- GloVe에 대해 이해할 수 있다.
- NLP를 이용해 주가를 분석할 수 있다.

GloVe

- [x] 벡터 차이의 의미 인코딩

벡터 차이의 의미 인코딩

→ 동시발생 확률의 비율을 통해
의미 구성 요소를 인코딩

GloVe

☑ 벡터 차이의 의미 인코딩

x = solid	x = solid	x = gas	x = water	x = random
$P(x \mid \text{ice})$	large	small	large	small
$P(x \mid \text{steam})$	small	large	large	small
$\frac{P(x \mid \text{ice})}{P(x \mid \text{steam})}$	large	small	~1	~1

GloVe

☑ 벡터 차이의 의미 인코딩

x = solid	x = solid	x = gas	x = water	x = random
$P(x \mid \text{ice})$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(x \mid \text{steam})$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$\frac{P(x \mid \text{ice})}{P(x \mid \text{steam})}$	8.9	8.9×10^{-2}	1.36	0.96

GloVe

☑ 벡터 차이의 의미 인코딩

? 단어 벡터 공간에서 선형 의미 구성 요소로
동시 발생 확률의 비율을 사용하는 방법

Log-bilinear 모형

$$w_i \cdot w_j = \log P(i|j)$$

Log-bilinear 모형과 벡터의 차이

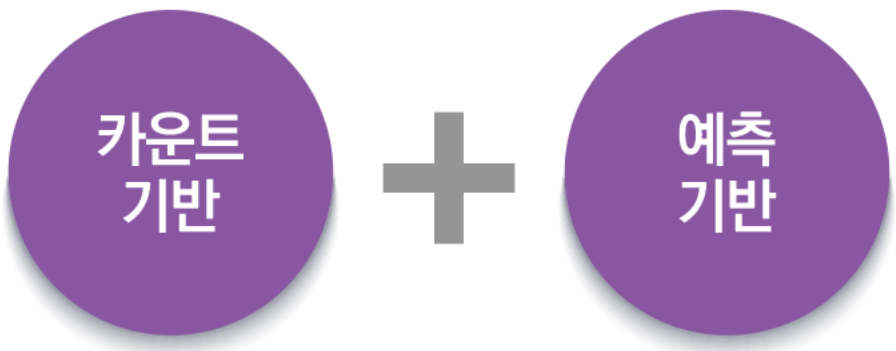
$$w_x \cdot (w_a - w_b) = \log \frac{P(x|a)}{P(x|b)}$$

GloVe

GloVe란?

GloVe

Global Vectors for Word Representation



GloVe

☑ GloVe란?

GloVe Pennington et al.(2014)

「 GloVe: Global Vectors for Word Representation 」

카운트 기반의 LSA와 예측 기반의 Word2Vec, 두 Approach의 단점을 보완

- Word2Vec와 GloVe 유사한 성능
- 필요성에 따라 선택하여 사용

GloVe

GloVe란?

카운트 기반 : LSA

- 각 단어의 빈도수를 카운트 한 행렬로 받아 차원을 축소
- 특이값분해를 통해 의미를 끌어내는 방법론
- 카운트 기반으로 말뭉치의 전체적인 통계 정보를 고려
- 단어 의미의 유추 작업에는 성능이 떨어짐

GloVe

GloVe란?

예측 기반 : Word2Vec

- 실제값과 예측값에 대한 오차를 손실 함수를 통해 줄여가며 학습
- 단어 간 유추 작업에는 LSA보다 뛰어난 성능
- 임베딩 벡터가 윈도우 크기 내에서만 주변 단어를 고려
- 말뭉치의 전체적인 통계 정보를 반영하지 못함

GloVe는 단점을 보완하며 두 가지 예측기반 방법 **모두 사용**

실질적으로는 **Word2Vec와 유사한 성능**

NLP와 주가

☑ 전처리 과정

전처리 과정

- 특성을 추출하기까지 **작업의 양이 방대함**
- 자연어처리는 특별히 **전처리 과정이 힘든 과정**
  - 예 감성분석의 경우 비꼬는 말의,최신 언어, 은어 등은 기계가 파악하기 어려움
- 사전처리의 정확도에 따라 **예측력 향상에 영향**

NLP와 주가

특성추출

트윗	▪ "@"기호는 "PERSON"으로 대체 ▪ "#"은 "TOPIC"으로 대체	슬랭	▪ 정규화된 단어로 대체
정제	▪ 말뭉치의 노이즈 데이터를 제거	불용어	▪ 더 이상 사용되지 않으므로 제거
정규화	▪ 표현 방법이 다양한 단어들을 하나의 단어로 통합	Negation	▪ 부정표현을 Negation이라는 토큰으로 대체

NLP와 주가

☑ 특성추출

품사 태거 (Parts of Speech)	▪ 명사, 동사, 접속사 등 문법적인 표시로 주석 첨부 ▪ 문장의 의미 유추에 도움
Bag-of-Words	▪ 단어들의 출현 빈도에 집중한 텍스트 데이터의 수치화 표현 방법 ▪ 단어의 unigram, bigram, n-gram 단계를 고려하여 감정 어휘를 사용해서 주관성 점수 제공
Feature Hashing(FH)	▪ 해시 태그가 지정된 단어는 작가가 직접 삽입한 감정과 레이블 → 매우 유용

NLP와 주가

기계학습의 방법론

전처리 과정

↓

기계학습, 딥러닝 기법

- Support Vector Machine
- Naive Bayes
- logistic regressions
- Random Forest

NLP와 주가

☑ 뉴스 감성분석과 주가 예측

주가 데이터	▪ 시가, 고가, 저가, 종가, 수정종가, 거래량
뉴스	▪ 구글뉴스, 로이터, 야후파이낸스 ▪ 2013년 2월 ~ 2016년 4월(약 3년)

수집된 뉴스 텍스트의 긍정/부정을 판단

↓

주가에 미치는 영향 파악

출처 로고 이미지 : https://namu.wiki/w/Apple

NLP와 주가

[긍정 2,360개 단어, 부정 7,383개 단어]

기계학습 방법론으로 테스트

- Random Forest : 약 88~92%
- Support Vector Machine : 약 86%
- Naïve Bayse : 약 83%

테스트 데이터로 예측한 결과

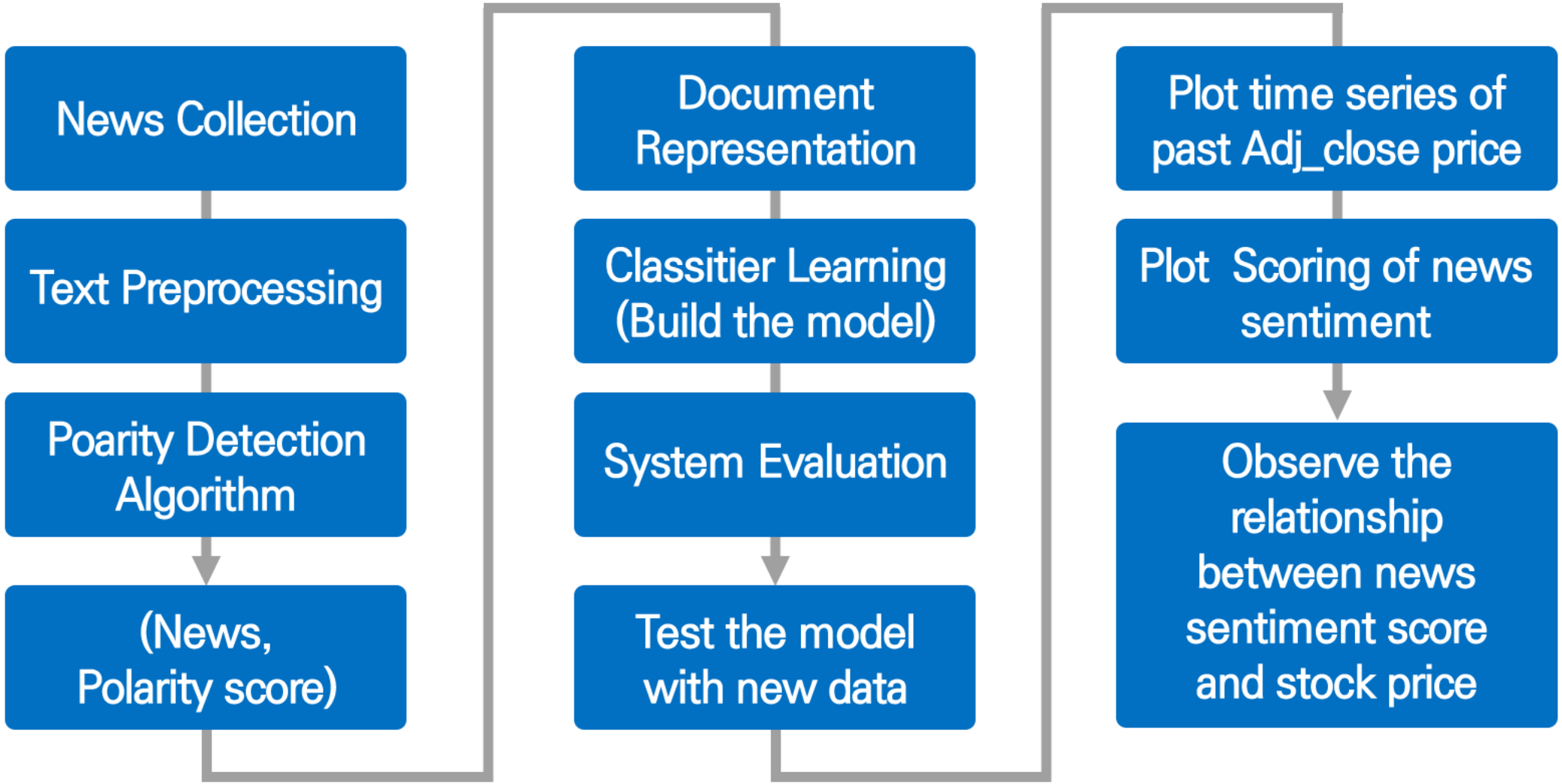
- Support Vector Machine : 90%
- Random Forest : 80%
- Naïve Bayse : 75%

높은 예측력을 보임!

출처 로고 이미지 : https://namu.wiki/w/Apple

NLP와 주가

☑ 뉴스 감성분석과 주가 예측



NLP와 주가

☑ 트위터 메시지와 주가 예측

[트위터(StockTwits) 메시지도 정보가 될 수 있다!]

감성분석 실시

- N-grams와 BN synsets를 조합한 감성분석의 예측력 : 약 **72%**

가장 높게 나타남!

출처 로고 이미지 : https://namu.wiki/w/%ED%8A%B8%EC%9C%84%ED%84%B0?from=Twitter

NLP와 주가

☑ 문장구조 분석과 주가 예측

금융 뉴스 헤드라인에서 문장 패턴을 추출

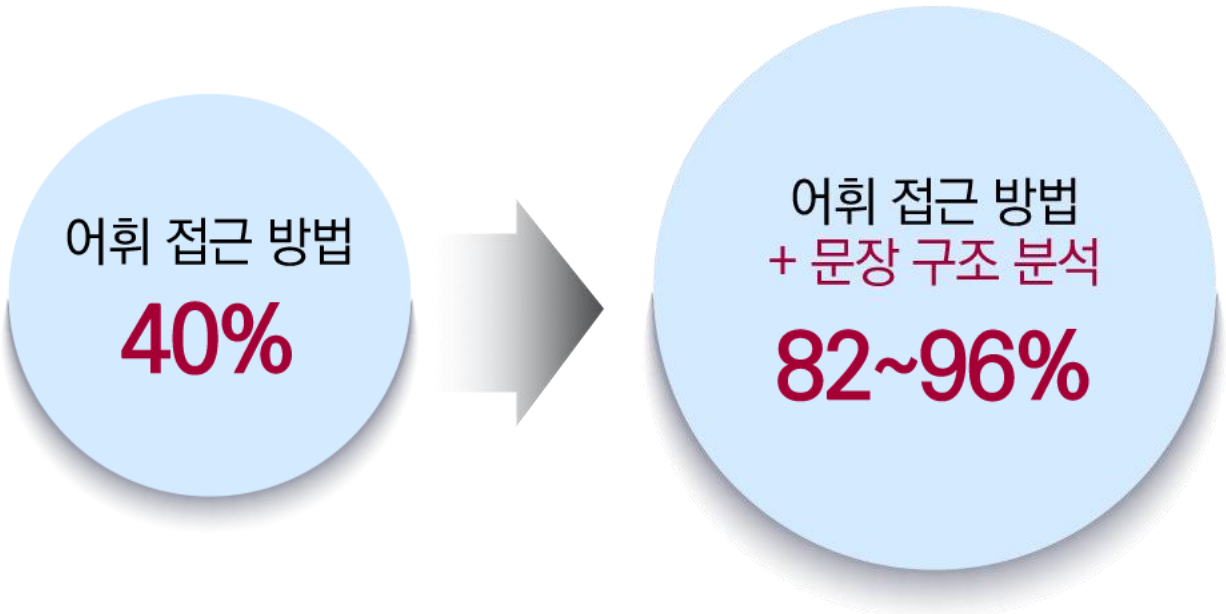
↓

주가 예측력 파악

어휘 접근 방법 + 문장 구조 분석

NLP와 주가

☑ 문장구조 분석과 주가 예측



문장 구조를 분석하는 것이 어휘만 분석하는 것보다 **더 높은 정보**를 가지고 있음

NLP와 주가

☑ BERT와 주가 예측

BERT Bidirectional Encoder Representations for Transformers

Google

구글에서 2018년 개발한 딥러닝 모델

- 자연어처리 분야에서 가장 우수한 성능
- 트렌스포머(Transformer)에 기반을 둔 모델
- 사전학습 후 특정 목적을 위해 Fine-Tuning하여 적용
- 양방향 모델이 문장의 앞뒤 문맥을 동시에 고려

BERT방법론과 **거시경제 데이터**를 함께 활용해 **예측력이 우수**

출처 로고 이미지 : https://ko.wikipedia.org/wiki/%EA%B5%AC%EA%B8%80

NLP와 주가

☑ 자연어처리의 활용

" 자연어처리에 관한
많은 논문이 연구,발표되고 있고
투자에도 활용되고 있다. "

Unit 1 Word Embedding & Word2vec

요약

1. NLP
2. Embedding
3. Word2Vec
4. GloVe
5. 전처리 과정
6. NLP를 활용한 주가 예측

요약

? NLP를 통한 뉴스가 주가 예측에
새로운 펙터가 될 수 있을까

설명력이 있다!

- 기존의 Pricing Theory에서 사용되지 않았던
새로운 차원의 정보를 다루고 있는 영역

- 많은 사람들이 관심을 가지고 활용하며
발전하고 있는 영역