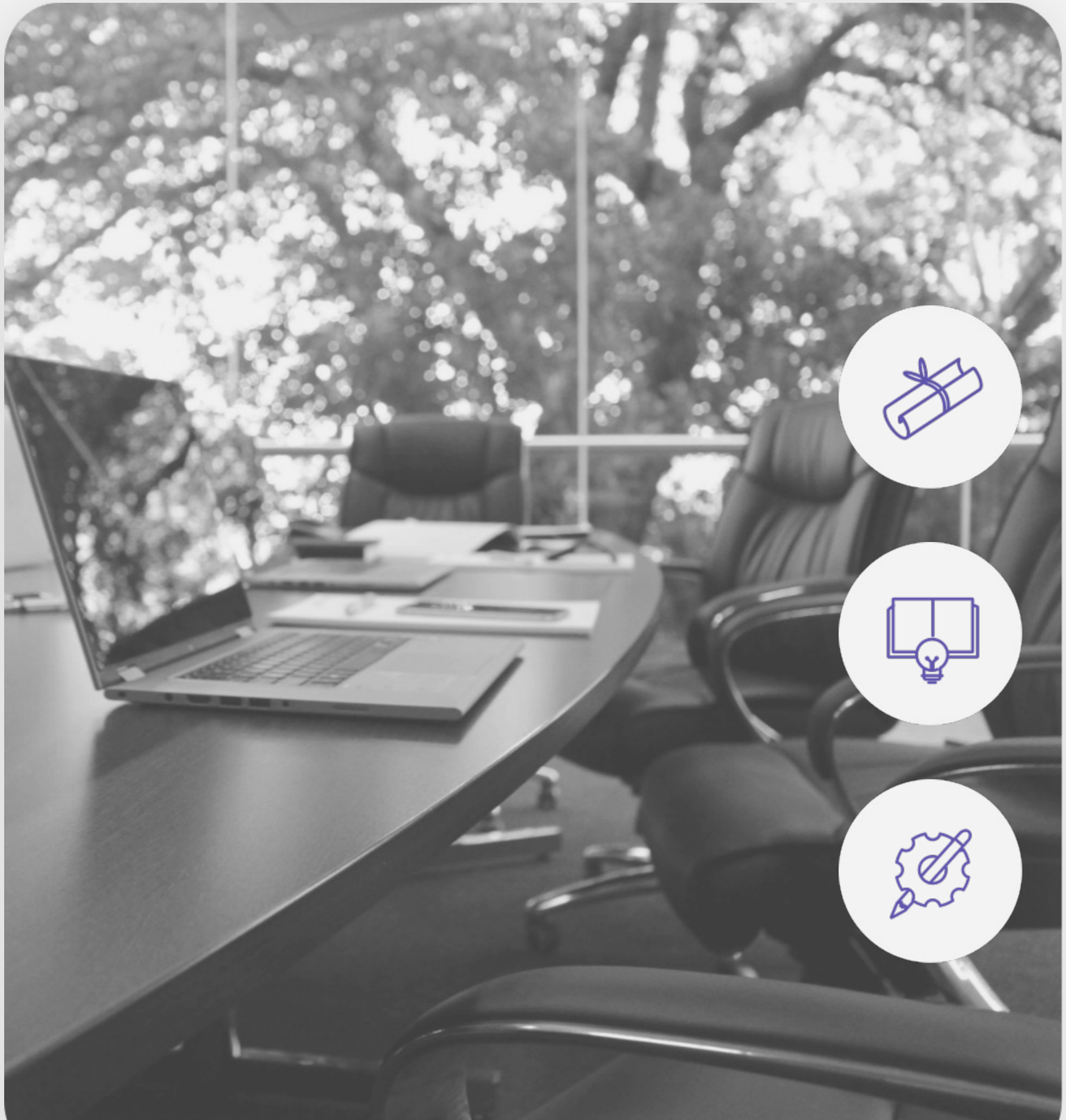
데이터마이닝_적용기법 제안

데이팅앱 대표 3사 비교: Google Play Store 리뷰를 중심으로

비즈니스인포매틱스학과 허지연 (2021100538)
비즈니스인포매틱스학과 조지윤 (2022166254)

2022.10.07



Contents

I. 토픽모델링 : Top2vec

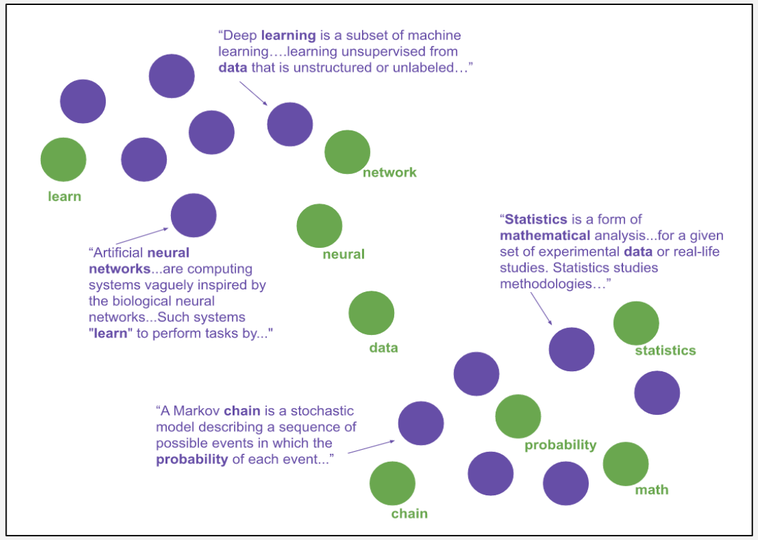
II. 감정분석 : Vader Algorithm

토픽모델링이란?

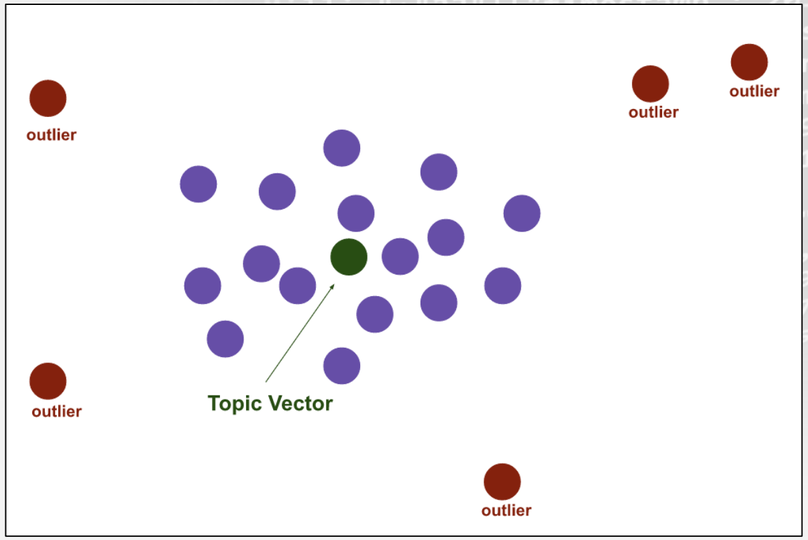
- ✓ 텍스트 데이터를 자동으로 분석하여 문서 세트의 클러스터 단어를 결정하는 기계 학습 기술.
- ✓ 사전 정의된 태그 목록이나 train data set이 필요하지 않기 때문에 **비지도 머신러닝.**
- ✓ 단어 빈도 및 단어 간 거리와 같은 패턴을 감지하여 유사한 피드백과 가장 자주 나타나는 단어 및 표현을 클러스터링.

Top2Vec?

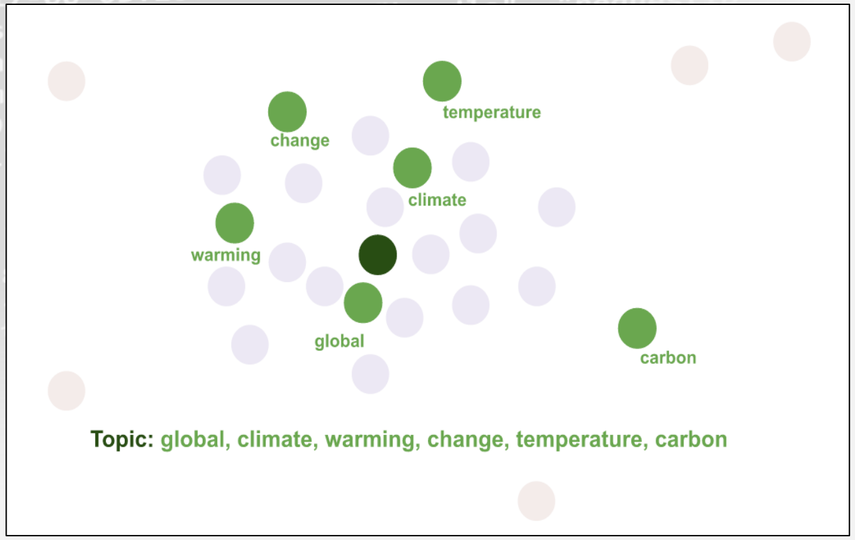
- ✓ 토픽모델링을 하는 동시에 의미론적 서치가 가능한 알고리즘.
- ✓ 텍스트에 있는 주제를 자동으로 임베딩시켜 감지.



Doc2Vec, Bert Sentence Transformer,
Universal Sentence Encoder



Topic vector
각 밀집되어 있는 document vector들의 중심 값



N-closest word vectors = topic words

Top2Vec?

- ✓ 따라서 Top2Vec을 사용하면 다음과 같은 결과를 뽑는 것이 가능함.
  1) 주제 수
  2) 주제
  3) 각 주제의 사이즈
  4) 계층적 주제 → 클러스터를 합치는 것이 가능해짐
  5) 주제별 문서, 키워드 검색 가능
  6) 유사한 문서, 키워드 검색 가능

감성분석이란?

텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 분석하는 과정

1. Lexicon 기반
2. Machine Learning 기반

Vader(Valence Aware Dictionary and sEntiment Reasoner) Algorithm

- ✓ Lexicon 기반 Rule -based sentiment analysis tool
- ✓ 단어마다 10명의 사람들에게 조사한 평균 sentiment score 가 기록되어있음
- ✓ 한 글자로 이루어진 단어는 제외시키고 구두점이나 느낌표, 물음표 등등의 글자가 있는 지 확인
- ✓ Rule - based 코딩
  1) Aren't 나 cannot 과 같은 부정적 동사를 체크하여 -0.74 가중치를 곱함
  2) 대문자는 강조의 의미로 보고 positive 일 경우 0.733을 더하고 negative의 경우 뺌
  3) Amazingly, barely와 같은 단어들은 수식어로 명사에 가중치를 부여함. 긍정일 경우 0.293을 더하고, 부정일 경우 뺌
  4) 추가적으로 수식어의 위치에 따라 다른 가중치가 부여됨. 단어의 앞앞단어라면 0.9를 곱하고 단어의 앞 단어라면 0.95의 가중치를 곱함
  5) But 의 경우 반전을 의미함으로 단어 앞에 있을 경우 0.5 뒤에 있을 경우 1.5를 곱함
  6) 느낌표의 경우 강조의 의미이므로 느낌표 하나당 0.295점을 줌 . 최대 4개까지 점수를 부여한 후 단어 점수에 더함
  7) -1~1 사이로 normalization = compound score 로 나타남
  8) 7번까지 계산된 sentiment socre를 비율로 나타낸다. pos score/ (pos score + neu score + neg score)

감성분석이란?

텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 분석하는 과정

1. Lexicon 기반
2. Machine Learning 기반

Vader(Valence Aware Dictionary and sEntiment Reasoner) Algorithm

[예시]

```
sentence = "VADER is pretty good at identifying the underlying sentiment of a text!"
```

```
# Output of example1
{'neg': 0.0, 'neu': 0.585, 'pos': 0.415, 'compound': 0.75}
```

Thank you!