블로그 포스팅 주제 분류 13주차

데이터분석 캡스톤디자인
2017100898 박혜원

분류기 변경

Decision Tree → K Means 분류기 변경

-지도 학습에 대한 부담감
-방법론에 대한 의심

K Means 공부

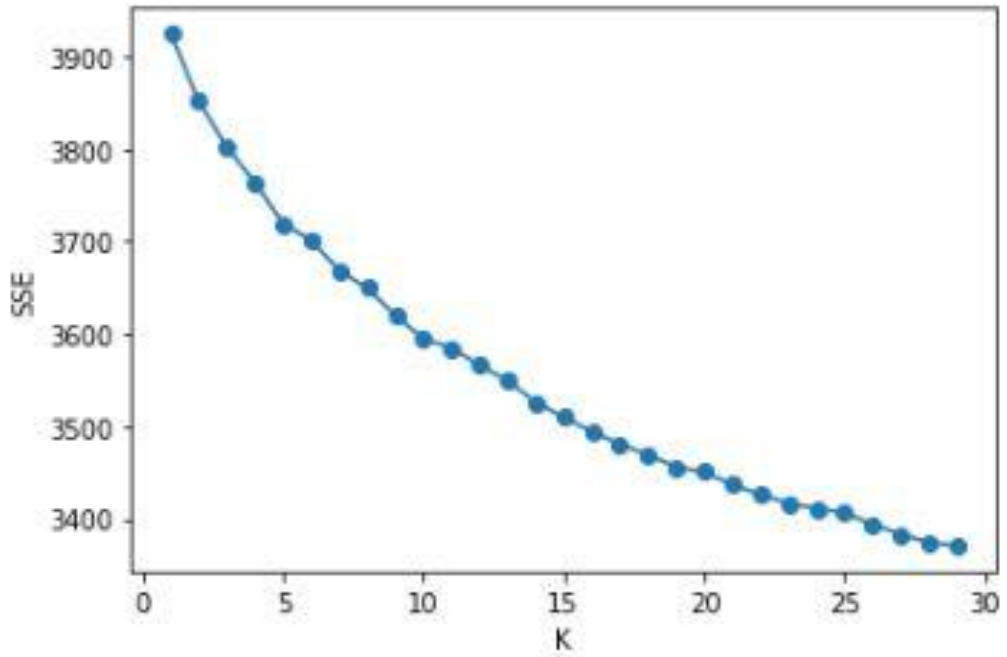
Kmeans 란?

1. 군집 개수 K 개수 정해주면, 랜덤으로 K개의 중심 선정
2. 각 개체별로 K개의 중심 중 제일 가까운 중심으로 clustering 됨
3. 나눠진 Cluster 내 객체별로 다시 중심점 선택
4. 선택된 중심점으로 다시 Clustering 재편
5. Label 변화 없을 때까지 반복

비지도 학습, 레이블 없이 학습 가능, 계산비용 낮음, 안정적 성능 => 큰 규모의 데이터 군집화에 적합

LDA와의 차이 : 문서의 토픽 개수에 대한 가정
LDA는 한 문서에 여러 종류 토픽 존재 가능 가정
Kmeans는 하나의 문서, 하나의 토픽 가정

K Means



K 개수 선정에 대한 어려움

클러스터내의 오차제곱합(SSE)를 이용해 K 결정하는 elbow 기법 사용

SSE 급격하게 변화 거의 없는 곳으로 K 결정해주면 되지만 급격하게 줄어드는 곳 판단 어려움

K Means

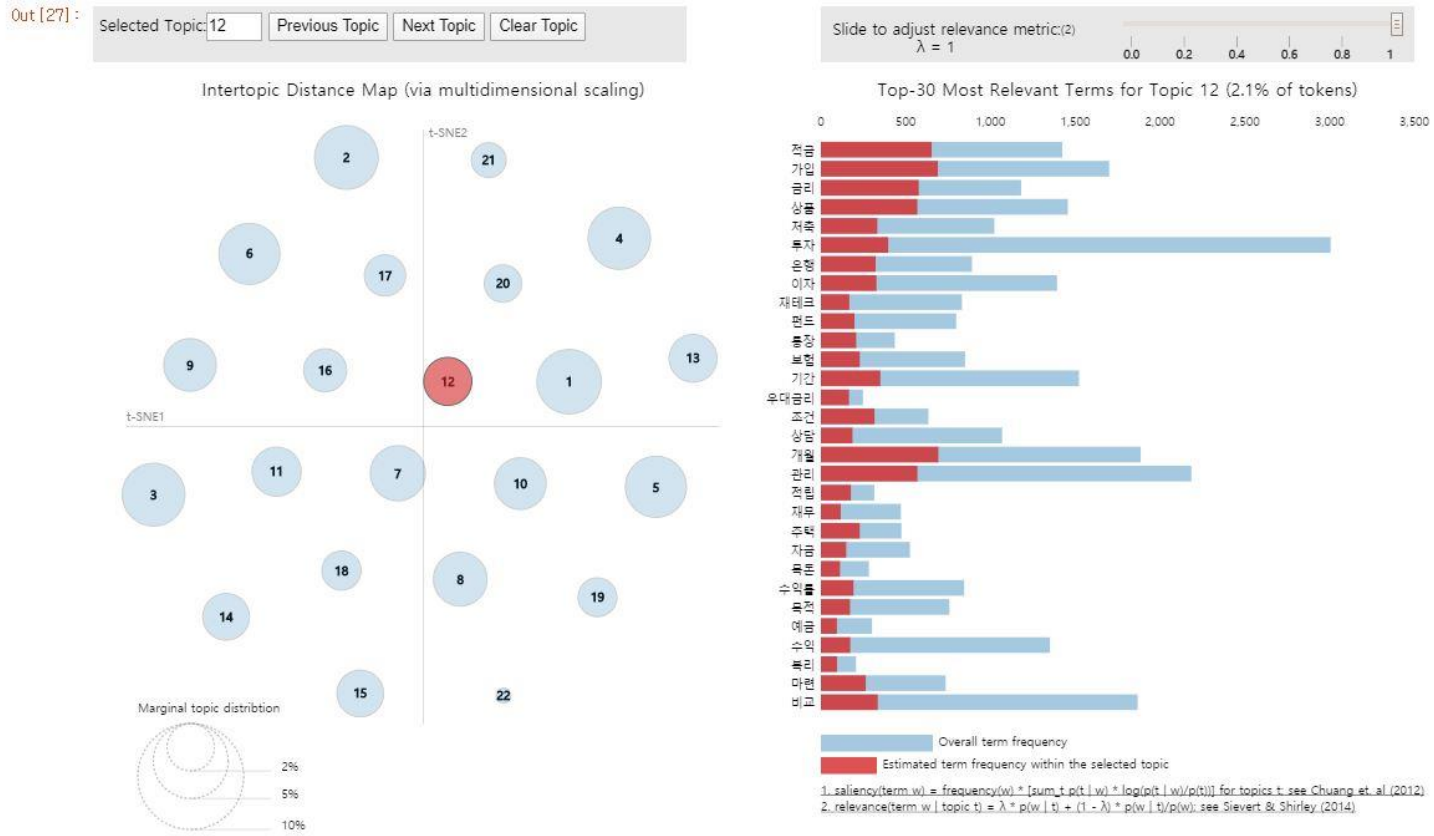
LDA와 동등한 조건 위해 우선 22개의 Topic으로 진행

Kmeans는 LDA와 다르게 하나의 포스팅 – 하나의 주제 라는 조건 갖고 있기 때문에
Pet/Baby , Beauty/Fashion, Drama/Movie 등 비슷한 주제도 모두 나눠 줌

```
other 추가 정답률:83 / 132  ============== 62.878787878787875%
other 제외 정답률:83 / 115  ============== 72.17391304347827%
```

결과

시각화 결과


Out [27] :
Selected Topic: 12
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
t-SNE2
t-SNE1
Marginal topic distribtion
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
Top-30 Most Relevant Terms for Topic 12 (2.1% of tokens)
적금
가입
금리
상품
저축
투자
은행
이자
재테크
펀드
통장
보험
기간
우대금리
조건
상담
개월
관리
적립
재무
주택
자금
목돈
수익률
목적
예금
수익
특리
마련
비교
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

다음주 계획

K Means 통한 타 블로그 주제 분류 및 내 블로그 주제 결정

K Means accuracy 평가

분류기 보완 및 수정 (K 개수 설정 / LDA topic 개수 설정)

LDA 와 K Means의 비교 분석 보고서 작성