블로그 포스팅 주제 분류 14주차

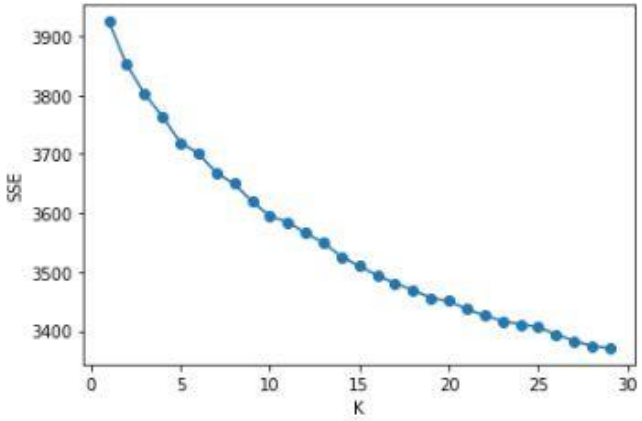
데 이 터 분 석 캡 스 톤 디 자 인
2 0 1 7 1 0 0 8 9 8 박 혜 원

13주차까지의 정리

Decision Tree -> Kmeans 분류기 변경
Kmeans 공부

K개수 선정 위해 elbow 기법 사용
→ 오차제곱합(SSE) 변화로
K개수 결정하는 방법

But 급격한 변화가 안 보인다는 문제



13주차까지의 정리

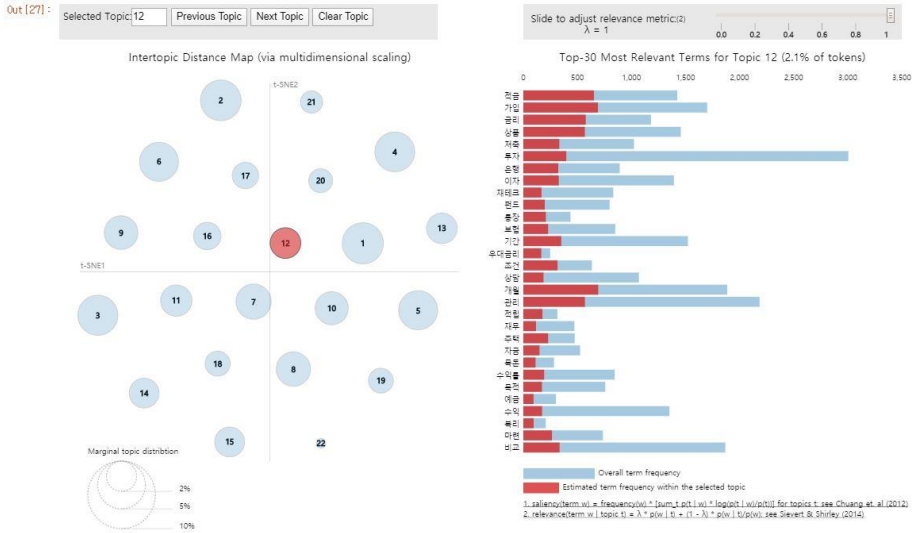
임의로 LDA 모델과 같은 **22개 Topic**으로 분석 진행해 봄

- LDA 모델과 마찬가지로 없는 Topic 몇 개 발생 (DIY, music, book)
- Kmeans는 LDA와 다르게 1문서당 1개의 주제만 있다고 가정
- 단어들이 조금 더 명확함 (영화/드라마 Topic, 육아/애견 Topic 구별)

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

13주차까지의 정리

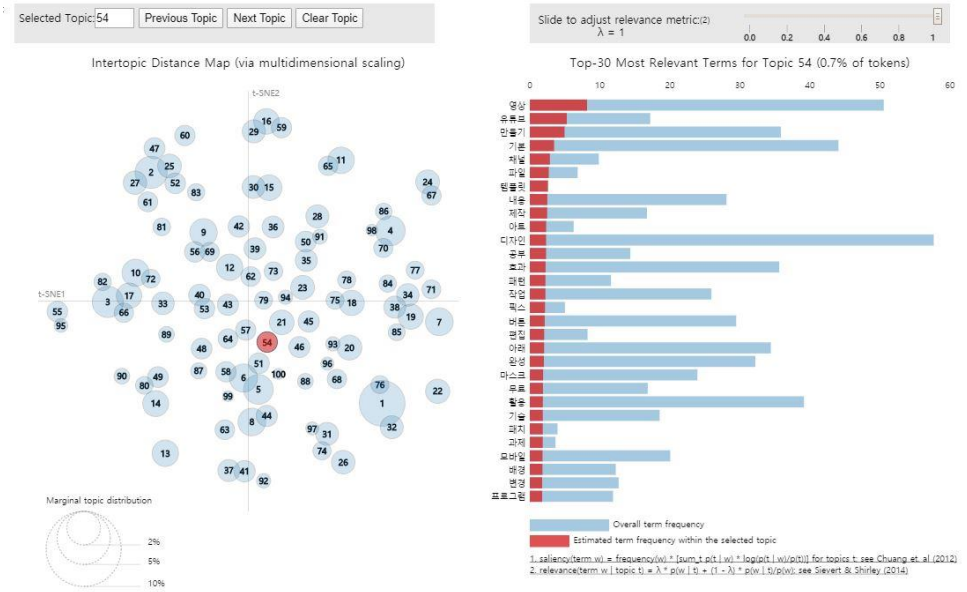
시각화 결과



14주차 KMeans

K를 임의적으로
매우 큰 수로 설정 ->

후처리를 통해 K 축소하는
방법으로 진행해 봄



14주차 KMeans

훨씬 다양한 주제 나옴 (LDA에서 도출되지 않았던 animation, music, book 포함)
후처리 : 100개 k 주제별로 묶고 라벨링
-> other 포함 18개의 topic

	blogger	topic	real_topic	check
50		delicious	delicious	True
51		IT	IT	True
52		baby	baby	True
53		fashion	fashion	True
54		movie	movie	True
55		car	car	True
56		delicious	travel	False

```
other 추가 정답률:98 / 132  ================ 74.24242424242425%
other 제외 정답률:98 / 128  ================ 76.5625%
```

14주차 KMeans

내 블로그 주제
LDA와 같은 결과인 delicious (맛집)

```
{'delicious': 52, 'IT': 21, 'travel': 10, '
delicious
```

다음주 계획

Coherence score (주제일관성 평가)

코드 정리하기

LDA / Kmeans 비교 보고서 작성

깃허브 / 클래스룸 업로드