10주차 과제

Topic modeling

담당교수님: 윤장혁 교수님

과목명: Data Analytics

이름: 박민성

학번: 201611145

전공: 산업공학과

제출일: 2020.05.22

Object

- 주어진 6개의 문장들(data_week10.txt)을 이용하여 Topic Modeling 수행
    - LDA(Latent Dirichlet Allocation)를 활용
    - 최적의 Topic 수를 고민 후 결정
        - 단, 최적의 Topic 수로 선정한 이유를 설명 (수리적으로 접근할 필요는 없음)
    - 최종 Topic들의 이름을 Labeling
        - 연관된 Keyword들을 고려하여 적합한 이름을 부여

LDA(Latent Dirichlet Allocation)

python module을 사용하여 분석을 진행하였다. 널리 쓰이는 genism, tomotopy 등 모듈 존재

불용처리를 할 필요가 없고, 데이터가 간단한 단어들로만 이루어져 있다고 생각해서 비교적 간단한 tomotopy로 분석을 진행했다.

```
> Users > minisong > Desktop > 10주차.py > ...
1   import tomotopy as tp
2   mdl = tp.LDAModel(k=3)
3   for line in open('C:\\Users\\minisong\\Desktop\\data_week10.txt'):
4       mdl.add_doc(line.strip().split())
5   for i in range(100):
6       mdl.train()
7       #print('Iteration: {}\tLog-likelihood: {}'.format(i, mdl.ll_per_word))
8   for k in range(mdl.k):
9       print('Top 2 words of topic #{}'.format(k))
10      print(mdl.get_topic_words(k, top_n=2))
```

최적의 Topic 수를 고민 후 결정

Topic의 수는 문장의 수가 많지 않고 몇 가지 'k'값과 'top_n'값을 설정해서 코드를 실행해본 경험을 토대로 3가지 정도가 적당하다고 생각해서 최적의 Topic 수를 3으로 정의했다.

최종 Topic들의 이름을 Labeling

```
Top 2 words of topic #0
[('eat', 0.372986376285553), ('cake', 0.372986376285553)]
Top 2 words of topic #1
[('bread', 0.5936883687973022), ('rice', 0.3964496850967407)]
Top 2 words of topic #2
[('kitty', 0.4938574731349945), ('hamster', 0.4938574731349945)]
```

총 term이 7개이고, document의 길이도 짧기 때문에 주제 안에서 값이 가장 높은 top_2 word만 조합해서 Labling을 하기로 하였다.

Result

Topic 0 – eat cake

Topic 1 – rice bread

Topic 2 – hamster kitty

(순서는 주제의 좋고 나쁨의 의미가 없음)

더 복잡한 모델이라면, Topic Coherence, Perplexity의 측정을 통해 내가 선정한 주제가 적절한지 알아봐야 한다. 하지만 이 분석에서는 고려하지 않았다.

Document 안에 많은 단어가 포함되어 있지 않고 data 자체가 단순하기 때문에 이러한 결과가 나온 것 같다.