

Count기반 단어 표현 모델

Bag of Words(BOW) 단어의 빈도로 텍스트 표현

Bag of words 문서 내 모든 단어들(토큰)을 가방 하나에 모두 집어넣고 사용 자주 언급된 단어일 수록 가방에서 나올 확률이 높아진다 (높은 빈도의 단어일 수록 중요하다고 판단)

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

워드 클라우드

단어-가중치 테이블을 시각화로 표현

단어 점수	
0	수학 70
1	과학 20
2	국어 100
3	영어 40
4	사회 60
5	체육 90



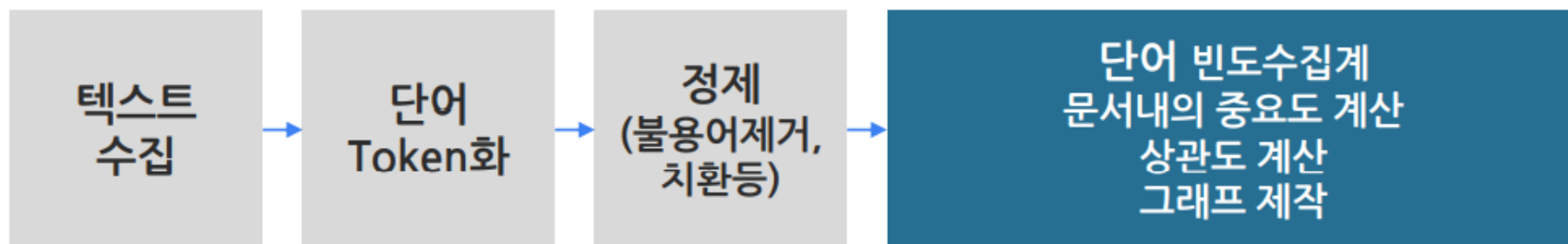
과학
국어
수학
영어
체육

프로그램이 단어를 가중치를 통해 인식한다

TF-IDF 해당 표현이 문서에서 얼마나 중요한가?

단어의 빈도(Term Frequency)와 역 문서 빈도(Inverse Document Frequency)를 토대로 **특정 문서 내에 어떤 단어가 얼마나 중요한 지**를 나타내는 통계적 수치
어떤 문서군(서류 더미)을 기준으로 삼느냐에 따라 단어의 중요도가 달라진다.

- TF-IDF는 주로 문서의 유사도를 구하는 작업
- 검색 시스템에서 검색 결과의 중요도를 정하는 작업
- 문서 내에서 특정 단어의 중요도를 구하는 작업 등



TF-IDF 해당 표현이 문서에서 얼마나 중요한가?

TF-IDF

$$tf\ idf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

TF

$tf(t, d)$ = 문서 d 에서 단어 t 가 언급된 횟수

보정1 - 불린 빈도

$tf(t, d)$ = 문서 d 에서 단어 t 가 한 번이라도 언급되었다면 1, 아니면 0

보정2 - 로그 스케일 빈도

$$tf(t, d) = \log(f(t, d) + 1)$$

보정3 - 증가 빈도

$$tf(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max\{f(w, d): w \in d\}}$$

IDF

$$idf(t, D) = \frac{\text{단어 } t \text{가 포함된 문서의 수}}{\text{전체 문서}(D) \text{의 수}}$$

$$idf(t, D) = \frac{\text{전체 문서의 수}}{\text{단어 } t \text{가 포함된 문서의 수}}$$

보정 - zero division + 로그 스케일

$$idf(t, D) = \log\left(\frac{\text{전체 문서의 수}}{1 + \text{단어 } t \text{가 포함된 문서의 수}}\right)$$

TF-IDF 해당 표현이 문서에서 얼마나 중요한가?

TF

	해리 포터	생일	호그와트	도비	마법	디멘터	마법사의 돌	비밀의 방	시리우스 블랙	빗자루	불사조 기사단	볼드모트
해리포터와 마법사의 돌	4	1	4	0	2	0	2	0	0	3	0	2
해리포터와 비밀의 방	5	0	4	1	3	0	0	2	0	3	0	3
해리포터와 불사조 기사단	5	0	3	1	3	2	0	0	1	3	3	4

IDF

	해리 포터	생일	호그와트	도비	마법	디멘터	마법사의 돌	비밀의 방	시리우스 블랙	빗자루	불사조 기사단	볼드모트
IDF 계산	=3/3	=3/1	=3/3	=3/2	=3/3	=3/1	=3/1	=3/1	=3/1	=3/3	=3/1	=3/3
IDF	1	3	1	1.5	1	3	3	3	3	1	3	1

	해리 포터	생일	호그와트	도비	마법	디멘터	마법사의 돌	비밀의 방	시리우스 블랙	빗자루	불사조 기사단	볼드모트
해리포터와 마법사의 돌	4	3	4	0	2	0	6	0	0	3	0	2
해리포터와 비밀의 방	5	0	4	1.5	3	0	0	6	0	3	0	3
해리포터와 불사조 기사단	5	0	3	1.5	3	6	0	0	3	3	9	4

TF-IDF 해당 표현이 문서에서 얼마나 중요한가?

	해리 포터	호그와트	마법	트롤	절대반지	사우론	호빗	모르도르
마법사의 돌	4	4	2	1	0	0	0	0
비밀의 방	5	4	3	0	0	0	0	0
반지 원정대	0	0	3	1	4	3	3	4
두개의 탑	0	0	3	0	4	3	1	4

	해리 포터	호그와트	마법	트롤	절대반지	사우론	호빗	모르도르
IDF계산	=4/2	=4/2	=4/4	=4/2	=4/2	=4/2	=4/2	=4/2
IDF	2	2	1	2	2	2	2	2

	해리 포터	호그와트	마법	트롤	절대반지	사우론	호빗	모르도르
마법사의 돌	8	8	2	2	0	0	0	0
비밀의 방	10	8	3	0	0	0	0	0
반지 원정대	0	0	3	2	8	6	6	8
두개의 탑	0	0	3	0	8	6	2	8

어떤 문서군을 지정 하느냐에 따라
키워드의 가중치가 달라진다

[참고]

워드클라우드

```
txt='강아지 산책 강아지 목욕 강아지 미용 강아지 쇼핑 친구와 저녁 먹음 가족과 점심 먹음 혼자 저녁 먹음 친구와 쇼핑'
```

연관분석

```
dataset=[['사과', '치즈', '생수'],  
          ['생수', '호두', '치즈', '고등어'],  
          ['수박', '사과', '생수'],  
          ['생수', '호두', '치즈', '옥수수']]
```

TF-IDF

```
docs=['파이썬 차트 파이썬 머신러닝',  
      '차트 파이썬 R 차트',  
      'R 분석 시각화']
```