한글 문장의 유사도 계산 (4)

사용된 대용량 말뭉치: KCCq28_Q01.txt (31.29MB, 20만줄)

사용된 입력 문장: 안녕하세요 이것은 테스트 문장입니다.

사용된 유사도 계산식: 코사인 유사도

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Ex. 코사인 유사도를 예시를 들어 해석

1. 나는 사과 좋아요 : 나는(1), 사과(1), 좋아요(1)
2. 나는 좋아요 좋아요: 나는(1), 좋아요(2)

코사인 유사도 분모:

sqrt(나는 개수^2 + 사과 개수^2 + 좋아요 개수^2) * sqrt(나는 개수^2 + 좋아요 개수^2)

= sqrt((1^2+1^2+1^2) * (1^2+2^2)

= sqrt(3 * 5)

= sqrt(15)

코사인 유사도 분자:

((문장1: 나는 개수) * (문장2: 나는 개수)) + ((문장1: 사과 개수) * (문장2: 사과 개수))

+ ((문장1: 좋아요 개수) * (문장2: 좋아요 개수))

= 1*1 + 1*0 + 1*2

= 1+2

= 3

⇨ 코사인 유사도 3/sqrt(15)

⇨ 문장 유사도 100 * 3/sqrt(15)

코드 구현:

➔ 멀티 프로세싱을 이용해 형태소 분석 및 유사도 측정을 하기 위해 코드 내부에서 대용량 말뭉치 문장들을 코어 개수에 맞게 분할

Ex. CPU 코어 개수가 4개, 말뭉치 문장이 20만줄이 있다고 가정

> 200000/4 = 50000 : 5만줄짜리 텍스트 리스트 4개를 생성

➔ 각 코어당 분할된 문장 리스트를 주고 형태소 분석 및 문장 유사도 검사를 시킨다.

➔ 유사도 검사가 끝나면 각 코어에서 유사도가 높은 상위 2문장씩 출력해준다.

실행 방법:

python sync_test.py [input-text] [-option]

➔ Option: -okt , -komoran , -kkma

어떤 형태소 분석기를 사용할 지 옵션으로 선택

➔ 입력 문장으로는 "안녕하세요 이것은 테스트 문장입니다." 를 터미널에 타이핑하였다.

➔ python sync_test.py KCCq28_Q01.txt -komoran

- 실행시간: 약 104초 -> 약 1분 44초

```
E:₩학교₩대학교₩4학년 1학기₩빅데이터₩4주자_문장_유사도_검사>python sync_test.py KCCq28_Q01.txt
 -komoran
입력 문장: 안녕하세요 이것은 테스트 문장입니다.
Process ID:  21672
권미진은 3일 자신의 블로그에 "안녕하세요.권미진입니다. :  48.67 %
이때 제갈량은 유비에게 "위연은 반골의 상입니다. :  44.19 %
Process ID:  21380
한편 ID crystalmovie님은 "제 시각은 부정적입니다. :  44.19 %
이 대변인은 "착각하지 마라.새누리당이 집권당이다. :  43.3 %
Process ID:  18488
이에 김 장관은 "네 주거복지 로드맵입니다. :  45.64 %
만약 그런 일이 실제 있다면 그것은 우연입니다"매우 역설적으로 들립니다. :  43.3 %
Process ID:  15052
이 의원은 또 "이것은 국민의당도 똑같다. :  47.25 %
이 의원은 "이것이 최선의 방법이라고 생각하지는 않습니다. :  43.3 %
--- 104.46190881729126 seconds ---
```

➔ python sync_test.py KCCq28_Q01.txt -okt

- 실행시간: 약 251초 -> 약 4분 11초

```
E:₩학교₩대학교₩4학년 1학기₩빅데이터₩4주자_문장_유사도_검사>python sync_test.py KCCq28_Q01.txt
 -okt
입력 문장: 안녕하세요 이것은 테스트 문장입니다.
Process ID:  21580
이 의원은 "이 모든 것이 표 때문에 그렇다. :  48.67 %
늙기 싫어하는 것은 인지상정.늙지 않는 것은 불가능한 것일까.빌 앤드루스는 늙지 않는 것이 과학
적으로 "가능한 일"이라고 말한다. :  46.42 %
Process ID:  9348
이것이 의미하는 것이 있을 것"이라고 설명했다. :  48.54 %
한 트위터리안은 "150만원이 중저가라는 것이 놀랍다. :  45.64 %
Process ID:  8988
박 대통령은 "중요한 것은 말이 아니라 실천이다. :  45.64 %
이것은 꼬리자르기를 하겠다는 것"이라고 주장했다. :  45.64 %
Process ID:  6380
이 의원은 또 "이것은 국민의당도 똑같다. :  53.03 %
이 의원은 "이것이 최선의 방법이라고 생각하지는 않습니다. :  46.29 %
--- 251.19628596305847 seconds ---
```

➔ python sync_test.py KCCq28_Q01.txt -kkma

- 실행시간: 약 2883초 -> 약 45분

```
E:₩학교₩대학교₩4학년 1학기₩빅데이터₩4주자_문장_유사도_검사>python sync_test.py KCCq28_Q01.txt
 -kkma
입력 문장: 안녕하세요 이것은 테스트 문장입니다.
Process ID:  21224
이 의원은 "이것이 최선의 방법이라고 생각하지는 않습니다. :  45.18 %
다음은 일문일답.  -새해 들어 이건희 회장이 주문한 내용은.  "회장님은 끝없이 도전하는 분이다.
:  44.72 %
Process ID:  16836
권 위원장은 "이 재판관 후임은 대법원장이 추천하고 대통령이 임명하는 구조다. :  46.17 %
한국은 어떤 에너지 정책을 도입해야 할까."원자력 발전은 계속해서 많은 국가들에 중요한 에너지원
이 될 것이다. :  44.91 %
Process ID:  8872
권미진은 3일 자신의 블로그에 "안녕하세요.권미진입니다. :  48.69 %
선거운동원들은 안 대표가 인사를 하면 "안녕하십니까.기호 3번 안철수입니다. :  45.44 %
Process ID:  17860
골고루 먹어야 해요." "가공하지 않은 신선한 음식을 많이 먹는 것도 중요하죠.이 부분은 부모님의
도움이 필요합니다. :  48.45 %
파란색 점퍼에 더불어민주당 배지와 어깨띠를 착용한 김 의원은 이곳을 찾은 주부를 향해 "안녕하세
요.김현미입니다. :  46.63 %
--- 2883.035451412201 seconds ---
```

결론:

- 형태소 분석기마다 다르게 형태소를 나누기 때문에 유사도가 다 다르게 나온다.
- 가장 빠른 형태소 분석기는 실행시간에서 알 수 있듯이 komoran이다.