한글 문장의 유사도 계산 (3)

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

사용된 입력 문장: 안녕하세요 이것은 유사도 검사 테스트입니다.

사용된 유사도 계산식: 코사인 유사도

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

실행 방법:

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

- ➔ Option은 입력과 가장 유사한 문장을 몇 개 추출할 것인지 정하는 것이므로 유사도 높은 상위 10개를 출력하고 싶으면 -10을 주면 된다.
- ➔ 여기서는 KCCq28_Q01.txt파일을 input으로 주었으며 유사도가 높은 상위 10개를 출력하기 위해 옵션으로 -10을 주었음
- ➔ 입력 문장으로는 "안녕하세요 이것은 유사도 검사 테스트입니다." 를 터미널에 타이핑하였다.
- ➔ python sync_test.py KCCq28_Q01.txt -10
- ➔ 실행시간: 약 2725초 -> 약 45분 41초
- ➔ 만약 100만줄 input text를 주게 되면 20만줄에 46분쯤 걸리므로 230분, 즉 3시간 50분쯤 걸린다.

```
E:₩학교₩대학교₩4학년 1학기₩빅데이터₩3주차_문장_유사도_검사>python sync_test.py KCCq28_Q01.txt -10
안녕하세요 이것은 유사도 검사 테스트입니다.
앞서 김 의원은 이날 특별검사 추천권을 야당이 갖도록 한 특검법안 원안 수정을 요구하며 "촛불은 촛불일 뿐이지 결국 바람이 불면 다 꺼지게 돼 있다"고 촛불 민심을 깎아내리는 듯한 발언을 해 논란을 예고했다. :  24.21 %
여기 이 홀로그램이 움직이는지, 빛에 비추면 숨은 그림이 보이는지도 확인하세요."31일 오전 서울 남대문시장에서 이 국장 등 23명의 한은 직원들이 상인들을 대상으로 위조지폐 판별법을 알려주는 캠페인을 가졌다. :  24.06 %
이에 대해 새정치민주연합 등 야당은 "용산 기지 이전 계획과 연합 토지 관리계획은 국회의 동의를 받은 한미 양국의 협정인 만큼 이를 변경하려면 반드시 국회의 동의를 받아야한다"고 강조하고 있다. :  23.94 %
한은은 "내년에는 석유류 가격의 물가하락 영향이 다소 약해지면서 올해보다 소비자 물가가 높아질 것"이라고 예상했다. :  23.93 %
관련기사"고체 엔진·탄두 짱짱 생산하라"최근까지 북한에 대해 말폭탄을 던졌던 트럼프와 틸러슨이 잇따라 유화적인 발언을 쏟아냈다. :  23.88 %
그는 "임기 4년을 8년처럼 일하는 농협 회장이 되겠다"고 포부를 밝혔다. :  23.87 %
다 함께 카메라 앞에 모여 생일축하 노래를 부른 뒤 "우리는 하나다"를 외친 뒤, 주먹감자를 날린 것이다. :  23.82 %
그런 점을 의식한 듯 문 위원장은 이 후보자에게 덕담을 하면서도 "청문회는 협상의 대상이 아닌 비판의 대상이란 점을 잊지 말라"고 충고했다. :  23.74 %
같은 당 안민석 의원도 "메르스와 가뭄이 추경의 요인인데 난데 없이 철도 이런게 32개가 들어와 있다. :  23.73 %
한 비박계 재선 의원은 "유 원내대표를 이렇게 몰아내려는 것은 명분이 없다"면서 "당청 관계 복원 노력 등을 포함해 유승민 원내대표의 결단을 기다려야 한다"고 말했다. :  23.7 %
time:  2725.0027112960815
```