String 숙제 (2023년)

1. 주제: TF-IDF와 Cosine 유사도를 이용한 문서 유사도 분석

○ BoW(Bag of Words) 모델
- 단어들의 순서는 고려하지 않고, 단어들의 출현 빈도수로 문서의 특징을 표현
- 문서에 포함된 단어들의 TF-IDF 값을 구한 후, cosine 유사도를 이용하여 문서들 사이의 거리를 측정

○ TF-IDF의 정의
- $TF(x, y) = \frac{\text{단어 } x\text{가 문서 } y\text{에 나타나는 빈도수}}{\text{문서 } y\text{의 모든 단어가 나타나는 빈도수의 합}}$
- $IDF(x) = \log_e(\frac{\text{전체 문서의 수}}{\text{단어 } x\text{가 나타나는 문서의 수}})$
- TFIDF(x, y) = TF(x, y) * IDF(x)

○ 입력: N개의 문서 제목과 문서의 내용

```
d1
eat apple eat grape
d2
eat banana eat grape
d3
sweet green grape
d4
pretty fruit box
```

○ 각 문서마다 TFIDF 벡터를 계산
- 문서 d1의 TF 값

eat	apple	grape
2/4 = 0.5	1/4 = 0.25	1/4 = 0.25

- eat, apple, grape의 IDF 값

eat	apple	grape
log(4/2) = 0.693	log(4/1) = 1.386	log(4/2) = 0.288

- 문서 d1의 TFIDF 벡터

eat	apple	grape
0.347	0.347	0.072

- 문서 d2의 TFIDF 벡터

eat	banana	grape
0.347	0.347	0.072

○ 문서의 특징을 행벡터로 표현할 수 있으므로, Cosine 유사도 사용

- Cosine 유사도 = $\frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$
  - 문서 1과 2의 유사도 = $\frac{0.347^2 + 0.072^2}{\sqrt{0.347^2 + 0.347^2 + 0.072^2} * \sqrt{0.347^2 + 0.347^2 + 0.072^2}}$
    = 0.51054
  - d1과 나머지 문서 사이의 유사도
    (d2, 0.51054) (d3, 0.02108) (d4, 0.00000)

○ 유사한 문서 선정

- Target 문서와 유사도가 가장 높은 k개의 문서 제목을 출력
- 단, TF-IDF 계산에 불용어는 포함하지 않을 것
- 예: Target 문서가 d1이며, k = 2일 경우 다음과 같이 출력
  1. d2 (유사도=0.51054)
  2. d3 (유사도=0.02108)

○ 입력:

- 파일 이름, 추천 문서 수, Target 문서 제목? s.txt 2 d1
  - 입력 파일의 구성:
    문서 1의 제목
    문서 1의 본문
    문서 2의 제목
    문서 2의 본문
    ...

○ 프로그램을 작성할 때 유의 사항

- 문서의 본문은 소문자로 변환한 후, TFIDF 계산할 것 (제목은 소문자로 변환된 상태임)
- 본문을 단어로 분리하기 위하여 String.split() 메소드를 사용하며, split()의 인자로 다음과 같은 정규식을 사용할 것: split("[,.?!:\"\\s]+"))
- 불용어는 "stopwords.txt" 파일로 제공되므로, 프로젝트 폴더에 복사한 후 프로그램에서 open하여 사용할 것
- 단어는 String.hashCode()로 변환한 후 TF-IDF 벡터에 사용할 것

2. 제출 내용: HW2.java 하나의 파일만 제출

① public class HW2 (파일 내에 나머지 클래스들은 public이 아님)
② default package 사용
③ **프로그램의 첫줄에 자신의 학번과 이름을 주석으로 추가**할 것. 그 외의 모든 주석은 삭제하기를 바랍니다.
④ 한글 encoding은 MS949로 설정 (제발~~)
← **위의 조건들을 만족하지 않는 과제물은 심사하지 않음!!**

3. 평가: 50점 만점

- Target 문서의 (단어의 hashCode(), TF-IDF 값)를 hashCode의 오름차순으로 출력 (30점)
- Target과 유사도가 가장 높은 k개의 문서에 대해 문서 제목와 유사도 출력. (20점)

- 앞부분이 틀리면, 그 이후는 0점 처리
- JDK 8로 컴파일하며, 프로그램 구성이나 성능에 심각한 문제가 있으면 실행 결과와 관계없이 감점 처리함 (**과제 1의 feedback 반영할 것!**)
- 무작위 샘플링을 통하여 선택된 학생에 대해서는 대면 평가를 수행하며, 자신의 코드를 제대로 설명하지 못하면 감점 (과제 0점 처리 및 최종성적 한 등급 하향 조정까지 가능) 처리함
- Copy로 판단되는 과제에 대해서는 성적을 한 등급 하향 조정

4. 동작의 예

실행의 예 1:

파일 이름, k, 문서 제목: s.txt 2 d1 결과 1. "d1"의 TF-IDF 벡터 [(100184, 0.347) (93029210, 0.347) (98615627, 0.072)] 결과 2. "d1"과(와) 유사한 2개의 문서 1. d2 (유사도=0.51054) 2. d3 (유사도=0.02108)	d1 eat apple eat grape d2 eat banana eat grape d3 sweet green grape d4 pretty fruit box

실행의 예 2:

파일 이름, k, 문서 제목: half.txt 10 the dark knight rises

결과 1. “the dark knight rises“의 TF-IDF 벡터

[(-2110163466, 0.123) (-2025444940, 0.189) (-1988594550, 0.225) (-1631971150, 0.194) (-1396165339, 0.310) (-1352163903, 0.128) (-1335386765, 0.225) (-1292876679, 0.130) (-1274446437, 0.169) (-1240095096, 0.169) (-1224441231, 0.150) (-1206091918, 0.142) (-1126830451, 0.146) (-1106754295, 0.100) (-1022424254, 0.153) (-982670050, 0.080) (-906017470, 0.200) (-704305337, 0.149) (-412625767, 0.184) (-309012785, 0.219) (-228897266, 0.143) (108960, 0.049) (117694, 0.030) (3016246, 0.209) (3053931, 0.148) (3075958, 0.100) (3079687, 0.194) (3307367, 0.164) (3314342, 0.097) (3456416, 0.109) (95457908, 0.072) (96653192, 0.114) (102744716, 0.086) (114851798, 0.061) (137728614, 0.184) (288961422, 0.128) (552120030, 0.124) (765915793, 0.102) (848184146, 0.126) (965542266, 0.150) (1827208126, 0.082)]

결과 2. “the dark knight rises“과(와) 유사한 10개의 문서

1. batman returns (유사도=0.27010)
2. the dark knight (유사도=0.24973)
3. batman: under the red hood (유사도=0.22613)
4. batman unmasked: the psychology of the dark knight (유사도=0.22046)
5. batman (유사도=0.21698)
6. batman: the dark knight returns, part 2 (유사도=0.20015)
7. batman forever (유사도=0.17594)
8. batman beyond: return of the joker (유사도=0.16933)
9. batman: mask of the phantasm (유사도=0.16738)
10. batman: the dark knight returns, part 1 (유사도=0.15758)

실행의 예 3:

파일 이름, k, 문서 제목: half.txt 5 star wars: episode i - the phantom menace
결과 1. “star wars: episode i - the phantom menace“의 TF-IDF 벡터
[(-1413356420, 0.608) (-891980137, 0.381) (-683151209, 0.559) (-377016647, 0.707) (-306987569, 0.389) (3125652, 0.280) (3258112, 0.592) (3443937, 0.312) (97618667, 0.325) (109519319, 0.416) (115168792, 0.145) (537538120, 0.365) (1099842154, 0.294) (1643358201, 0.707)]

결과 2. “star wars: episode i - the phantom menace“과(와) 유사한 5개의 문서
1. star wars: episode iii - revenge of the sith (유사도=0.30554)
2. star wars: episode ii - attack of the clones (유사도=0.27165)
3. return of the jedi (유사도=0.17849)
4. the empire strikes back (유사도=0.13768)
5. the journey of august king (유사도=0.12238)