과제 #7

201904479 컴퓨터 전자 시스템 학과 임서연

1. 3쪽에서 생성된 naver_review.txt 파일을 읽어서 2쪽에 있는 데이터들을 계산. 단어의 수는 공백으로 분리된 어절의 수를 의미

	단어수	문장수	고유 단어
naver_ review.txt	총 단어수 : 1509809	[193390 rows x 1 columns] 총 문장수 : 193390	고유단어수 : 56723

2. 단어수를 10,000개와 20,000개로 늘려 4쪽의 sentencepiece를 실행하고 90001~90010번의 평에 대한 결과를 얻음

단어수 10,000

```
박성웅은 역시 연기를 너무 잘해
['_박', '성', '웅', '은', '_역시', '_연기를', '_너무', '_잘', '해']
[351, 8377, 8912, 8297, 288, 1241, 23, 63, 8323]

지구 탄생 이래 최초의 인류의 공동 프로젝트
['_지구', '_탄생', '_이래', '_최초', '의', '_인', '류의', '_공', '동', '_프로', '젝', '트']
[2997, 4375, 2271, 4137, 8294, 77, 2886, 120, 8371, 1202, 9508, 8472]

여자 개미 호바가 진짜 귀여움
['_여자', '_개', '미', '_호', '바', '가', '_진짜', '_귀여움']
[312, 74, 8317, 552, 8448, 8285, 54, 5940]

화려한 배우들 모아놓고 액션도 좋고 스토리도 좋았고 눈이 즐거웠다,,한국판 오션스일레븐?
['_화려한', '_배우들', '_모아', '놓고', '_액션도', '_좋고', '_스토리도', '_좋았', '고', '_눈이', '_즐거', '웠', '다', ',,', '한국', '판', '_오', '션', '스', '일', '레', '븐', '?']
[2135, 581, 6621, 962, 1881, 845, 809, 7476, 8280, 2359, 1605, 9410, 8278, 245, 2810, 8559, 72, 8489, 8312, 8368, 8411, 9221, 8329]

어디론가 다녀온 듯한 여운
['_어디', '론가', '_다녀', '온', '_듯한', '_여운']
[646, 1327, 6404, 8579, 1998, 549]

새로운 느낌의 드라마...음악이 한 몫 했다...
['_새로운', '_느낌의', '_드라마', '...', '음악', '이', '_한', '_', '몫', '_했다', '...']
[1601, 4343, 114, 8, 1543, 8277, 37, 8275, 9420, 1505, 8]

아저씨를 이을 최고의 액션영화
['_아저씨', '를', '_이', '을', '_최고의', '_액션영화']
[2488, 8331, 6, 8301, 200, 2989]

00신다리~ㅋㅋㅋ 귀여워
['_00', '신', '다리', '~', 'ㅋㅋㅋ', '_귀여워']
[411, 8385, 7620, 8341, 326, 4924]


.............
['_', '.........', '.....']
[8275, 585, 559]

정말 좋은영화네요...
['_정말', '_좋은영화', '네요', '...']
[42, 1790, 39, 8]
```

단어수 20,000

```
박성웅은 역시 연기를 너무 잘해
['_박', '성', '웅', '은', '_역시', '_연기를', '_너무', '_잘', '해']
[351, 8377, 8912, 8297, 288, 1241, 23, 63, 8323]

지구 탄생 이래 최초의 인류의 공동 프로젝트
['_지구', '_탄생', '_이래', '_최초', '의', '_인', '류의', '_공', '동', '_프로', '젝', '트']
[2997, 4375, 2271, 4137, 8294, 77, 2886, 120, 8371, 1202, 9508, 8472]

여자 개미 호바가 진짜 귀여움
['_여자', '_개', '미', '_호', '바', '가', '_진짜', '_귀여움']
[312, 74, 8317, 552, 8448, 8285, 54, 5940]

화려한 배우들 모아놓고 액션도 좋고 스토리도 좋았고 눈이 즐거웠다,,한국판 오션스일레븐?
['_화려한', '_배우들', '_모아', '놓고', '_액션도', '_좋고', '_스토리도', '_좋았', '고', '_눈이', '_즐거', '웠', '다', ',,', '한국', '판', '_오', '션', '스', '일', '레', '븐', '?']
[2135, 581, 6621, 962, 1881, 845, 809, 7476, 8280, 2359, 1605, 9410, 8278, 245, 2810, 8559, 72, 8489, 8312, 8368, 8411, 9221, 8329]

어디론가 다녀온 듯한 여운
['_어디', '론가', '_다녀', '온', '_듯한', '_여운']
[646, 1327, 6404, 8579, 1998, 549]

새로운 느낌의 드라마...음악이 한 몫 했다...
['_새로운', '_느낌의', '_드라마', '...', '음악', '이', '_한', '_', '몫', '_했다', '...']
[1601, 4343, 114, 8, 1543, 8277, 37, 8275, 9420, 1505, 8]

아저씨를 이을 최고의 액션영화
['_아저씨', '를', '_이', '을', '_최고의', '_액션영화']
[2488, 8331, 6, 8301, 200, 2989]

00신다리~ㅋㅋㅋ 귀여워
['_00', '신', '다리', '~', 'ㅋㅋㅋ', '_귀여워']
[411, 8385, 7620, 8341, 326, 4924]


..............
['_', '.........', '.....']
[8275, 585, 559]

정말 좋은영화네요...
['_정말', '_좋은영화', '네요', '...']
[42, 1790, 39, 8]
```

3. naver_review.txt 파일 내용에 대해 Okt 형태소 분석기를 실행시킴. 결과에서 나타난 고유 단어수를 계산. 90001~90010번의 평에 대한 결과에 대해 Okt를 실행하고 위 2번의 결과와 비교함.

Okt 형태소 분석기를 실행

```
박성웅은 역시 연기를 너무 잘해
['박성웅', '은', '역시', '연기', '를', '너무', '잘', '해']

지구 탄생 이래 최초의 인류의 공동 프로젝트
['지구', '탄생', '이래', '최초', '의', '인류', '의', '공동', '프로젝트']

여자 개미 호바가 진짜 귀여움
['여자', '개미', '호바', '가', '진짜', '귀', '여', '움']

화려한 배우들 모아놓고 액션도 좋고 스토리도 좋았고 눈이 즐거웠다,,한국판 오션스일레븐?
['화려한', '배우', '들', '모아놓고', '액션', '도', '좋고', '스토리', '도', '좋았고', '눈', '이', '즐거웠다', ',,', '한국판', '오션스일레븐', '?']

어디론가 다녀온 듯한 여운
['어디', '론', '가', '다녀온', '듯', '한', '여운']

새로운 느낌의 드라마...음악이 한 몫 했다...
['새로운', '느낌', '의', '드라마', '...', '음악', '이', '한', '몫', '했다', '...']

아저씨를 이을 최고의 액션영화
['아저씨', '를', '이을', '최고', '의', '액션영화']

00신다리~ㅋㅋㅋ 귀여워
['00', '신다리', '~', 'ㅋㅋㅋ', '귀여워']


..............
['..............']

정말 좋은영화네요...
['정말', '좋은', '영화', '네', '요', '...']
```

4. 위의 1~3의 결과와 무관하게 새로 <구어체(2).txt> 파일에서 영어와 한국어에 대해 sentencepiece를 적용하여 32,000단어를 추출함. 이 경우 영어와 한국어 단어를 각각 32,000개씩 추출함. <구어체(2)> 파일에서 한국어 부분에는 영어 단어도 포함되어 있으니 영어를 제외하고 한국어 단어만을 사용하여야 함.

5. <구어체(2)> 문장 중 110,500~110,510 번째 문장에 대해 sentencepiece를 수행하고 그 결과를 구함.

110,500~110,510 번째 문장에 대해 sentencepiece

```
정말 신기하게도 공통점이 전혀 없는 형이에요
['_정말', '_신기', '하게도', '_공통점이', '_전혀', '_없는', '_형', '이에요']
[218, 5625, 5399, 19254, 1321, 598, 551, 96]

정말 심각하게 좋은 날이네요
['_정말', '_심각하게', '_좋은', '_날이', '네요']
[218, 11455, 222, 2484, 355]

정말 심각한 문제는 이들이 식물의 뿌리를 갉아 먹어서 식물의 뿌리에 알을 낳는 붕어가 알을 낳지 못해요
['_정말', '_심각한', '_문제는', '_이들이', '_식물의', '_뿌리를', '_', '갉', '아', '_먹어서', '_식물의', '_뿌', '리에', '_알을', '_낳', '는', '_붕', '어가', '_알을
[218, 3610, 1907, 9368, 28679, 22679, 30781, 0, 30807, 10495, 28679, 2633, 962, 19779, 4593, 30784, 9446, 1243, 19779, 4593, 30791, 3555]

정말 아름다운 전시회입니다
['_정말', '_아름다운', '_전시회', '입니다']
[218, 1542, 8461, 20]

정말 염치없지만 말하고 싶은 게 있습니다
['_정말', '_염', '치', '없', '지만', '_말하고', '_싶은', '_게', '_있습니다']
[218, 2471, 30916, 30873, 67, 3240, 616, 186, 41]

정말 예측 불가능하죠 그렇지 않아요
['_정말', '_예측', '_불가능', '하죠', '_그렇지', '_않아요']
[218, 3295, 1817, 4287, 2625, 689]

정말 오랜 친구인 것처럼 느꼈고 자매처럼 보였어요
['_정말', '_오랜', '_친구인', '_것처럼', '_느꼈고', '_자매', '처럼', '_보였어요']
[218, 1714, 14636, 1398, 19983, 19810, 440, 7541]

정말 오랜만에 느끼는 한가한 오후입니다
['_정말', '_오랜만에', '_느끼는', '_한가한', '_오후', '입니다']
[218, 3663, 4342, 22964, 1294, 20]


 정말 오랜만에 연락드리네요
 ['_정말', '_오랜만에', '_연락', '드리', '네요']
 [218, 3663, 337, 1247, 355]

 정말 오랜만에 찾아뵙게 되었습니다
 ['_정말', '_오랜만에', '_찾아뵙', '게', '_되었습니다']
 [218, 3663, 11095, 30811, 792]

 정말 오랜만에 최고의 휴식을 했습니다
 ['_정말', '_오랜만에', '_최고의', '_휴식을', '_했습니다']
 [218, 3663, 1442, 4082, 658]
```