README

[과제 4] 한글 문장의 유사도 계산 (4)

4. 문장 유사도 계산을 워드 임베딩 기법 등 다양한 방법으로 구현

- 입력: 한글 문장 1개
- 방법: 형태소 분석기 2가지와 WPM, 총 3가지 이상의 방법으로 실험

  정확도가 높은, 또는 실행 속도 빠른 방법은 무엇인지(각 방법들의 특징 또는 문제점) 비교 또는 실험 결과 설명
- 3분 내외 동영상 제작(유투브에 업로드): 구현 방법 설명 및 실행 내용
- 제출물: 소스코드 및 실행파일, 보고서(실행속도 및 실행화면 스샷 포함),

  “유투브 동영상 링크”

<참고1> 입력문장과 말뭉치의 각 문장들을 워드 임베딩 기법 등을 이용하여 문장 벡터를 만들고, 벡터 유사도 계산 기법을 이용하는 방법을 추가한 경우 가산점을 부여함.

<참고2> 실행속도를 매우 빠른 새로운 방법을 사용하여 구현한 경우에

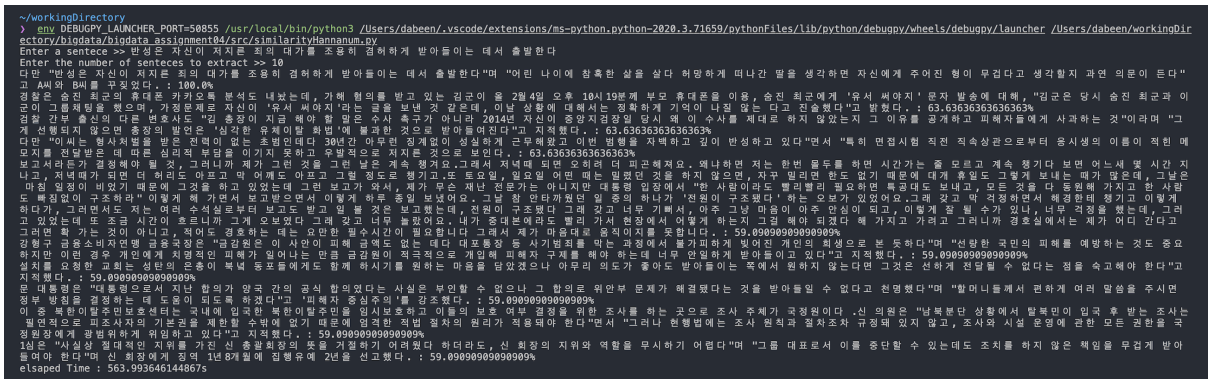
가산점을 부여함. 멀티 쓰레딩 기법, DBMS 등 다양한 기법 사용 가능.

YOUTUBE 링크

```
https://youtu.be/9yDd7ApHtqA
```

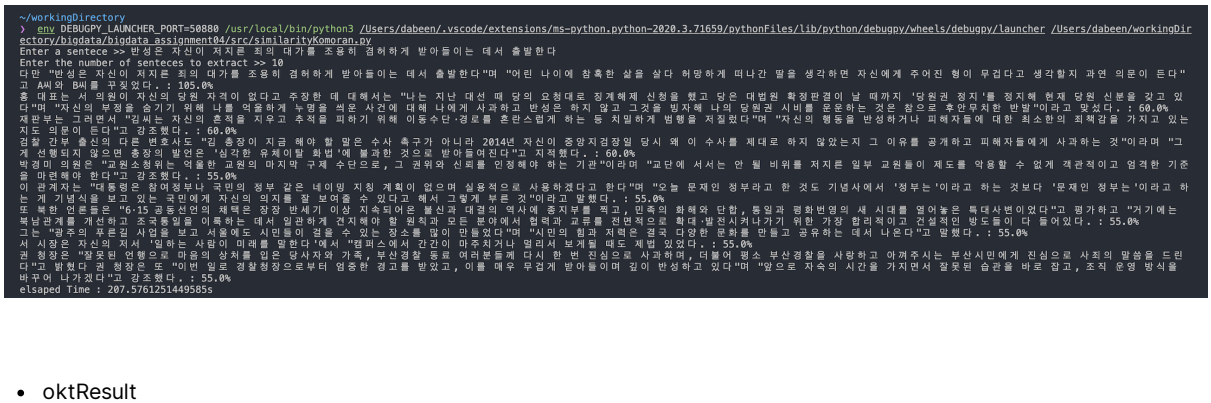
실행 결과

- HannanumResult



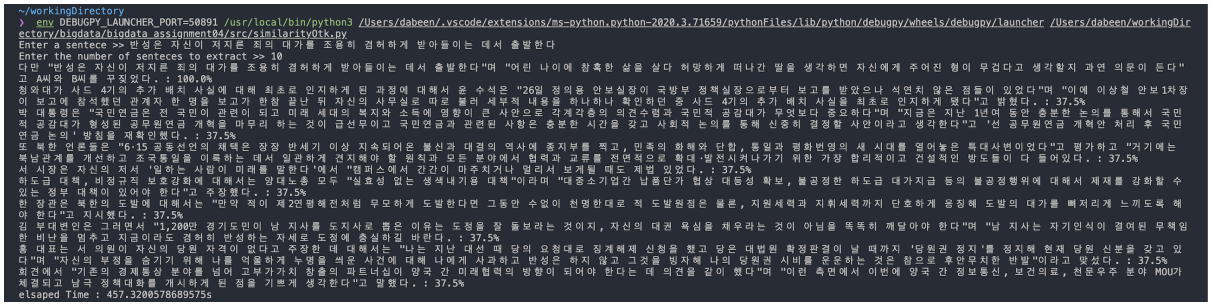
```
~/workingDirectory
❯ env DEBUGPY_LAUNCHER_PORT=50855 /usr/local/bin/python3 /Users/dabeen/.vscode/extensions/ms-python.python-2020.3.71659/pythonFiles/lib/python/debugpy/wheels/debugpy/launcher /Users/dabeen/workingDirectory/bigdata/bigdata_assignment04/src/similarityHannanum.py
Enter a sentence >> 반성은 자신이 저지른 죄의 대가를 조용히 겸허하게 받아들이는 데서 출발한다
Enter the number of senteces to extract >> 10
다만 "반성은 자신이 저지른 죄의 대가를 조용히 겸허하게 받아들이는 데서 출발한다"며 "어린 나이에 참혹한 삶을 살다 허망하게 떠나간 딸을 생각하면 자신에게 주어진 형이 무겁다고 생각할지 과연 의문이 든다"고 A씨와 B씨를 꾸짖었다. : 100.0%
경찰은 숨진 최근의 휴대폰 카카오톡 분석도 내놨는데, 가해 혐의를 받고 있는 김군이 올 2월4일 오후 10시19분께 부모 휴대폰을 이용, 숨진 최군에게 '유서 써야지' 문자 발송에 대해, "김군은 당시 숨진 최군과 이군이 그룹채팅을 했으며, 가정문제로 자신이 '유서 써야지'라는 글을 보낸 것 같은데, 이날 상황에 대해서는 정확하게 기억이 나질 않는 다고 진술했다"고 밝혔다. : 63.636363636363636%
검찰 간부 출신의 다른 변호사도 "김 총장이 지금 해야 할 말은 수사 촉구가 아니라 2014년 자신이 중앙지검장일 당시 왜 이 수사를 제대로 하지 않았는지 그 이유를 공개하고 피해자들에게 사과하는 것"이라며 "그게 선행되지 않으면 총장의 발언은 '심각한 유체이탈 화법'에 불과한 것으로 받아들여진다"고 지적했다. : 63.636363636363636%
다만 "이씨는 형사처벌을 받은 전력이 없는 초범인데다 30년간 아무런 징계없이 성실하게 근무해왔고 이번 범행을 자백하고 깊이 반성하고 있다"면서 "특히 면접시험 직전 직속상관으로부터 응시생의 이름이 적힌 메모지를 전달받은 데 따른 심리적 부담을 이기지 못하고 우발적으로 저지른 것으로 보인다. : 63.636363636363636%
보고서라든가 결정해야 될 것, 그러니까 제가 그런 것을 그런 날은 계속 했거요.그래서 저녁때 되면 오히려 더 피곤해져요. 왜냐하면 저는 한번 몰두를 하면 시간가는 줄 모르고 계속 했기다 보면 어느새 몇 시간 지나고, 저녁때가 되면 더 허리도 아프고 막 어깨도 아프고 그럴 정도로 했기고,또 토요일, 일요일 이런 때는 밀렸던 것을 하지 않으면, 자꾸 밀리면 한도 없기 때문에 대개 휴일도 그렇게 보내는 때가 많은데, 그날은 마침 일정이 비었기 때문에 그것을 하고 있었는데 그런 보고가 와서, 제가 무슨 재난 전문가는 아니지만 대통령 입장에서 '한 사람이라도 빨리빨리 필요하면 특공대도 보내고, 모든 것을 다 동원해 가지고 한 사람도 빠짐없이 구조하라" 이렇게 해 가면서 보고받으면서 이렇게 하루 종일 보냈어요. 그날 참 안타까웠던 일 중의 하나가 '전원이 구조됐다' 하는 오보가 있었어요.그래 갖고 막 격정하면서 해경한테 챙기고 이렇게 하다가, 그러면서도 저는 여러 수석실로부터 보고도 받고 일 볼 것은 보고했는데, 전원이 구조됐다 그래 갖고 너무 기뻐서, 아주 그냥 마음이 아주 안심이 되고, 이렇게 잘 될 수가 있나, 너무 걱정을 했는데, 그러고 있었는데 또 조금 시간이 흐르니까 그게 오보였다 그래 갖고 너무 놀랐어요. 내가 중대본에라도 빨리 가서 현장에서 어떻게 하는지 그럼 해야 되겠다 해 가지고 가려고 그러니까 경호실에서는 제가 어디 간다고 그러면 확 가는 것이 아니고, 적어도 경호하는 데는 요만한 필수시간이 필요합니다 그래서 제가 마음대로 움직이지를 못합니다. : 59.09090909090909%
김현구 금융소비자연맹 금융국장은 "금감원은 이 사안이 피해 금액도 없는 데다 대포통장 등 사기범죄를 막는 과정에서 불가피하게 빚어진 개인의 희생으로 본 듯하다"며 "선량한 국민의 피해를 예방하는 것도 중요하지만 이런 경우 개인에게 처절적인 피해가 일어나는 만큼 금감원의 적극적으로 개입해 피해자 구제를 해야 하는데 너무 안일하게 받아들이고 있다"고 지적했다. : 59.09090909090909%
설치를 요청한 교회는 성탄의 은총이 북녘 동포들에게도 함께 하시기를 원하는 마음을 담았겠으나 아무리 의도가 좋아도 받아들이는 쪽에서 원하지 않는다면 그것은 선하게 전달될 수 없다는 점을 숙고해야 한다"고 지적했다. : 59.09090909090909%
문 대통령은 "대통령으로서 지난 합의가 양국 간의 공식 합의였다는 사실은 부인할 수 없으나 그 합의로 위안부 문제가 해결됐다는 것을 받아들일 수 없다고 천명했다"며 "할머니들께서 권하게 여러 말씀을 주시면 정부 방침을 결정하는 데 도움이 되도록 하겠다"고 '피해자 중심주의'를 강조했다. : 59.09090909090909%
이 중 북한이탈주민보호센터는 국내에 입국한 북한이탈주민을 임시보호하고 이들의 보호 여부 결정을 위한 조사를 하는 곳으로 조사 주체가 국정원이다 .신 의원은 "남복분단 상황에서 탈북민이 입국 후 받는 조사는 필연적으로 피조사자의 기본권을 제한할 수밖에 없기 때문에 엄격한 적법 절차의 원리가 적용돼야 한다"면서 "그러나 현행법에는 조사 원칙과 절차조차 규정돼 있지 않고, 조사와 시설 운영에 관한 모든 권한을 국정원장에게 광범위하게 위임하고 있다"고 지적했다. : 59.09090909090909%
1심은 "사실상 절대적인 지위를 가진 신 총괄회장의 뜻을 거절하기 어려웠다 하더라도, 신 회장의 지위와 역할을 무시하기 어렵다"며 "그룹 대표로서 이를 중단할 수 있는데도 조치를 하지 않은 책임을 무겁게 받아들여야 한다"며 신 회장에게 징역 1년 8개월에 집행유예 2년을 선고했다. : 59.09090909090909%
elsaped Time : 563.9936461448675
```

- KomoranResult



```
~/workingDirectory
❯ env DEBUGPY_LAUNCHER_PORT=50880 /usr/local/bin/python3 /Users/dabeen/.vscode/extensions/ms-python.python-2020.3.71659/pythonFiles/lib/python/debugpy/wheels/debugpy/launcher /Users/dabeen/workingDirectory/bigdata/bigdata_assignment04/src/similarityKomoran.py
Enter a sentence >> 반성은 자신이 저지른 죄의 대가를 조용히 겸허하게 받아들이는 데서 출발한다
Enter the number of senteces to extract >> 10
다만 "반성은 자신이 저지른 죄의 대가를 조용히 겸허하게 받아들이는 데서 출발한다"며 "어린 나이에 참혹한 삶을 살다 허망하게 떠나간 딸을 생각하면 자신에게 주어진 형이 무겁다고 생각할지 과연 의문이 든다"고 A씨와 B씨를 꾸짖었다. : 105.0%
홍 대표는 서 의원이 자신의 당원 자격이 없다고 주장한 데 대해서는 "나는 지난 대선 때 당의 요청대로 징계해제 신청을 했고 당은 대법원 확정판결이 날 때까지 '당원권 정지'를 정지해 현재 당원 신분을 갖고 있다"며 "자신의 부정을 숨기기 위해 나를 억울하게 누명을 씌운 사건에 대해 나에게 사과하고 반성은 하지 않고 그것을 방자해 나의 당원권 '시비를 운운하는 것은 참으로 후안무치한 반발"이라고 맞섰다. : 60.0%
재판부는 그러면서 "김씨는 자신의 존적을 지우고 추적을 피하기 위해 이동수단·경로를 혼란스럽게 하는 등 치밀하게 범행을 저질렀다"며 "자신의 행동을 반성하거나 피해자들에 대한 최소한의 죄책감을 가지고 있는지도 의문이 든다"고 강조했다. : 60.0%
검찰 간부 출신의 다른 변호사도 "김 총장이 지금 해야 할 말은 수사 촉구가 아니라 2014년 자신이 중앙지검장일 당시 왜 이 수사를 제대로 하지 않았는지 그 이유를 공개하고 피해자들에게 사과하는 것"이라며 "그게 선행되지 않으면 총장의 발언은 '심각한 유체이탈 화법'에 불과한 것으로 받아들여진다"고 지적했다. : 60.0%
박경미 의원은 "교원소청위는 억울한 교원의 마지막 구제 수단으로, 그 권위와 신뢰를 인정해야 하는 기관"이라며 "교단에 서서는 안 될 비위를 저지른 일부 교원들이 제도를 악용할 수 없게 객관적이고 엄격한 기준을 마련해야 한다"고 강조했다. : 55.0%
이 관계자는 "대통령은 참여정부나 국민의 정부 같은 네이밍 지칭 계획이 없으며 실용적으로 사용하겠다고 한다"며 "오늘 문재인 정부라고 한 것도 기념사에서 '정부는'이라고 하는 것보다 '문재인 정부는'이라고 하는 게 기념식을 보고 있는 국민에게 자신의 의지를 잘 보여줄 수 있다고 해서 그렇게 부른 것"이라고 말했다. : 55.0%
또 북한 언론들은 "6·15 공동선언의 채택은 장장 반세기 이상 지속되어온 불신과 대결의 역사에 종지부를 찍고, 민족의 화해와 단합, 통일과 평화번영의 새 시대를 열어놓은 특대사변이었다"고 평가하고 "거기에는 북남관계를 개선하고 조국통일을 이룩하는 데서 일관하게 견지해야 할 원칙과 모든 분야에서 협력과 교류를 전면적으로 확대·발전시켜나가기 위한 가장 합리적이고 건설적인 방도들이 다 들어있다. : 55.0%
그는 "발주의 부분일 사업을 보고 서울에도 시민들이 걷을 수 있는 장소를 많이 만들었다"며 "시민의 힘과 자력은 결국 다양한 문화를 만들고 공유하는 데서 나온다"고 말했다. : 55.0%
서 시장은 자신의 저서 '일하는 사람이 미래를 말한다'에서 "캠퍼스에서 간간이 마주쳐거나 멀리서 보게될 때도 체벌 있었다. : 55.0%
권 청장은 "잘못된 언행으로 마음의 상처를 입은 당사자와 가족, 부산경찰 동료 여러분들께 다시 한 번 진심으로 사과하며, 더불어 평소 부산경찰을 사랑하고 아껴주시는 부산시민에게 진심으로 사죄의 말씀을 드린다"고 밝혔다 권 청장은 또 "이번 일로 경찰청장으로부터 엄중한 경고를 받았고, 이를 매우 무겁게 받아들이며 깊이 반성하고 있다"며 "앞으로 자숙의 시간을 가지면서 잘못된 습관을 바로 잡고, 조직 운영 방식을 바꾸어 나가겠다"고 강조했다. : 55.0%
elsaped Time : 207.5761251449585s
```

- oktResult



구현 방법

- 형태소 분석 방식에 의한 유사도 측정

  아래와 같은 순서로 형태소 분석에 의한 유사도 측정 알고리즘을 구현하였다.

  1. 텍스트 파일을 한 문장씩 리스트에 저장해 반환한다.
  2. 한 문장씩 형태소를 분석한다.
  3. collections.Counter를 이용하여 각 형태소에 대한 빈도를 측정한다.
  4. 형태소 갯수가 적은 문장을 기준으로 공통 형태소 갯수를 측정한다.
  5. 입력받은 문장의 형태소 갯수로 공통 형태소 갯수를 나누고 100을 곱하여 공통 형태소 유사도를 계산하여 sentenceDict에 저장한다.
  6. 유사도가 높은 순서대로 정렬 한 후 사용자 입력으로 받은 n 만큼 문장을 출력한다.

- 1을 구현한 openFile

```
def openFile(filePath):
  lines = None
  with open(filePath) as f:
    try:
      lines = f.read().splitlines()
    except:
      f.close()
  return lines
```

- 2를 구현한 코드

```
for line in sentenceList:
  if line == '':
    continue
  sentencePos = han.pos(line)
```

- 3을 구현한 코드

```
sentencePosCount = Counter(sentencePos)
```

- 4를 구현한 코드

```
common = 0
for morpheme in inputPosCount:
  if morpheme in sentencePosCount:
    common += min(inputPosCount[morpheme],sentencePosCount[morpheme])
```

- 5를 구현한 코드

```
similarity = 100 * common / inputLen
sentenceDict[line] = similarity
```

- 6을 구현한 코드

```
hanResult = konlpyHannanum(sen, sentenceList)
hanResult = sorted(hanResult.items(), key=lambda x: x[1], reverse=True)

for i in range(Num):
  print(f'{hanResult[i][0]} : {hanResult[i][1]}%')
```

문제점

- 공통사항 : sample.txt (15000줄)
- 실행시간 : Komoran > Okt > Hannanum

1. Komoran KoNLPy 패키지의 경우 다른 형태소분석기보다 실행시간이 적게 걸렸으나 다르게 같은 입력 문장이더라도 105%의 유사도를 반환하였다.

```
oktResult = [('반성은', 'Noun'), ('자신', 'Noun'), ('이', 'Josa'), ('저지른', 'Verb'), ('죄', 'Noun'), ('의', 'Josa'), ('대가', 'Noun'
komoResult = [('반성', 'NNG'), ('은', 'JX'), ('자신', 'NNG'), ('이', 'JKS'), ('저지르', 'VV'), ('ㄴ', 'ETM'), ('죄', 'NNG'), ('의', 'J
hanResult = [('반성', 'N'), ('은', 'J'), ('자신', 'N'), ('이', 'J'), ('저지르', 'P'), ('ㄴ', 'E'), ('죄', 'N'), ('의', 'J'), ('대가', '
```

2. 느린 실행시간