

Text Analysis 실습

비정형 데이터 분석을 통한 효율적인 의사결정

박진수 교수

**Big Data Institute,
Seoul National University**

실습 1 – 두 개의 영화 리뷰 텍스트 간 유사성 계산하기



실습 1-1. 두 영화 리뷰 간 유사성 계산하기 (1)

- 실습 2-2-4의 결과로 저장된 영화 리뷰 중 두 개를 임의로 선택해 두 영화 리뷰 간의 유사성을 계산해 출력해 본다

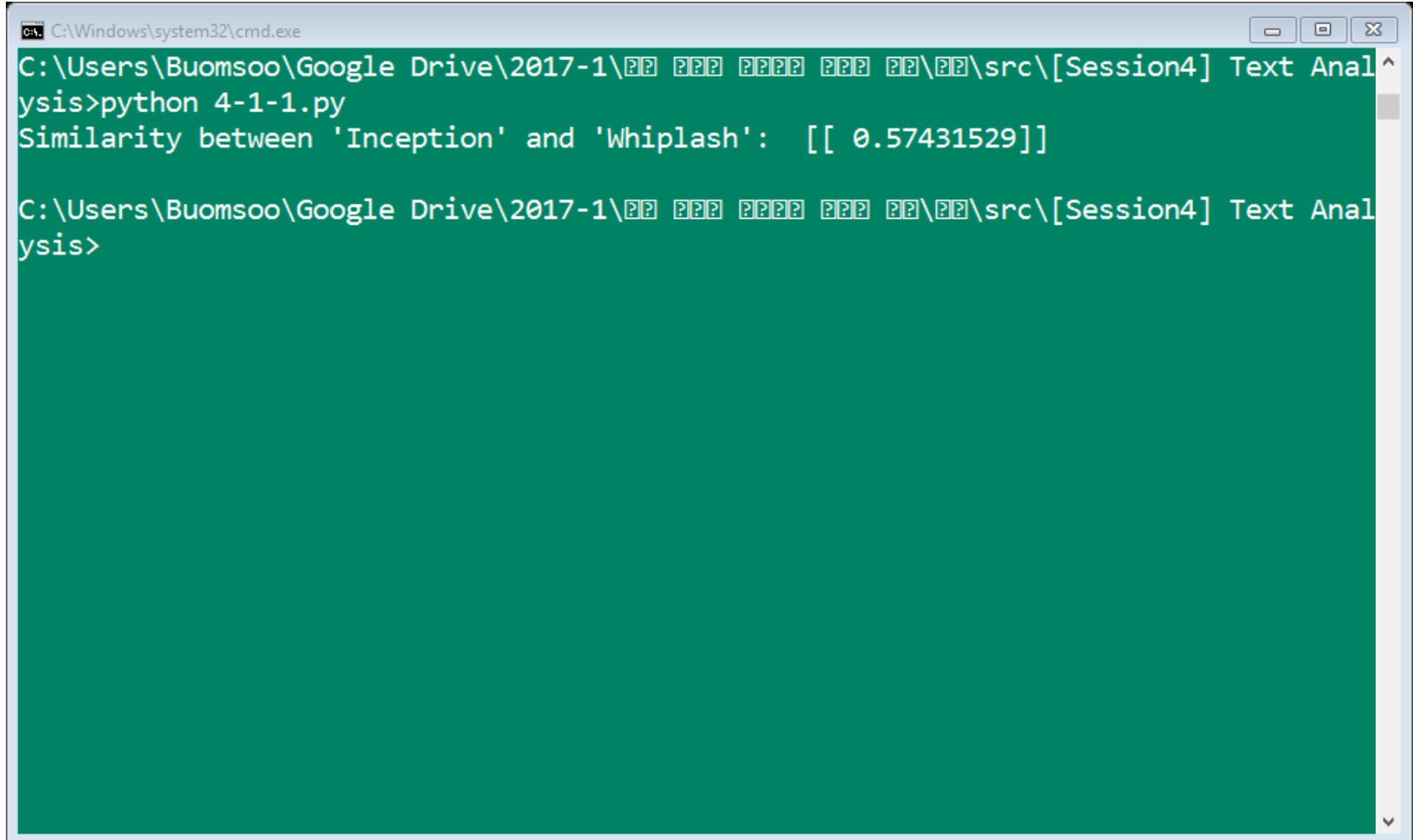
- TIPS

- 두 개의 영화 리뷰 텍스트만으로 corpus를 만들어 유사성을 계산한다

- sklearn 패키지를 활용한다

실습 1-1. 두 영화 리뷰 간 유사성 계산하기 (1)

- 출력 결과(인셉션과 위플래시 간의 유사성)



```
C:\Windows\system32\cmd.exe
C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>python 4-1-1.py
Similarity between 'Inception' and 'Whiplash': [[ 0.57431529]]

C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>
```

실습 1-2. 두 영화 리뷰 간 유사성 계산하기 (2)

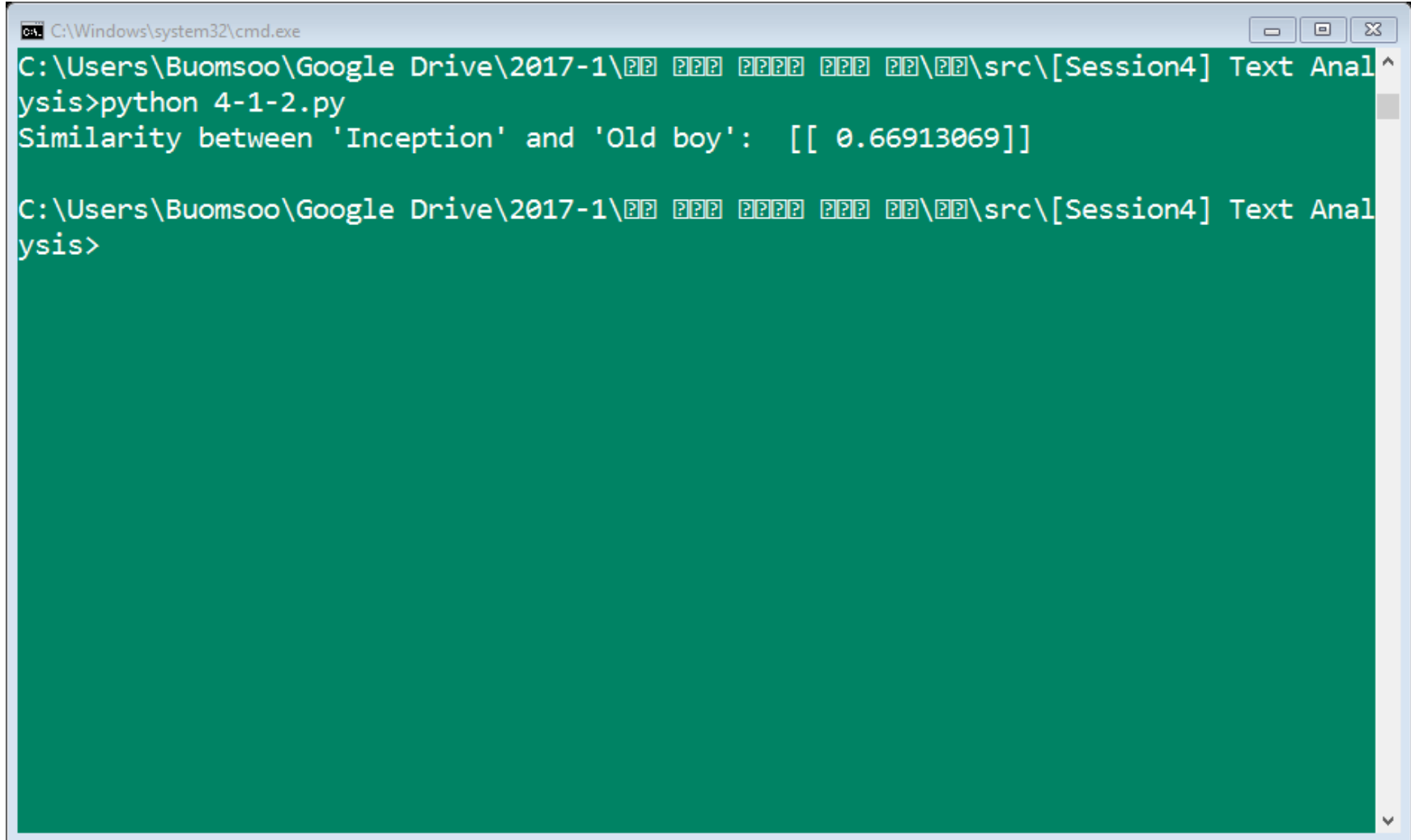
- 실습 2-2-4의 결과로 저장된 영화 리뷰 중 다른 두 개를 임의로 선택해 두 영화 리뷰 간의 유사성을 계산해 출력해 본다

- TIPS

- 두 개의 영화 리뷰 텍스트만으로 corpus를 만들어 유사성을 계산한다
- sklearn 패키지를 활용한다

실습 1-2. 두 영화 리뷰 간 유사성 계산하기 (2)

- 출력 결과(인셉션과 올드보이 간의 유사성)



```
C:\Windows\system32\cmd.exe
C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Analysis>python 4-1-2.py
Similarity between 'Inception' and 'Old boy': [[ 0.66913069]]

C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Analysis>
```

실습 1-3. 두 영화 리뷰 간 유사성 계산하기 (3)

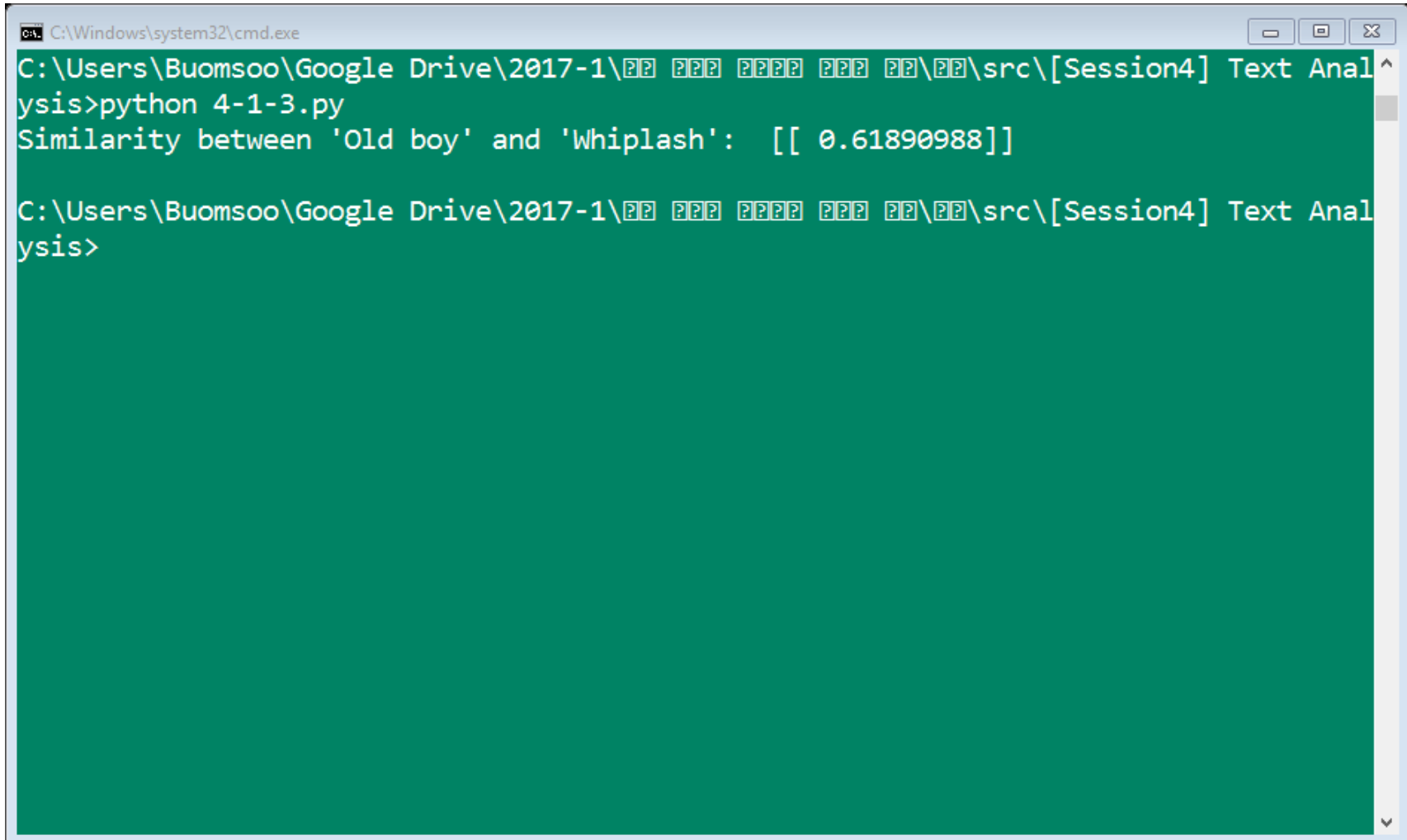
- 실습 2-2-4의 결과로 저장된 영화 리뷰 중 다른 두 개를 임의로 선택해 두 영화 리뷰 간의 유사성을 계산해 출력해 본다

- TIPS

- 두 개의 영화 리뷰 텍스트만으로 corpus를 만들어 유사성을 계산한다
- sklearn 패키지를 활용한다

실습 1-3. 두 영화 리뷰 간 유사성 계산하기 (3)

- 출력 결과(인셉션과 올드보이 간의 유사성)



```
C:\Windows\system32\cmd.exe
C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>python 4-1-3.py
Similarity between 'Old boy' and 'Whiplash': [[ 0.61890988]]

C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>
```


실습 2 – 여러 영화 리뷰 텍스트 간 유사성 계산하기



실습 2-1. 세 영화 리뷰 간 유사성 계산하기

- 실습 2-2-4의 결과로 저장된 영화 리뷰 중 세 개의 영화 리뷰를 선택해 영화 리뷰 간 유사성을 계산해 본다

- TIPS

- 세 개의 영화 리뷰 텍스트로 corpus를 만들어 유사성을 계산한다

- sklearn 패키지를 활용한다

실습 2-1. 세 영화 리뷰 간 유사성 계산하기

- 출력 결과

```
C:\Windows\system32\cmd.exe
C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>python 4-2-1.py
Similarity between 'Inception' and 'Whiplash': [[ 0.51751967]]
Similarity between 'Inception' and 'Old boy': [[ 0.63306892]]
Similarity between 'Whiplash' and 'Old boy': [[ 0.56556735]]

C:\Users\Buomsoo\Google Drive\2017-1\?? ???? ???? ???? ??\??\src\[Session4] Text Anal
ysis>
```

실습 2-1. 세 영화 리뷰 간 유사성 계산하기

· 참고

- 영화 리뷰를 두 개씩 pairing 했을 때

	Inception	Whiplash	Old Boy
Inception	-	0.5743	0.6691
Whiplash	0.5743	-	0.6189
Old Boy	0.6691	0.6189	-

- 영화 리뷰를 세 개씩 묶어서 계산했을 때

	Inception	Whiplash	Old Boy
Inception	-	0.5175	0.6330
Whiplash	0.5175	-	0.5655
Old Boy	0.6330	0.5655	-

- 결과가 미묘하게 달라지는 이유는?