

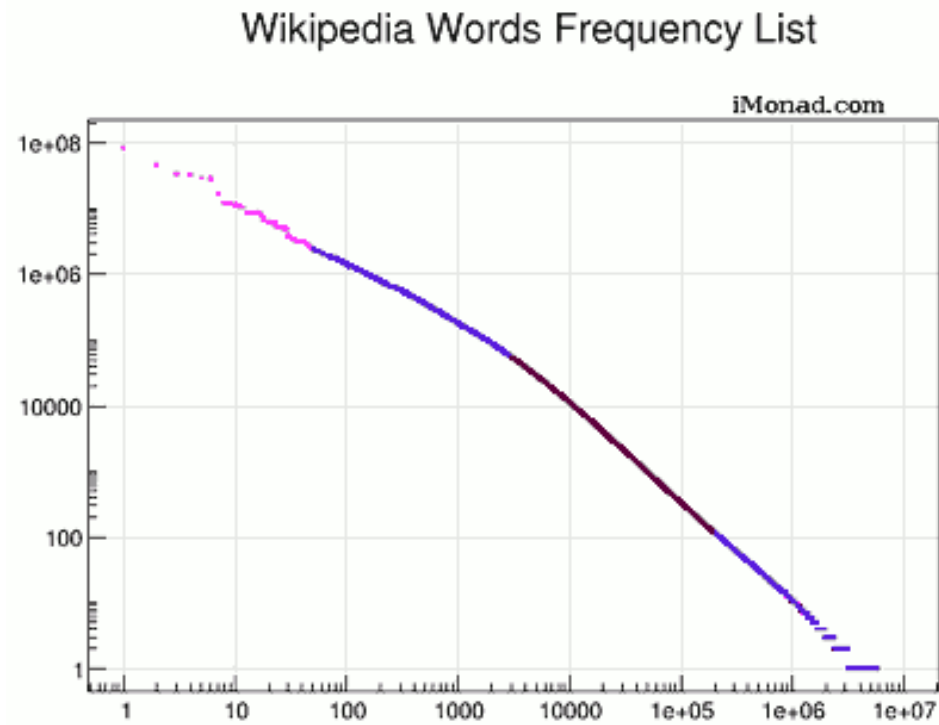
# Text Exploration & Visualization 실습

비정형 데이터 분석을 통한 효율적인 의사결정

박진수 교수

Big Data Institute,  
Seoul National University

# 실습 1 – 가장 많이 등장하는 단어 추출하기



# 실습 1-1. 리뷰에서 많이 등장하는 명사 추출하기

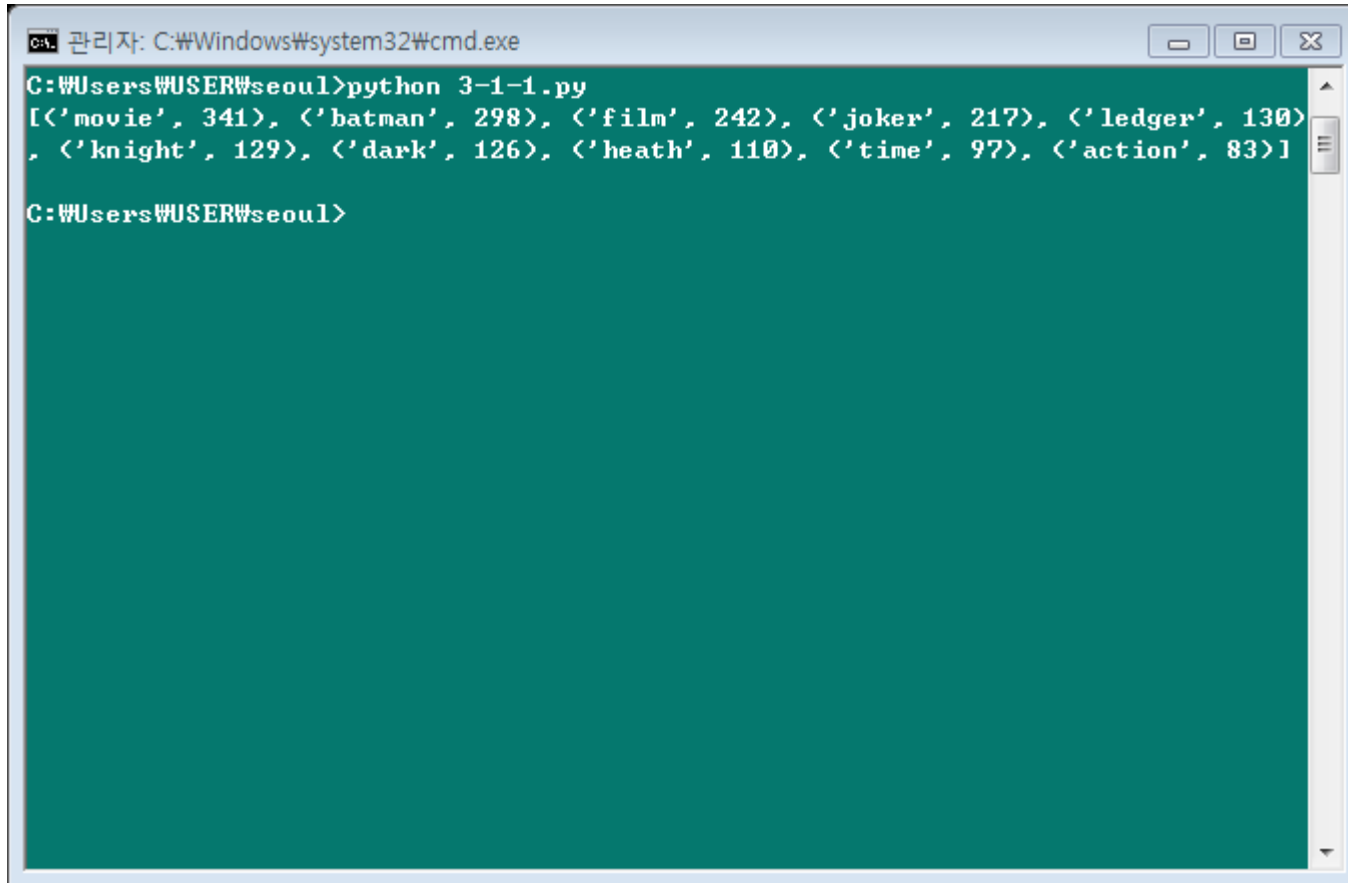
· 실습 1-3-4의 결과(영화 다크 나이트 리뷰 데이터) 파일을 불러와 가장 많이 사용된 명사 10개를 추출해 출력해 본다

## · TIPS

- 텍스트를 lower() 함수를 사용해 소문자로 변환 후 비교한다
- 파이썬의 collections 패키지의 Counter 함수를 사용해 많이 사용된 명사를 추출한다
- POS tag 가 'NN', 'NNS', 'NNP', 'NNPS'인 것을 추출한다

# 실습 1-1. 리뷰에서 많이 등장하는 명사 추출하기

- 출력 결과



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-1-1.py
[<'movie', 341>, <'batman', 298>, <'film', 242>, <'joker', 217>, <'ledger', 130>, <'knight', 129>, <'dark', 126>, <'heath', 110>, <'time', 97>, <'action', 83>]
```

## 실습 1-1. 리뷰에서 많이 등장하는 명사 추출하기

- 영화 '다크 나이트' 리뷰에서 자주 등장한 명사들

단어	빈도수	단어	빈도수
movie	341	knight	129
batman	298	dark	126
film	242	heath	110
joker	217	time	97
ledger	130	action	83

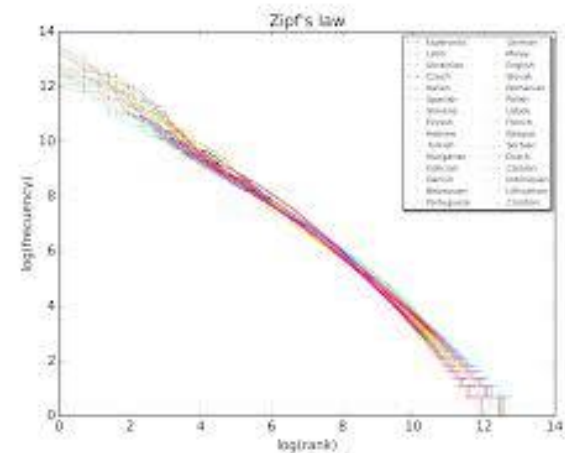
- 대부분 영화 및 영화 내용과 관련된 단어임을 알 수 있다

# 실습 1-1. 리뷰에서 많이 등장하는 명사 추출하기

· 참고: 지프의 법칙(Zipf's Law) [source: [https://en.wikipedia.org/wiki/Zipf%27s\\_law](https://en.wikipedia.org/wiki/Zipf%27s_law)]

- 20/80의 법칙이라고도 하며 우리가 사용하는 일상어에서 소수의 단어들이 빈도수의 대부분을 차지한다는 내용의 법칙이다
- 예를 들어, 미국 표준 영어에서 “the”가 가장 많이 쓰이며 전체 문서에서 약 7%의 빈도를 보여준다고 한다. 다음으로 많이 사용되는 단어는 “of”와 “and”로 각각 3.5%와 3%의 빈도를 보여준다

· 즉, 지프의 법칙에 따르면 텍스트에서 자주 사용되는 상위 단어 몇 개(즉, 키워드)만 보더라도 텍스트의 내용을 대략적으로 파악할 수 있다



## 실습 1-2. 리뷰에서 많이 등장하는 형용사 추출하기

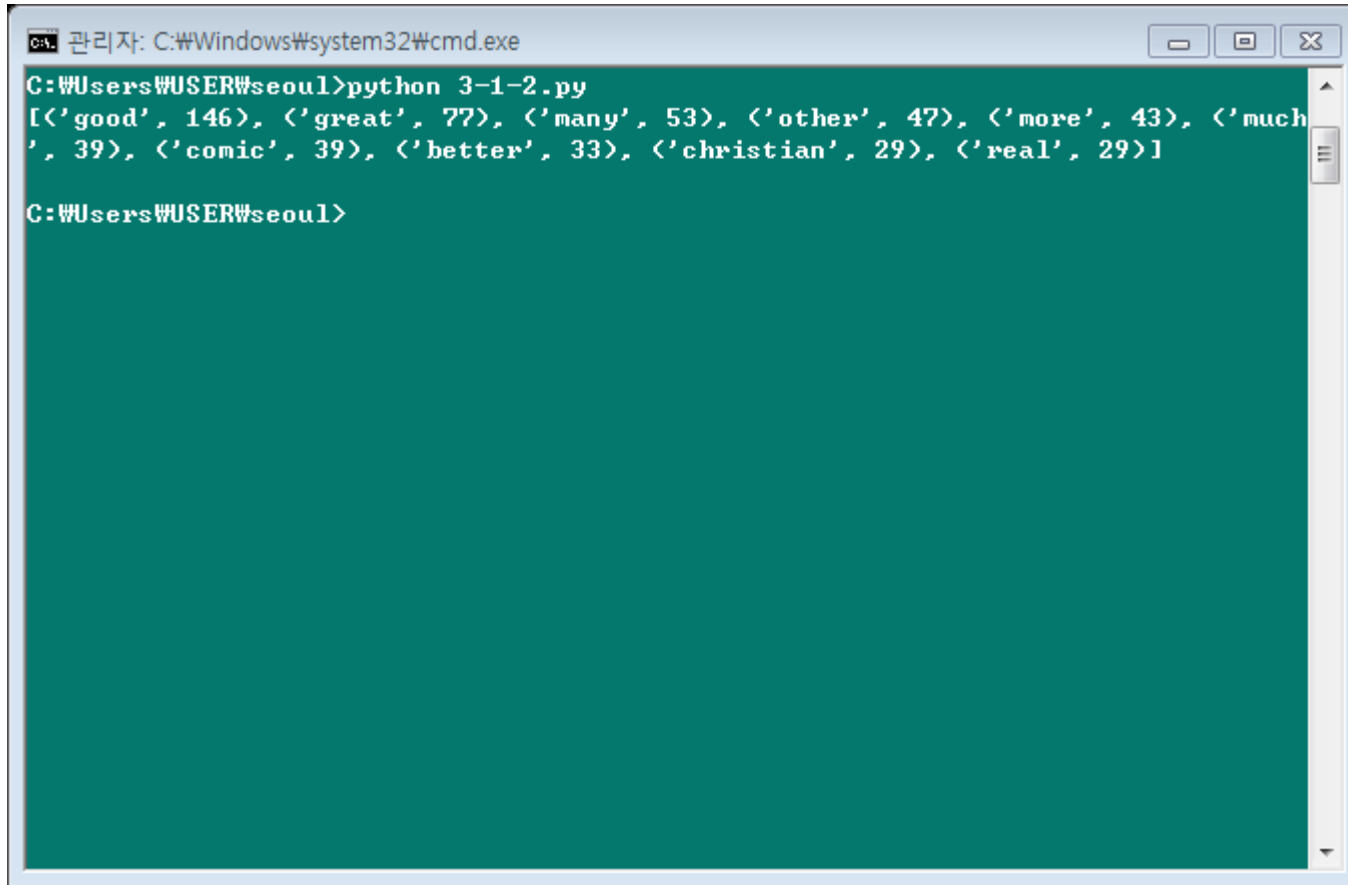
- 실습 1-3-4의 결과(영화 다크 나이트 리뷰 데이터) 파일을 불러와 가장 많이 사용된 형용사 10개를 추출해 출력해 본다

### · TIPS

- 텍스트를 lower() 함수를 사용해 소문자로 변환 후 비교한다
- 파이썬의 collections 패키지의 Counter 함수를 사용해 많이 사용된 명사를 추출한다
- POS tag 가 'JJ', 'JJR', 'JJS' 인 것을 추출한다

## 실습 1-2. 리뷰에서 많이 등장하는 형용사 추출하기

- 출력 결과



A screenshot of a Windows command prompt window. The title bar reads "관리자: C:\Windows\system32\cmd.exe". The command prompt shows the execution of a Python script: `C:\Users\WUSER\seoul>python 3-1-2.py`. The output is a list of tuples: `[('good', 146), ('great', 77), ('many', 53), ('other', 47), ('more', 43), ('much', 39), ('comic', 39), ('better', 33), ('christian', 29), ('real', 29)]`. The prompt then returns to `C:\Users\WUSER\seoul>`.

```
관리자: C:\Windows\system32\cmd.exe
C:\Users\WUSER\seoul>python 3-1-2.py
[('good', 146), ('great', 77), ('many', 53), ('other', 47), ('more', 43), ('much', 39), ('comic', 39), ('better', 33), ('christian', 29), ('real', 29)]
C:\Users\WUSER\seoul>
```



## 실습 1-2. 리뷰에서 많이 등장하는 형용사 추출하기

- 영화 '다크 나이트' 리뷰에서 자주 등장한 형용사들

단어	빈도수	단어	빈도수
good	146	much	39
great	77	comic	39
many	53	better	33
other	47	christian	29
more	43	real	29

- 형용사는 주로 영화에 대한 감정, 혹은 평가와 관련된 단어들이 많은 것을 알 수 있다
- 29번 등장한 '**Christian**': 주연 배우인 크리스천 배일(Christian Bale)의 이름이라 자주 등장한 것인데 컴퓨터는 이를 형용사로 인식함 => 컴퓨터 인식의 한계이므로 사람의 개입이 필요한 부분

# 실습 1-3. 리뷰에서 많이 등장하는 단어 추출하기

---

- 실습 1-3-5에서 수집했던 다른 영화들의 리뷰에서 자주 등장하는 단어를 추출해 본다

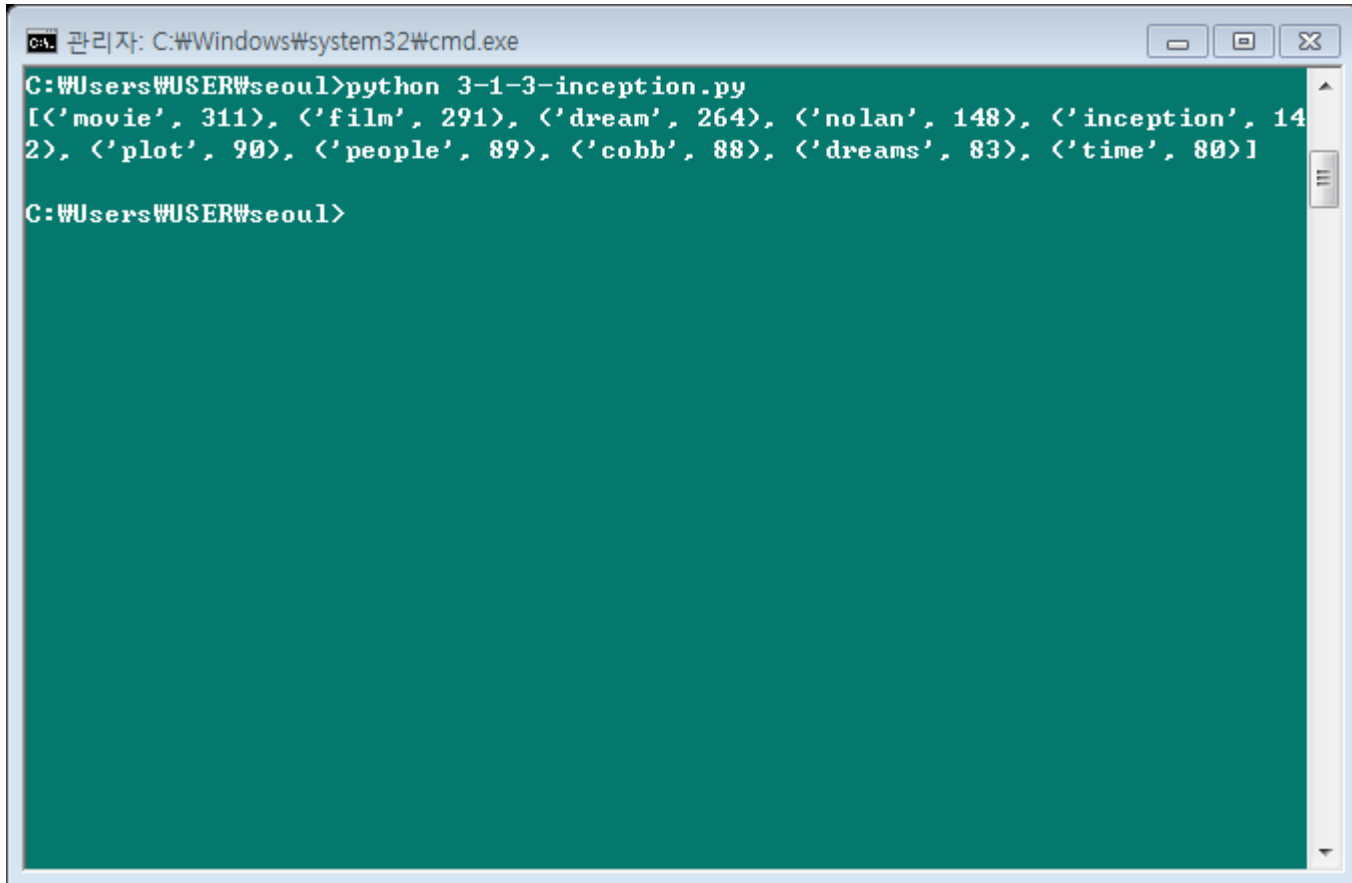
- 자신이 관심 있는 품사의 단어들을 자유롭게 추출해 출력해 본다

- TIPS

- 실습 2-1-2의 영어 문장 POS 태그를 참고한다

## 실습 1-3. 리뷰에서 많이 등장하는 단어 추출하기

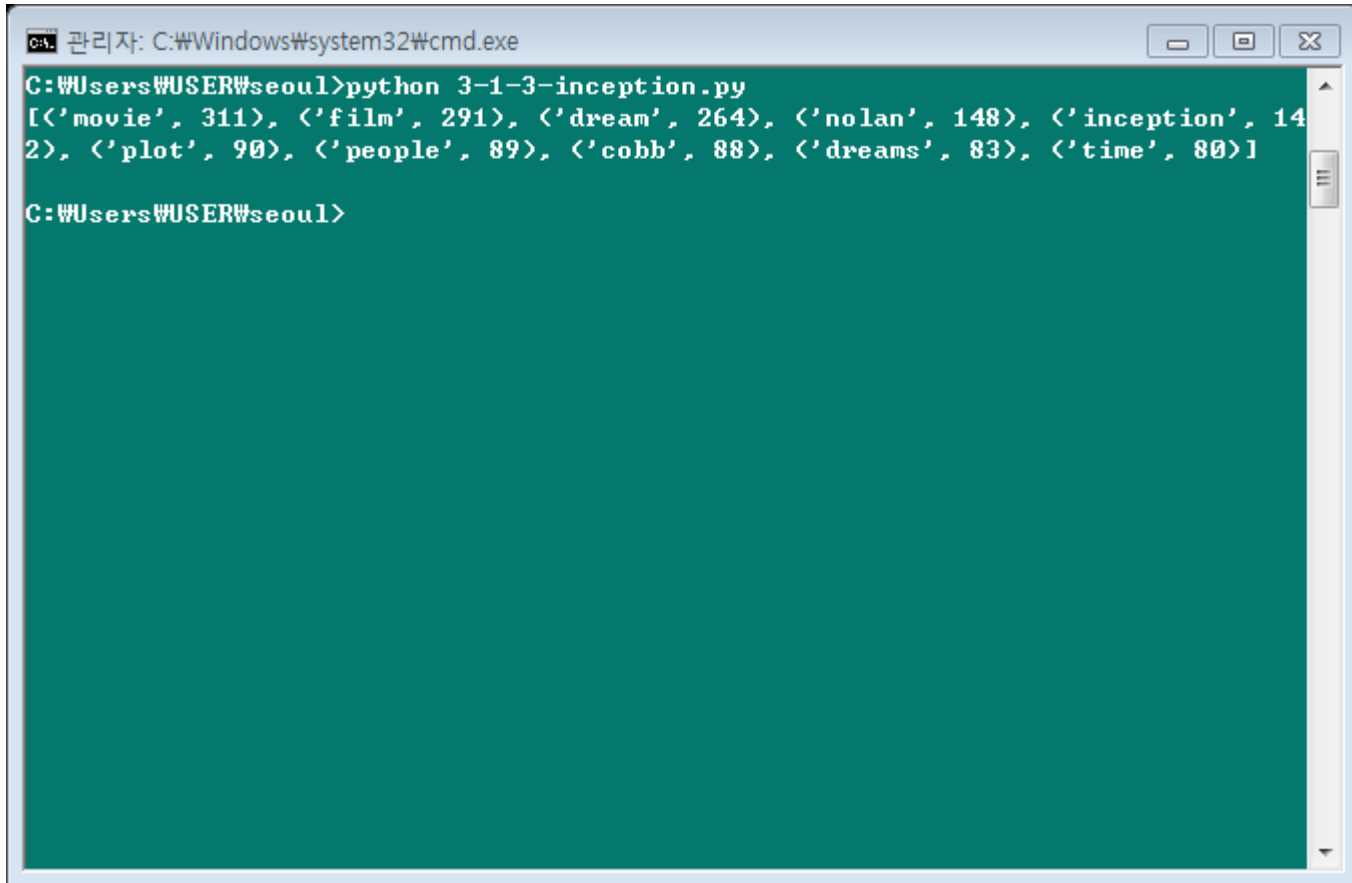
- 출력 결과: 영화 '인셉션' 리뷰에서 자주 등장하는 명사들



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-1-3-inception.py
[('movie', 311), ('film', 291), ('dream', 264), ('nolan', 148), ('inception', 142), ('plot', 90), ('people', 89), ('cobb', 88), ('dreams', 83), ('time', 80)]
C:\Users\USER\seoul>
```

## 실습 1-3. 리뷰에서 많이 등장하는 단어 추출하기

- 출력 결과: 영화 '인셉션' 리뷰에서 자주 등장하는 명사들

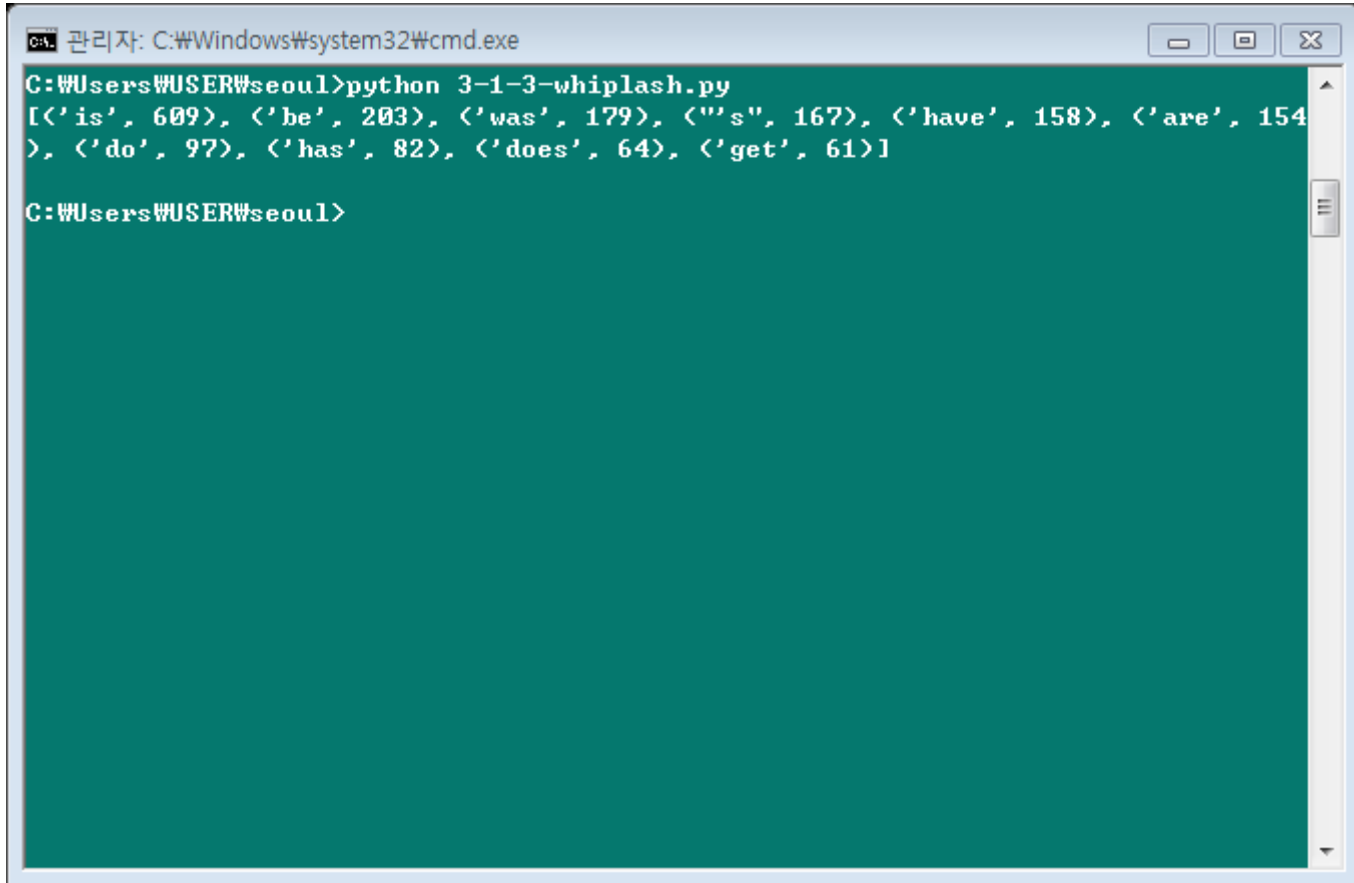


```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-1-3-inception.py
[('movie', 311), ('film', 291), ('dream', 264), ('nolan', 148), ('inception', 142), ('plot', 90), ('people', 89), ('cobb', 88), ('dreams', 83), ('time', 80)]

C:\Users\USER\seoul>
```

## 실습 1-3. 리뷰에서 많이 등장하는 단어 추출하기

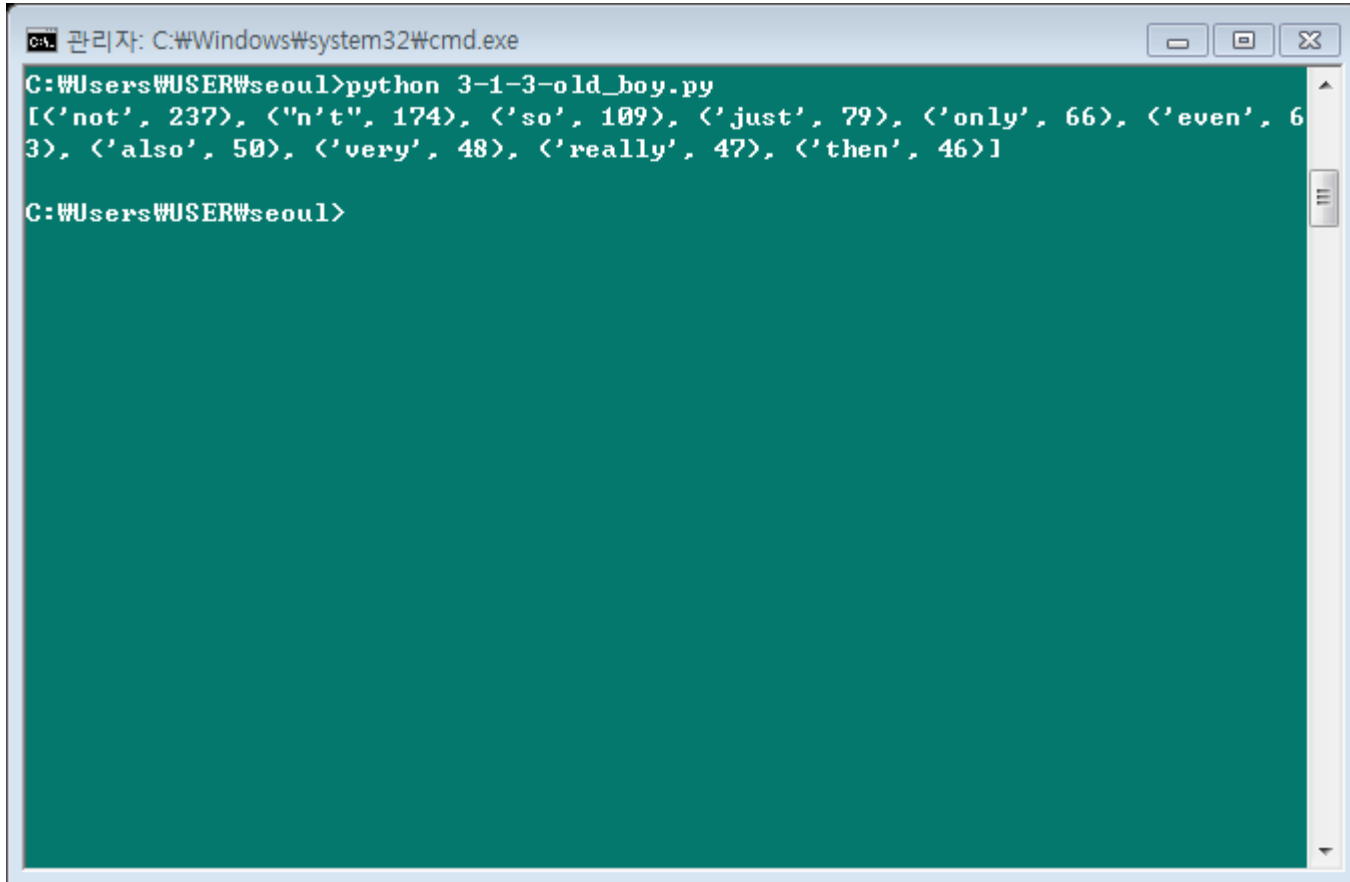
- 출력 결과: 영화 '위플래시' 리뷰에서 자주 등장하는 동사들



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-1-3-whiplash.py
[('is', 609), ('be', 203), ('was', 179), ('s', 167), ('have', 158), ('are', 154), ('do', 97), ('has', 82), ('does', 64), ('get', 61)]
C:\Users\USER\seoul>
```

## 실습 1-3. 리뷰에서 많이 등장하는 단어 추출하기






- 출력 결과: 영화 '올드보이' 리뷰에서 자주 등장하는 부사들




```
관리자: C:\Windows\system32\cmd.exe
C:\Windows\system32\cmd.exe>python 3-1-3-old_boy.py
[('not', 237), ('n't', 174), ('so', 109), ('just', 79), ('only', 66), ('even', 63), ('also', 50), ('very', 48), ('really', 47), ('then', 46)]

C:\Windows\system32\cmd.exe>
```

# 실습 2 – IMDb 리뷰 데이터 탐색하기

AllIMDbPro [Help](#)     
[Movies, TV & Showtimes](#) [Celebs, Events & Photos](#) [News & Community](#) [Watchlist](#) [Sign in with Facebook](#) [Other Sign in options](#)

IMDb > [Whiplash \(2014\)](#) > [Reviews & Ratings](#) - IMDb



Reviews & Ratings for

**Whiplash** [More at IMDbPro »](#)

[Write review](#)

Filter: Best Hide Spoilers: ☐

Page 1 of 83: [\[1\]](#) [\[2\]](#) [\[3\]](#) [\[4\]](#) [\[5\]](#) [\[6\]](#) [\[7\]](#) [\[8\]](#) [\[9\]](#) [\[10\]](#) [\[11\]](#) ▶

Index 821 reviews in total

[Own the rights?](#)  
[Buy it at Amazon](#)  
[More at IMDb Pro](#)  
[Add to Watchlist](#)  
[Update Data](#)


**Quicklinks**  
[reviews](#)

**Top Links**

- trailers and videos
- full cast and crew
- trivia
- official sites
- memorable quotes

**Overview**

489 out of 605 people found the following review useful:



**An incredibly powerful film!**

★★★★★

**Author:** [Gbert254](#) from Utah, United States  
28 January 2014

<http://switchingreels.com/2014/01/28/sundance-review-whiplash/>

Ever had a dream of being a great football player? A great dancer? A great singer? A great musician? Our protagonist has a dream of being a great drummer, a drummer that will be remembered forever. Maybe you are still fighting for your dream. Maybe you have given up on greatness. Greatness doesn't come easily, you need to practice at it. Andrew practices until his hands bleed.

Andrew (Miles Teller) is 19-year old student at a music conservatory in Manhattan. Terrence Fletcher (J.K. Simmons) is a teacher at the conservatory with a ruthlessly brutal teaching style. After picking Andrew to play in the school band, he pushes Andrew to his limits in order to realize his full potential, at the risk of his humanity.

I had a billiards teacher at one point in my life, who was close to becoming a pro in his craft but a grease fire accident changed all that. His

## 실습 2-1. 전체 리뷰의 토큰 개수 출력하기

- 전체 토큰의 개수와 중복되지 않는(unique) 토큰의 개수를 출력해 보자

The image shows a Notepad++ window with a file named 'result-1-3-4.txt'. The text content is a repetitive paragraph about Christopher Nolan's 'The Dark Knight' movie. The paragraph is repeated multiple times, with some lines appearing to be pasted multiple times. The status bar at the bottom indicates the file is a 'Normal text file' with a length of 242,563 and 101 lines. The cursor is at line 1, column 1. The encoding is UTF-8 and the line endings are Windows (CR LF).

D:\Google Drive\W07-1#도식 데이터 사이언스 연구소 교육원실습\wrcw[Session1] Web Crawling#result-1-3-4.txt - Notepad++ [Administrator]

파일(F) 편집(E) 찾기(S) 보기(V) 인코딩(N) 언어(L) 설정(T) Tools 매크로 실행 플러그인 장 관련 ?

result-1-3-4.txt

1 Christopher Nolan's second bundle of joy "The Dark Knight" EXCEEDED all of my expectations!!! With the success of 2005's  
2 I had the honor of watching TDK during a screening and was completely blown away! This isn't just the best Batman movie e  
3 I used to leave a theatre after seeing a highly anticipated movie, specifically a sequel, and be so revved up about what  
4 I must say I was excited for this movie since the instant BATMAN BEGINS appeared on screen at the end of the first film.  
5 I thought Batman Begins was a very well conceived and put together movie. We finally get Batman as a fully rendered cha  
6 Batman has always been my favourite superhero ever since the first time I heard about him because he is human with no po  
7 (Synopsis) Bruce Wayne/Batman (Christian Bale) continues to eliminate crime in Gotham City with the help of Lt. Jim Gordo  
8 I got to see The Dark Knight on Wednesday night, the reason though why I'm writing this movie comment this late is becaus  
9 I think the big question...or the question everyone actually cares about is this: Is Christopher Nolan's follow up to the  
10 I have just come out of the "Dark Curtain" Screening for The Dark Knight in Adelaide, and I'm blown away.Nolan's directin  
11 I had the honor of watching TDK during a screening and was completely blown away! This isn't just the best Batman movie e  
12 I used to leave a theatre after seeing a highly anticipated movie, specifically a sequel, and be so revved up about what  
13 I must say I was excited for this movie since the instant BATMAN BEGINS appeared on screen at the end of the first film.  
14 I thought Batman Begins was a very well conceived and put together movie. We finally get Batman as a fully rendered cha  
15 Batman has always been my favourite superhero ever since the first time I heard about him because he is human with no po  
16 (Synopsis) Bruce Wayne/Batman (Christian Bale) continues to eliminate crime in Gotham City with the help of Lt. Jim Gordo  
17 I got to see The Dark Knight on Wednesday night, the reason though why I'm writing this movie comment this late is becaus  
18 I think the big question...or the question everyone actually cares about is this: Is Christopher Nolan's follow up to the  
19 I have just come out of the "Dark Curtain" Screening for The Dark Knight in Adelaide, and I'm blown away.Nolan's directin  
20 I couldn't believe "The Dark knight" could live up to the hype. That's perhaps the biggest surprise. The secret, I believ  
21 I used to leave a theatre after seeing a highly anticipated movie, specifically a sequel, and be so revved up about what  
22 I must say I was excited for this movie since the instant BATMAN BEGINS appeared on screen at the end of the first film.  
23 I thought Batman Begins was a very well conceived and put together movie. We finally get Batman as a fully rendered cha  
24 Batman has always been my favourite superhero ever since the first time I heard about him because he is human with no po  
25 (Synopsis) Bruce Wayne/Batman (Christian Bale) continues to eliminate crime in Gotham City with the help of Lt. Jim Gordo  
26 I got to see The Dark Knight on Wednesday night, the reason though why I'm writing this movie comment this late is becaus  
27 I think the big question...or the question everyone actually cares about is this: Is Christopher Nolan's follow up to the  
28 I have just come out of the "Dark Curtain" Screening for The Dark Knight in Adelaide, and I'm blown away.Nolan's directin  
29 I couldn't believe "The Dark knight" could live up to the hype. That's perhaps the biggest surprise. The secret, I believ  
30 Dark, yes, complex, ambitious. Christopher Nolan and his co-writer Jonathan Nolan deserve a standing ovation. I don't usu  
31 I must say I was excited for this movie since the instant BATMAN BEGINS appeared on screen at the end of the first film.  
32 I thought Batman Begins was a very well conceived and put together movie. We finally get Batman as a fully rendered cha  
33 Batman has always been my favourite superhero ever since the first time I heard about him because he is human with no po  
34 (Synopsis) Bruce Wayne/Batman (Christian Bale) continues to eliminate crime in Gotham City with the help of Lt. Jim Gordo  
35 I got to see The Dark Knight on Wednesday night, the reason though why I'm writing this movie comment this late is becaus  
36 I think the big question...or the question everyone actually cares about is this: Is Christopher Nolan's follow up to the  
37 I have just come out of the "Dark Curtain" Screening for The Dark Knight in Adelaide, and I'm blown away.Nolan's directin  
38 I couldn't believe "The Dark knight" could live up to the hype. That's perhaps the biggest surprise. The secret, I believ  
39 Dark, yes, complex, ambitious. Christopher Nolan and his co-writer Jonathan Nolan deserve a standing ovation. I don't usu  
40 \*\* PLOT SPOILERS EVERYWHERE \*\* Christopher Nolan's "The Dark Knight" is one of the better films in the history of its gen  
41 I thought Batman Begins was a very well conceived and put together movie. We finally get Batman as a fully rendered cha  
42 Batman has always been my favourite superhero ever since the first time I heard about him because he is human with no po  
43 (Synopsis) Bruce Wayne/Batman (Christian Bale) continues to eliminate crime in Gotham City with the help of Lt. Jim Gordo

Normal text file length: 242,563 lines: 101 Ln:1 Col:1 Sel:0|0 Windows (CR LF) UTF-8 INS



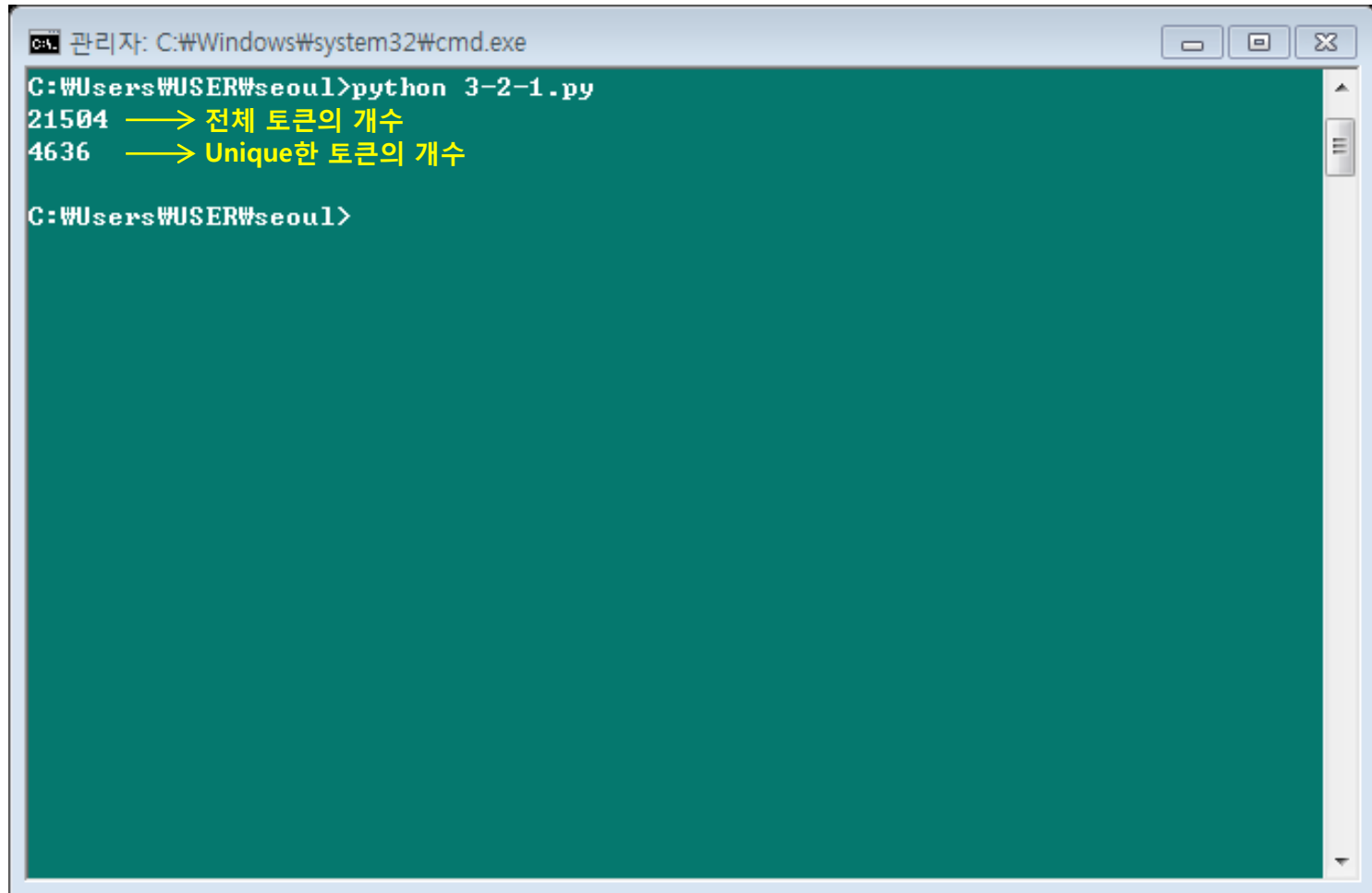
## 실습 2-1. 전체 리뷰의 토큰 개수 출력하기

### · TIPS

- 리뷰를 토큰화할 때 `lower` 함수를 사용해 소문자로 변환한 뒤 토큰화한다
- Nltk의 `Text` 함수를 사용한다
  - `nltk.Text()`는 전체 문서(corpus)를 쉽고 간단하게 탐색할 수 있도록 도와준다
- Unique한 토큰의 개수를 출력하기 위해서는 리스트를 셋(set)으로 변환한다
- 분석의 효율성을 위해 `Stopwords`를 제거한다

## 실습 2-1. 전체 리뷰의 토큰 개수 출력하기

- 출력 결과



```
C:\> 관리자: C:\Windows\system32\cmd.exe  
C:\Users\WUSER\seoul>python 3-2-1.py  
21504  —> 전체 토큰의 개수  
4636   —> Unique한 토큰의 개수  
C:\Users\WUSER\seoul>
```

The screenshot shows a Windows command prompt window titled "관리자: C:\Windows\system32\cmd.exe". The prompt is "C:\Users\WUSER\seoul>". The user has entered the command "python 3-2-1.py". The output of the script is displayed on two lines: "21504 —> 전체 토큰의 개수" and "4636 —> Unique한 토큰의 개수". The prompt "C:\Users\WUSER\seoul>" is shown again on the next line.

## 실습 2-2. 토큰의 등장 횟수 시각화하기

- 실습 1-3-4에서 수집하였던 영화 ‘다크 나이트(The Dark Knight)’ 리뷰를 토큰화한 뒤 가장 많이 등장하는 토큰 50개의 등장 횟수를 그래프로 시각화해본다

The image shows a Notepad++ window editing a file named 'result-1-3-4.txt'. The text in the editor is a repetitive paragraph about Christopher Nolan's 'The Dark Knight' movie, with line numbers 1 through 40 visible on the left margin. The text describes the reviewer's excitement and praise for the movie, mentioning Bruce Wayne/Batman (Christian Bale) and Jonathan Nolan. The bottom status bar shows 'Normal text file', 'length: 242,562 lines: 101', 'Ln:1 Col:1 Sel:0|0', 'Windows (CR LF) UTF-8', and 'INS'.

## 실습 2-2. 토큰의 등장 횟수 시각화하기

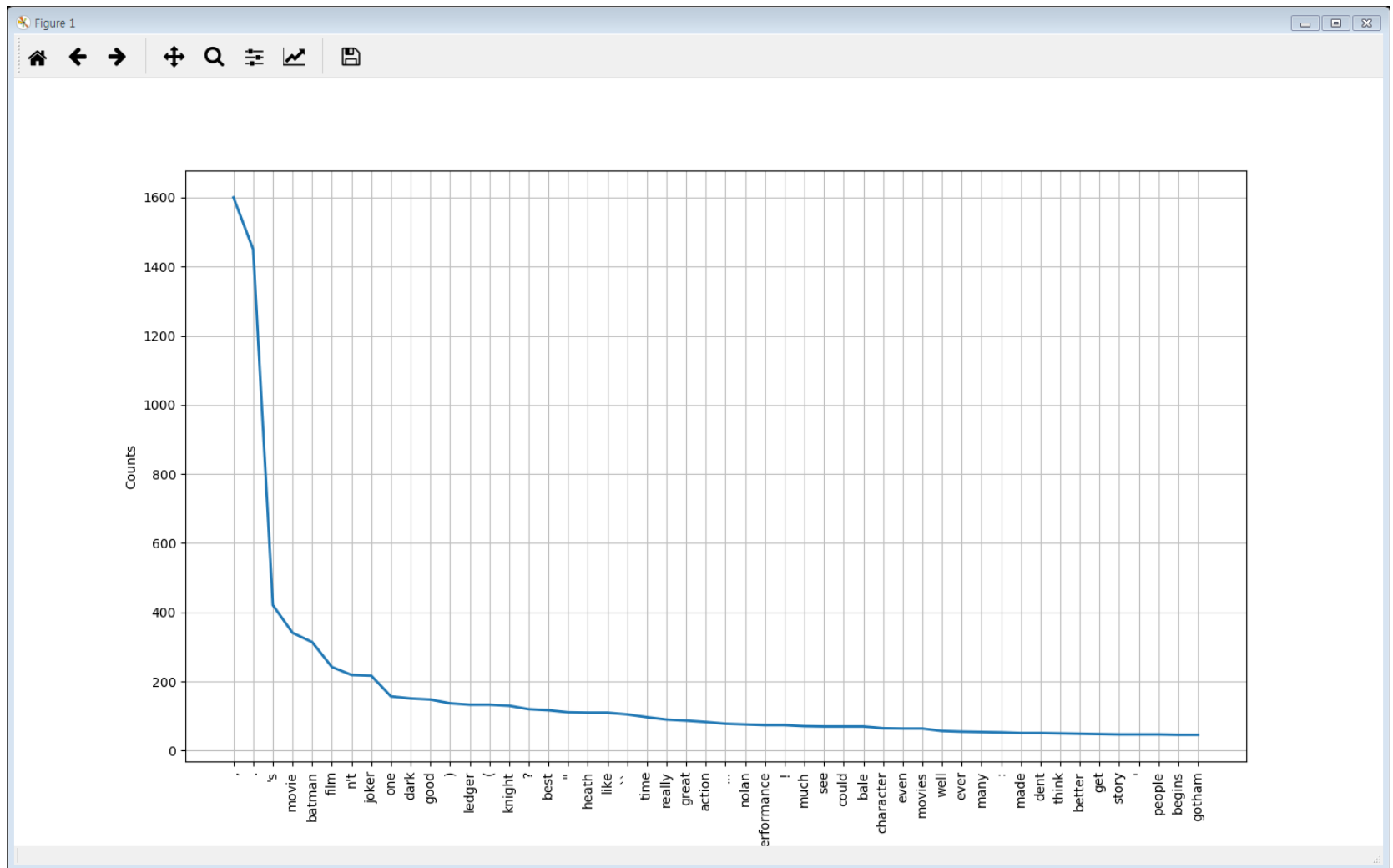
---

- TIPS

- nltk의 Text 함수를 사용해 토큰화한 결과를 저장한다
- plot 함수를 사용해 그래프를 시각화해본다

## 실습 2-2. 토큰의 등장 횟수 시각화하기

· 출력 결과



## 실습 2-3. 문맥상 유사한 단어 출력하기

- 실습 1-3-4에서 수집하였던 영화 '다크 나이트(The Dark Knight)' 리뷰에서 문맥상 'Batman'과 'Joker'와 유사한(similar) 단어를 출력해 본다

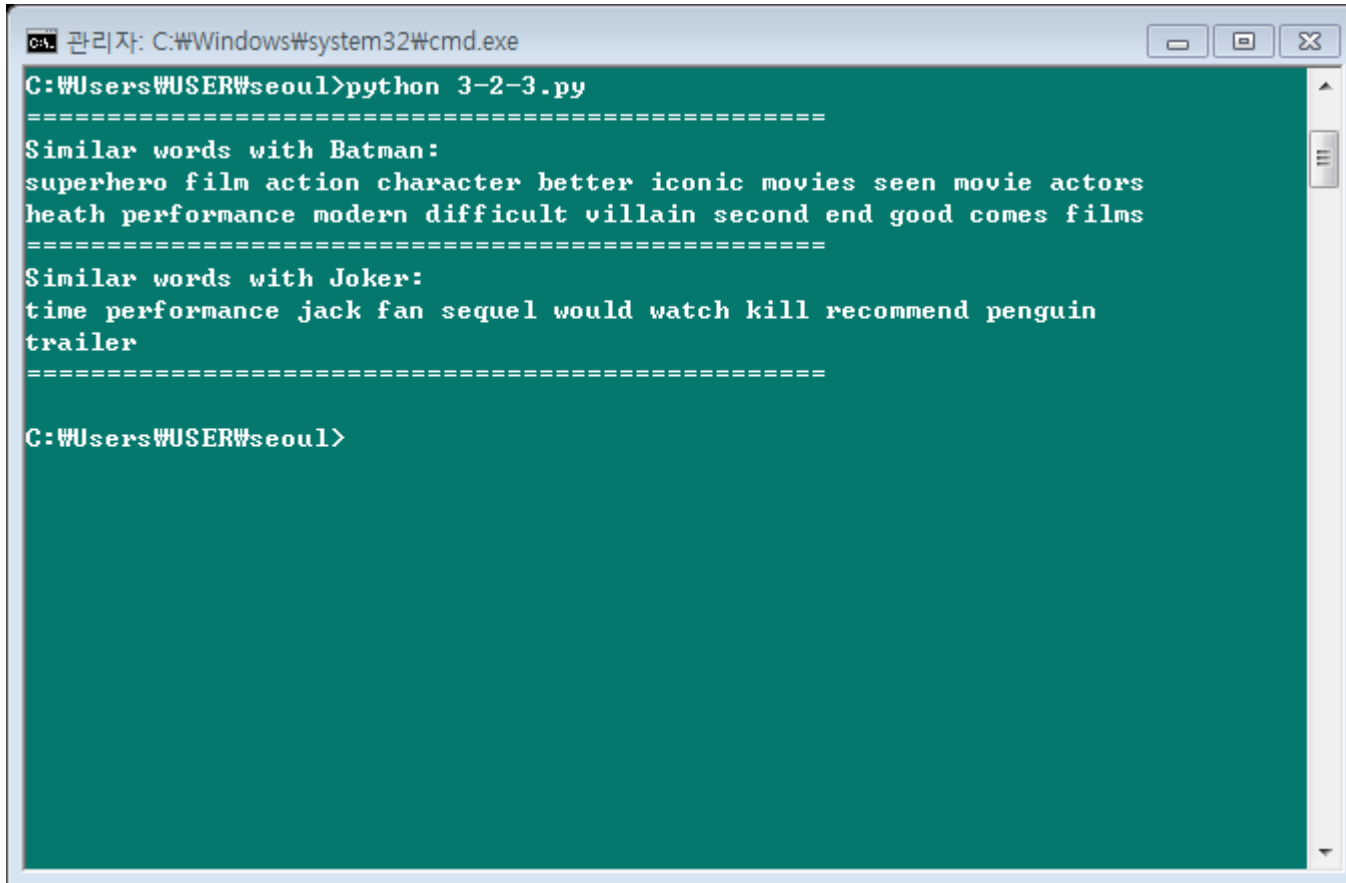


### · TIPS

- similar 함수를 사용해 유사한 단어를 출력한다

## 실습 2-3. 문맥상 유사한 단어 출력하기

- 출력 결과



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER>python 3-2-3.py
=====
Similar words with Batman:
superhero film action character better iconic movies seen movie actors
heath performance modern difficult villain second end good comes films
=====
Similar words with Joker:
time performance jack fan sequel would watch kill recommend penguin
trailer
=====
C:\Users\USER>
```

## 실습 2-4. 텍스트의 연어(collocation) 출력하기

- 연어(collocation) [source: <https://en.wikipedia.org/wiki/Collocation>]

- 문장 내에서 유의미하게 자주 함께 등장하는 단어의 쌍

- 예를 들어, 'dark'와 'black'이 거의 유사한 의미를 지니지만, 'black chocolate' 보다는 'dark chocolate'이라는 단어의 조합이 훨씬 선호된다. 이 경우 'dark'와 'chocolate' 사이에 연어 관계가 있다고 볼 수 있다

- 한국어에도 수많은 연어가 존재한다. 예를 들어, 일상적으로 '장갑을 끼다', '옷을 입다', '신발을 신다' 라고 표현하지 '장갑을 입다', '옷을 신다', '신발을 끼다' 라고 표현하지 않는다

- 실습 1-3-4에서 수집하였던 영화 '다크 나이트(The Dark Knight)' 리뷰 텍스트 내의 연어를 출력해 본다

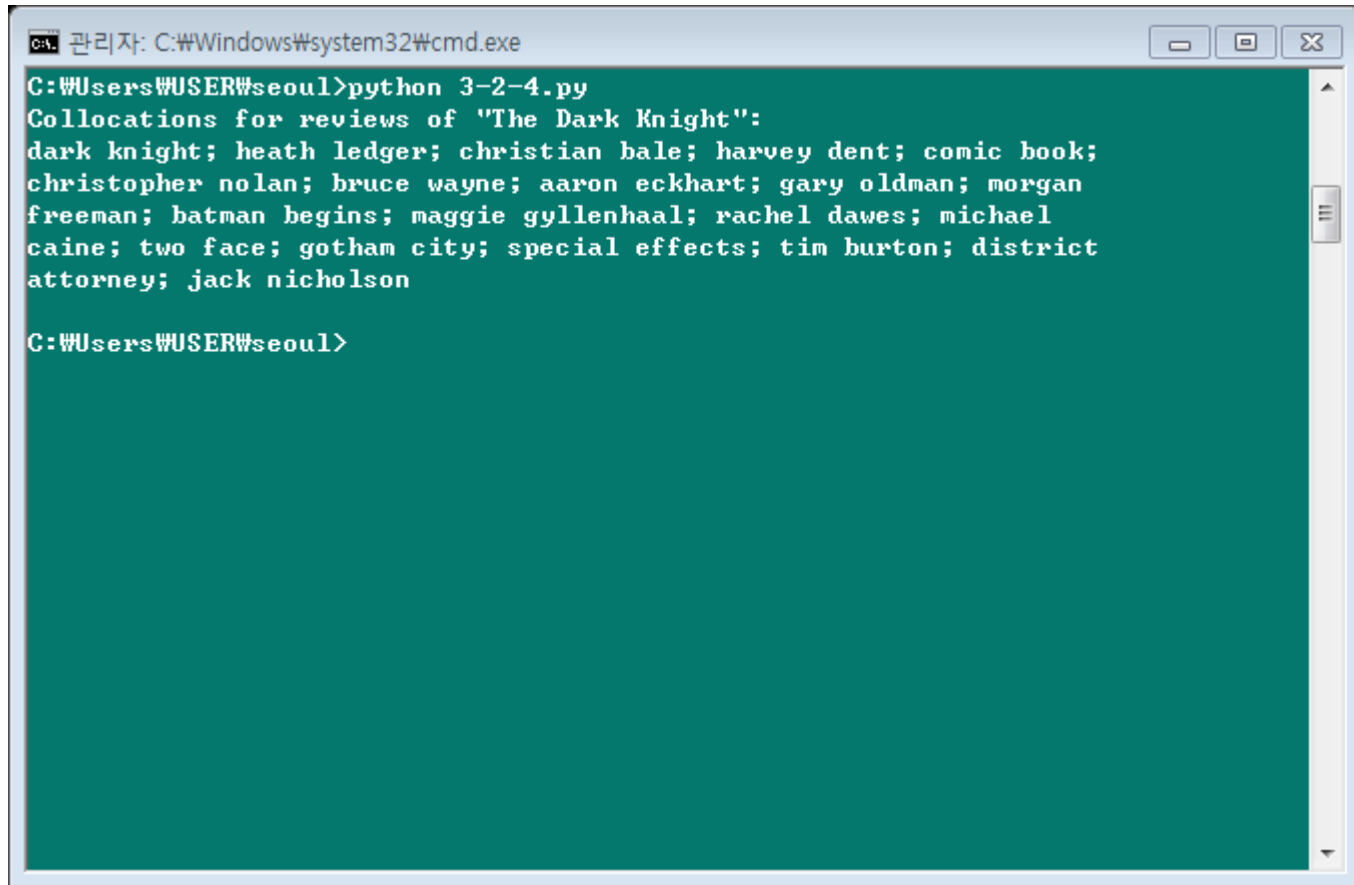
- TIPS

- collocations 함수를 사용해 연어를 출력한다



## 실습 2-4. 텍스트의 연어(collocation) 출력하기

- 출력 결과



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-2-4.py
Collocations for reviews of "The Dark Knight":
dark knight; heath ledger; christian bale; harvey dent; comic book;
christopher nolan; bruce wayne; aaron eckhart; gary oldman; morgan
freeman; batman begins; maggie gyllenhaal; rachel daves; michael
caine; two face; gotham city; special effects; tim burton; district
attorney; jack nicholson

C:\Users\USER\seoul>
```

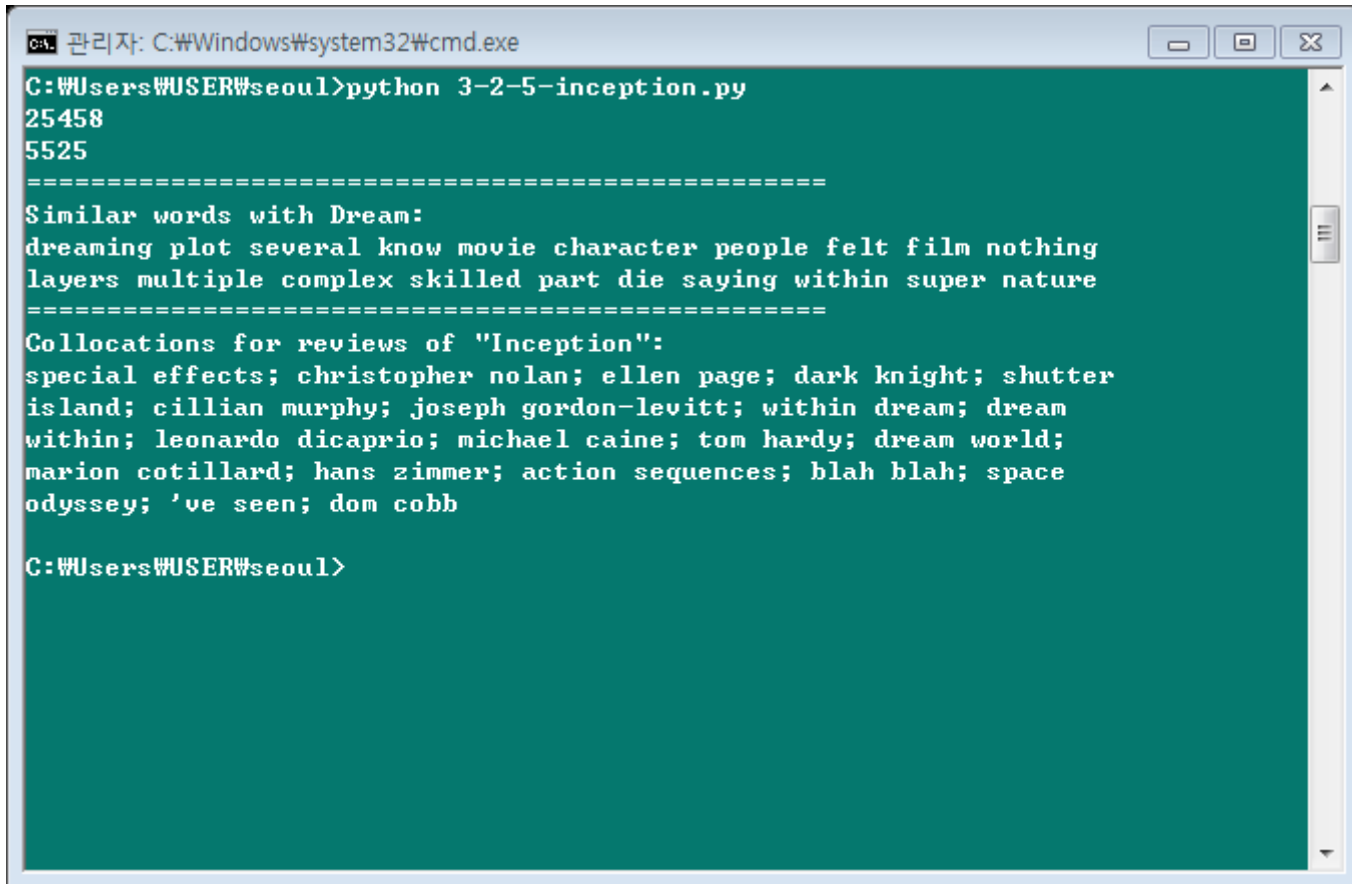
## 실습 2-5. 다른 리뷰 텍스트 탐색하기

---

- 2-1부터 2-4까지의 실습 내용을 실습 1-3-5에서 수집했던 다른 영화 리뷰 데이터에 적용해 보자
- 앞의 예제의 내용에 구매 받지 않고 자유롭게 적용해 본다

## 실습 2-5. 다른 리뷰 텍스트 탐색하기

- 출력 결과(인셉션)

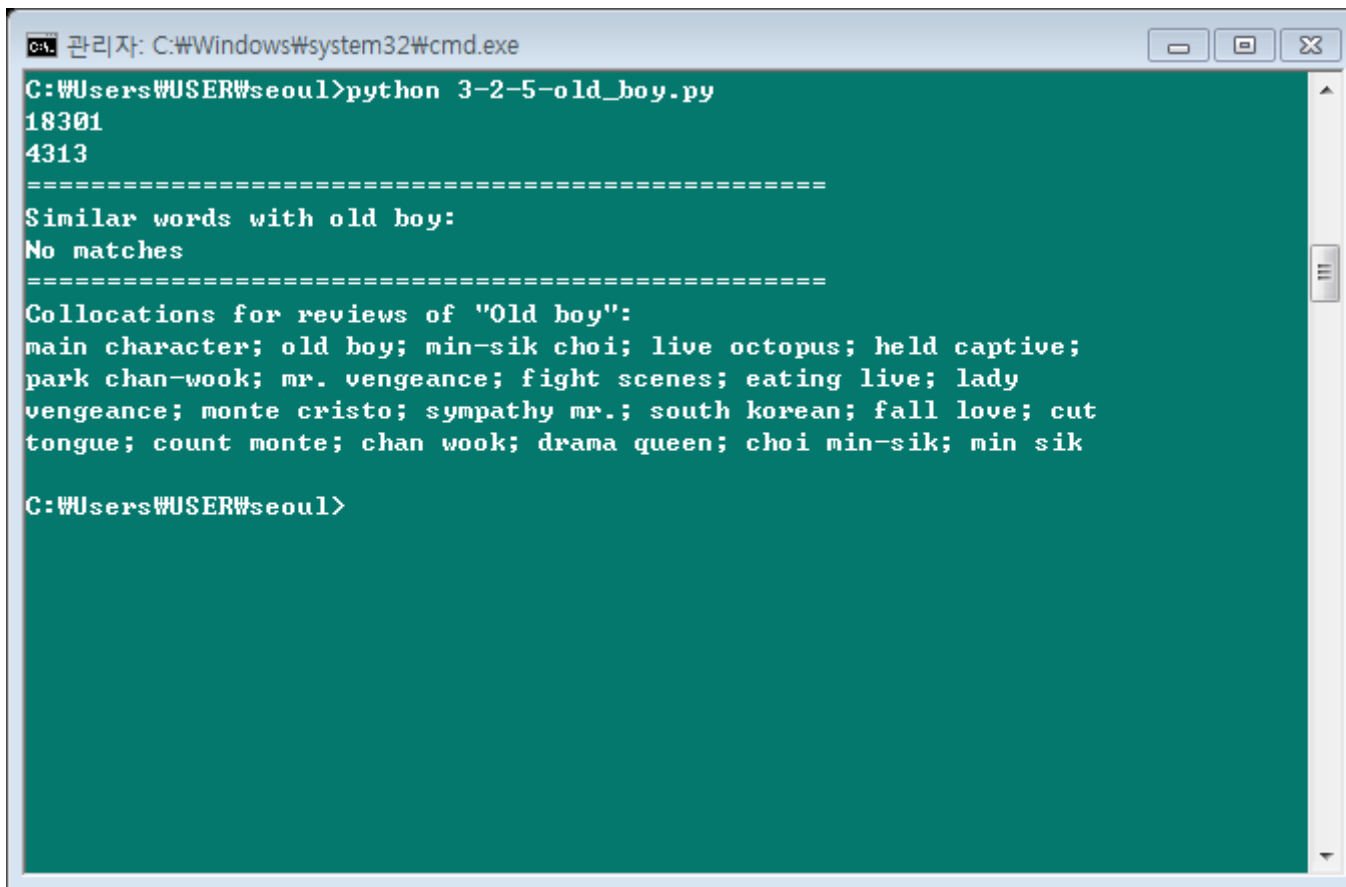


```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-2-5-inception.py
25458
5525
=====
Similar words with Dream:
dreaming plot several know movie character people felt film nothing
layers multiple complex skilled part die saying within super nature
=====
Collocations for reviews of "Inception":
special effects; christopher nolan; ellen page; dark knight; shutter
island; cillian murphy; joseph gordon-levitt; within dream; dream
within; leonardo dicaprio; michaelaine; tom hardy; dream world;
marion cotillard; hans zimmer; action sequences; blah blah; space
odyssey; 've seen; dom cobb

C:\Users\USER\seoul>
```

## 실습 2-5. 다른 리뷰 텍스트 탐색하기

- 출력 결과(올드보이)

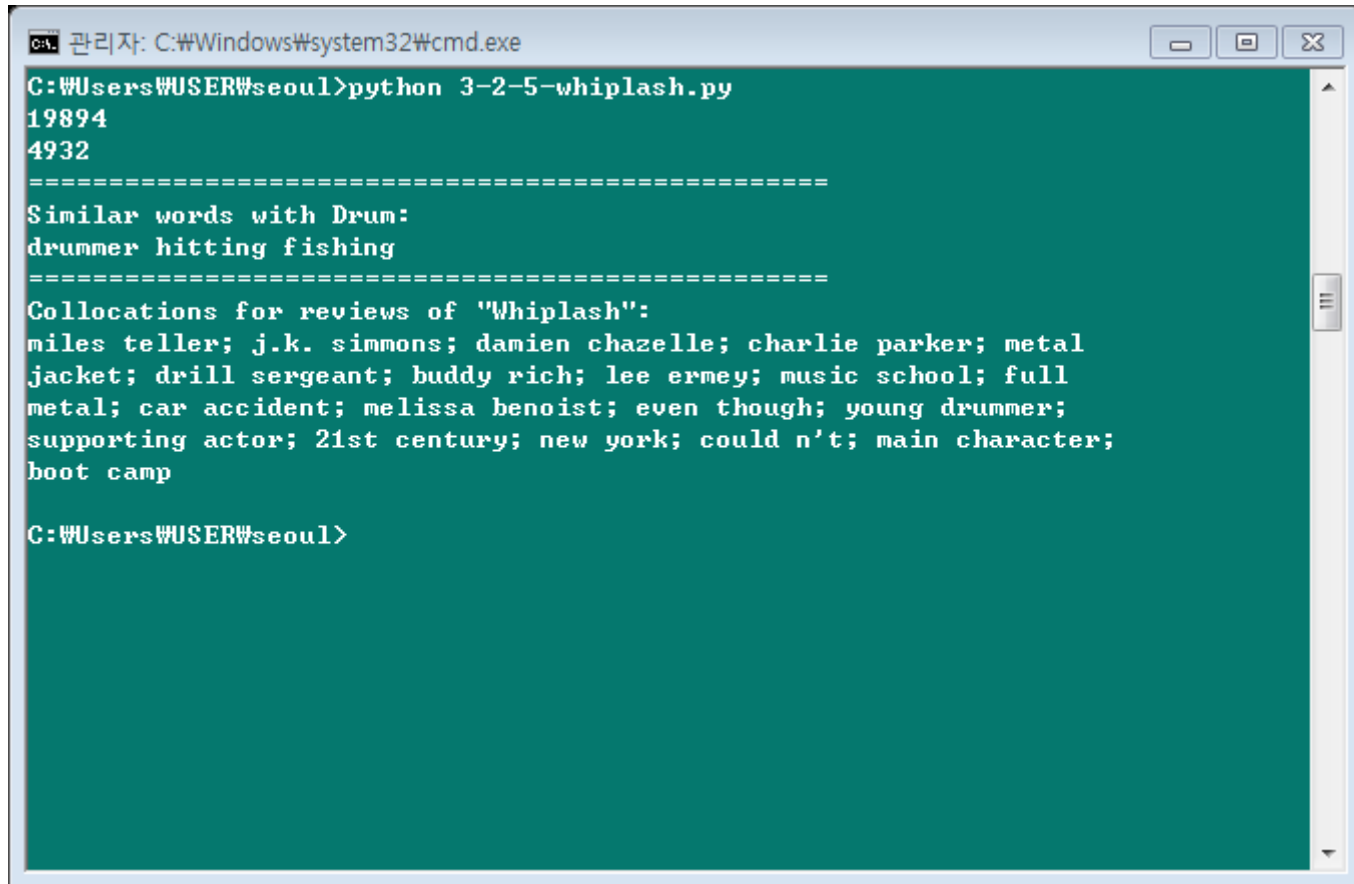


```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-2-5-old_boy.py
18301
4313
=====
Similar words with old boy:
No matches
=====
Collocations for reviews of "Old boy":
main character; old boy; min-sik choi; live octopus; held captive;
park chan-wook; mr. vengeance; fight scenes; eating live; lady
vengeance; monte cristo; sympathy mr.; south korean; fall love; cut
tongue; count monte; chan wook; drama queen; choi min-sik; min sik

C:\Users\USER\seoul>
```

## 실습 2-5. 다른 리뷰 텍스트 탐색하기

- 출력 결과(위플래시)



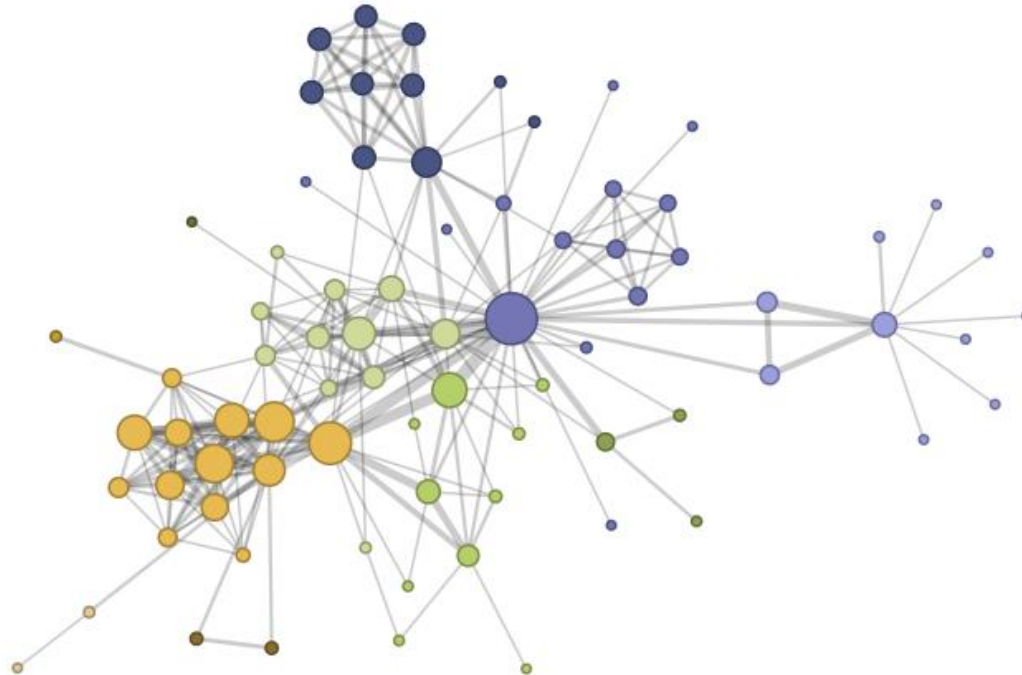
```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 3-2-5-whiplash.py
19894
4932
=====
Similar words with Drum:
drummer hitting fishing
=====
Collocations for reviews of "Whiplash":
miles teller; j.k. simmons; damien chazelle; charlie parker; metal
jacket; drill sergeant; buddy rich; lee ermey; music school; full
metal; car accident; melissa benoist; even though; young drummer;
supporting actor; 21st century; new york; could n't; main character;
boot camp

C:\Users\USER\seoul>
```

## 실습 3-1. 연관 단어 그래프 만들기(advanced)

· 실습 2-2-3에서 결과로 출력한 lemmatize된 영화 다크 나이트의 리뷰를 바탕으로, 연관된(같은 문장에서 등장하는) 명사들끼리 연결된 그래프를 만들어 출력해 본다

- 구현하기 너무 어려우면 완성된 코드를 참고만 해 연관 단어 그래프 만드는 법을 익힌다



## 실습 3-1. 연관 단어 그래프 만들기(advanced)

### · TIPS

- Unique한 명사들만 추출해 이를 리스트로 만든다
- (총 문장의 개수)를 행의 수로, (unique한 명사의 개수)를 열의 수로 하는 행렬을 만들어 이를 자신의 전치 행렬 (Transpose matrix)와 곱한 co-occurrence matrix를 만든다
- 비가중 무향 그래프(Unweighted, undirected graph)를 만들어 명사들 간의 연관 관계를 표현한다
- numpy와 networkx, 그리고 matplotlib 패키지를 활용한다

## 실습 3-1. 연관 단어 그래프 만들기(advanced)

· 참고: co-occurrence matrix

- 특정 문장에서 두 단어가 얼마나 자주 함께 등장하는지를 표현하기 위한 행렬로, occurrence matrix를 자신의 전치 행렬과 곱(matrix multiplication)해 계산한다

- 예시: 문장 2개(문장1, 문장2)와 명사 3개(고양이, 강아지, 치타)가 있고 각 단어가 각 문장에서 등장하는 횟수는 아래와 같다고 가정하자

	고양이	강아지	치타
문장1	1	0	1
문장2	0	1	0

■ occurrence matrix는 아래와 같다

$$M = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \text{ and } M^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$



## 실습 3-1. 연관 단어 그래프 만들기(advanced)

· 참고: co-occurrence matrix

■ co-occurrence matrix는 다음과 같이 쉽게 계산할 수 있다

$$M^T M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

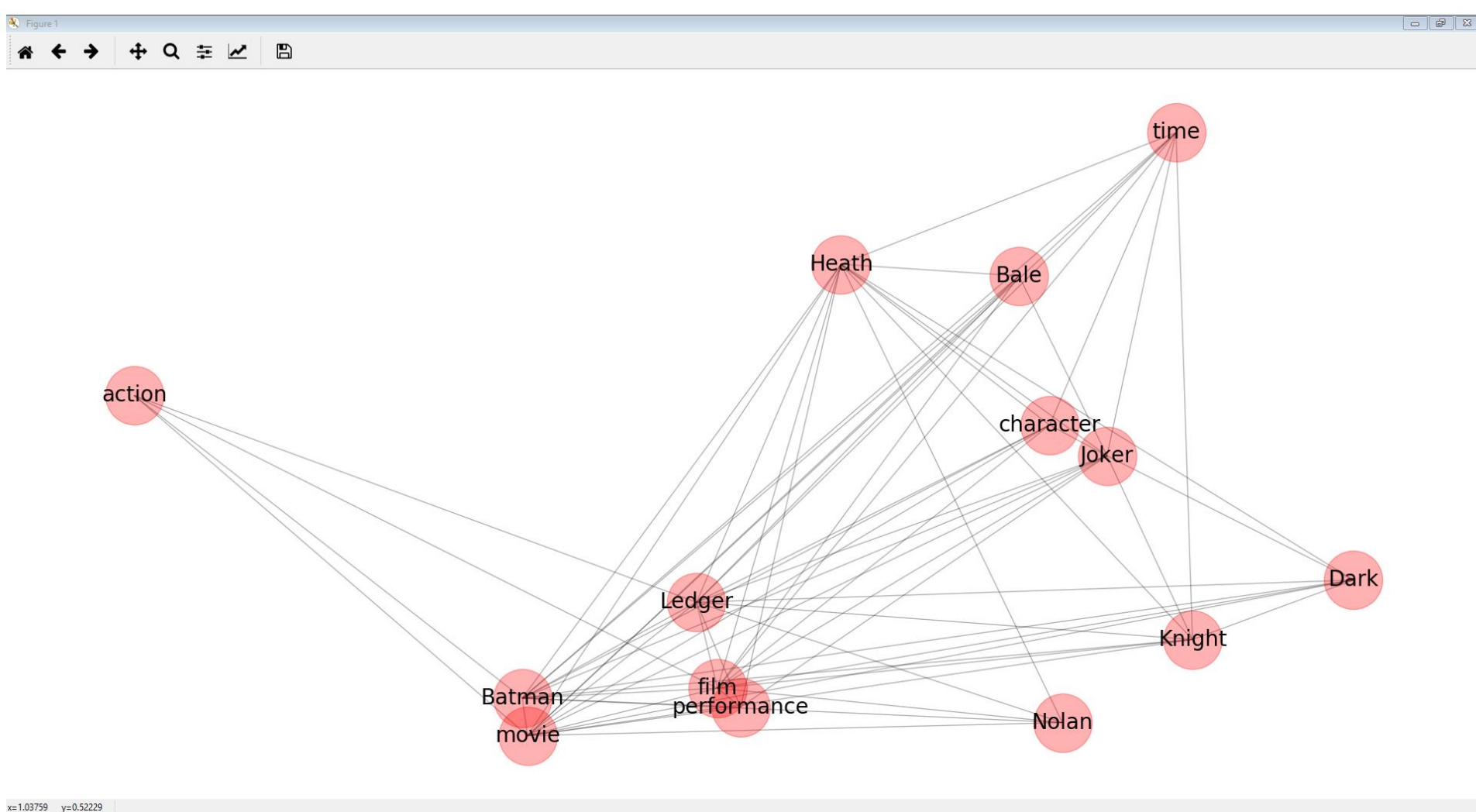
	고양이	강아지	치타
고양이	1	0	1
강아지	0	1	0
치타	1	0	1

■ 고양이와 치타가 문장1에서 함께 같이 등장하므로 co-occurrence matrix의 (1,3) 그리고 (3,1) 원소가 1이다

■ co-occurrence matrix는 무조건 대칭행렬(symmetric matrix)이고 대각선(diagonal) 원소의 값은 1이다 – 이유는?

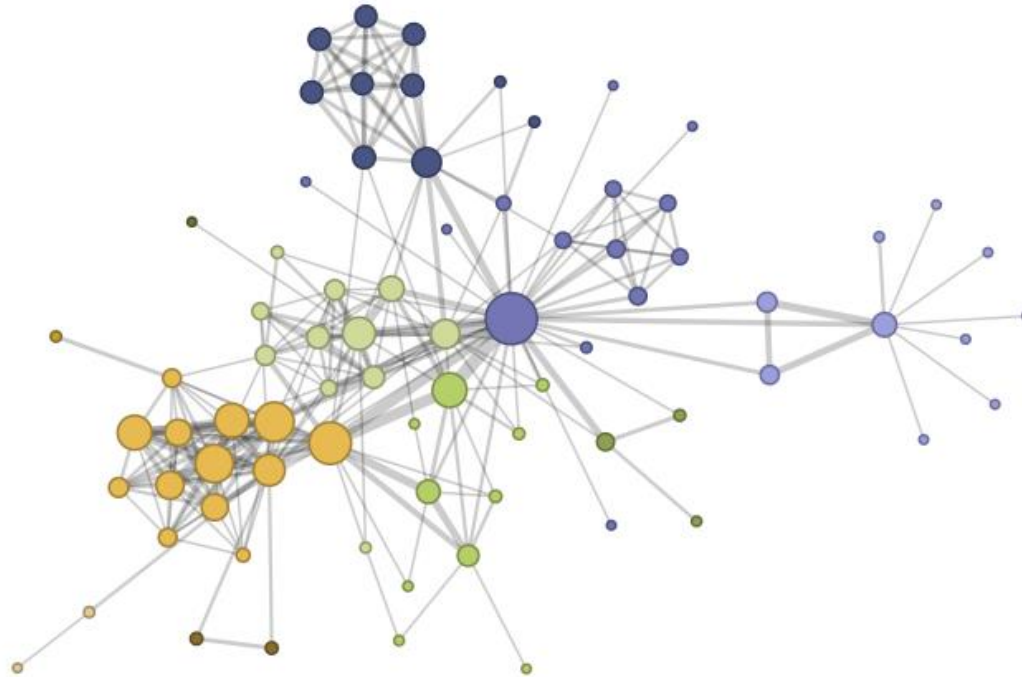
## 실습 3-1. 연관 단어 그래프 만들기(advanced)

· 출력 결과



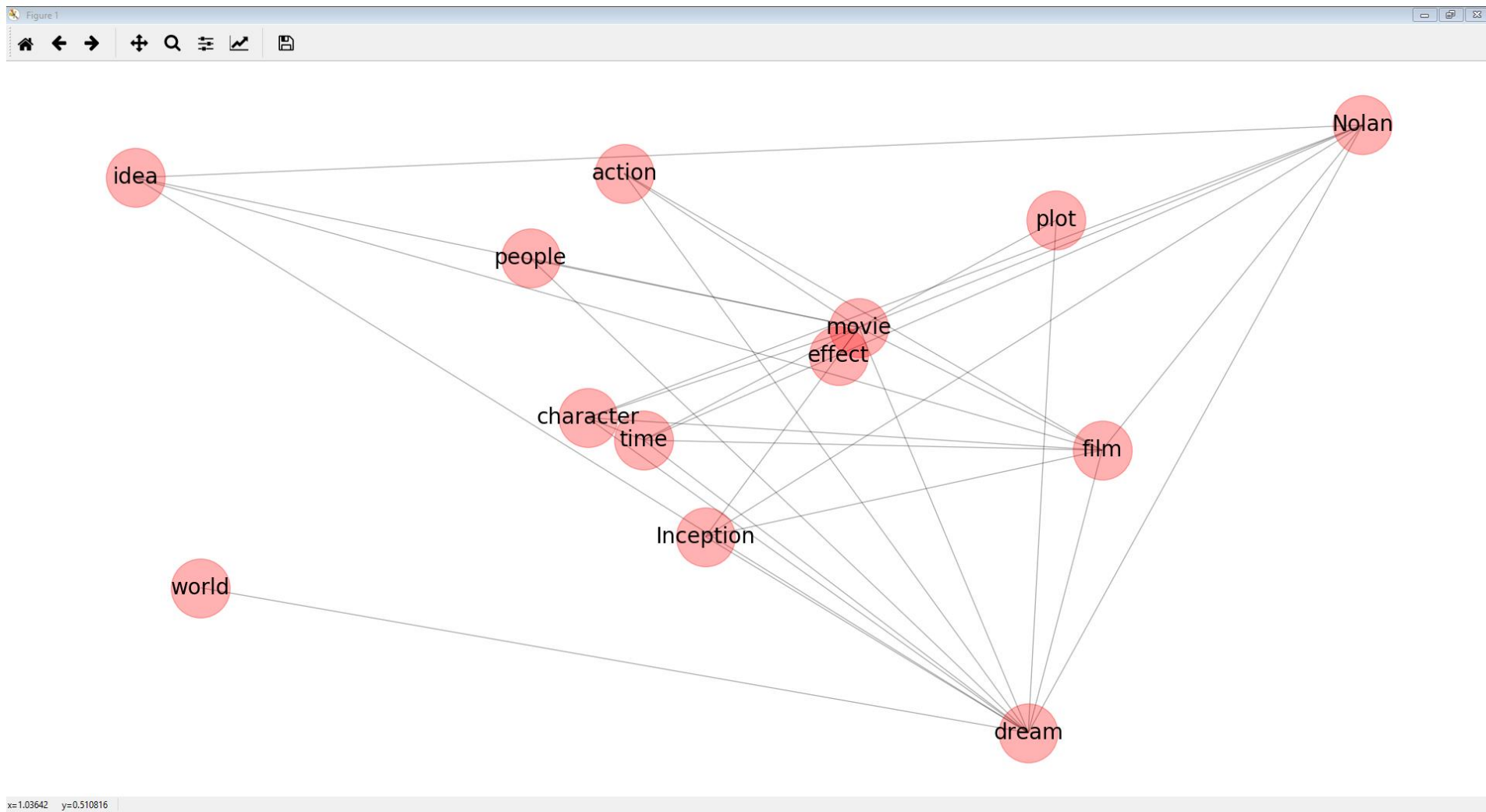
## 실습 3-2. 연관 단어 그래프 만들기(advanced) (2)

- 실습 2-2-4에서 결과로 나온 다른 영화 리뷰들로도 연관 단어 그래프를 만들어 본다
  - 구현하기 너무 어려우면 완성된 코드를 참고만 해 연관 단어 그래프 만드는 법을 익힌다
  - 출력 결과는 데이터와 파라미터 세팅에 따라 얼마든지 달라질 수 있다



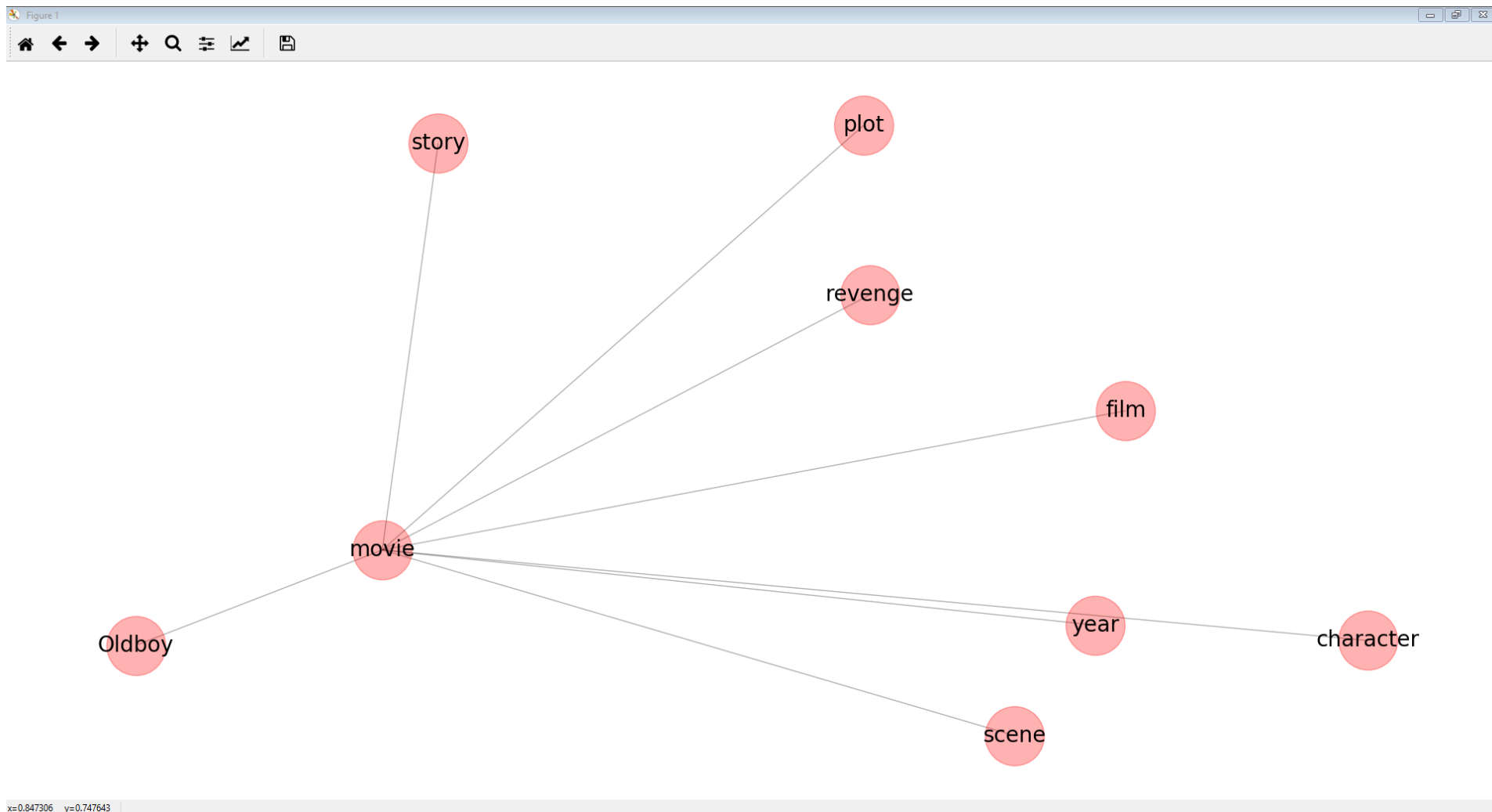
## 실습 3-2. 연관 단어 그래프 만들기(advanced) (2)

· 출력 결과(인셉션)



## 실습 3-2. 연관 단어 그래프 만들기(advanced) (2)

· 출력 결과(올드보이)



## 실습 3-2. 연관 단어 그래프 만들기(advanced) (2)

· 출력 결과(위플래시)

