

Text Preprocessing 실습

비정형 데이터 분석을 통한 효율적인 의사결정

박진수 교수

**Big Data Institute,
Seoul National University**

실습 1 – 영어 텍스트 전처리하기

There is nothing to
writing.
All you have to do is
sit down at a
typewriter and bleed.

Ernest
Hemingway

실습 1-1. 영어 문장 토큰화하기

- 파이썬의 nltk 모듈을 활용하여, 아래 문장들을 토큰화(Tokenization)한 결과를 출력한다

- Sentence 1: “My only regret in life is that I did not drink more wine.”

- Sentence 2: “I drink to make other people more interesting.”

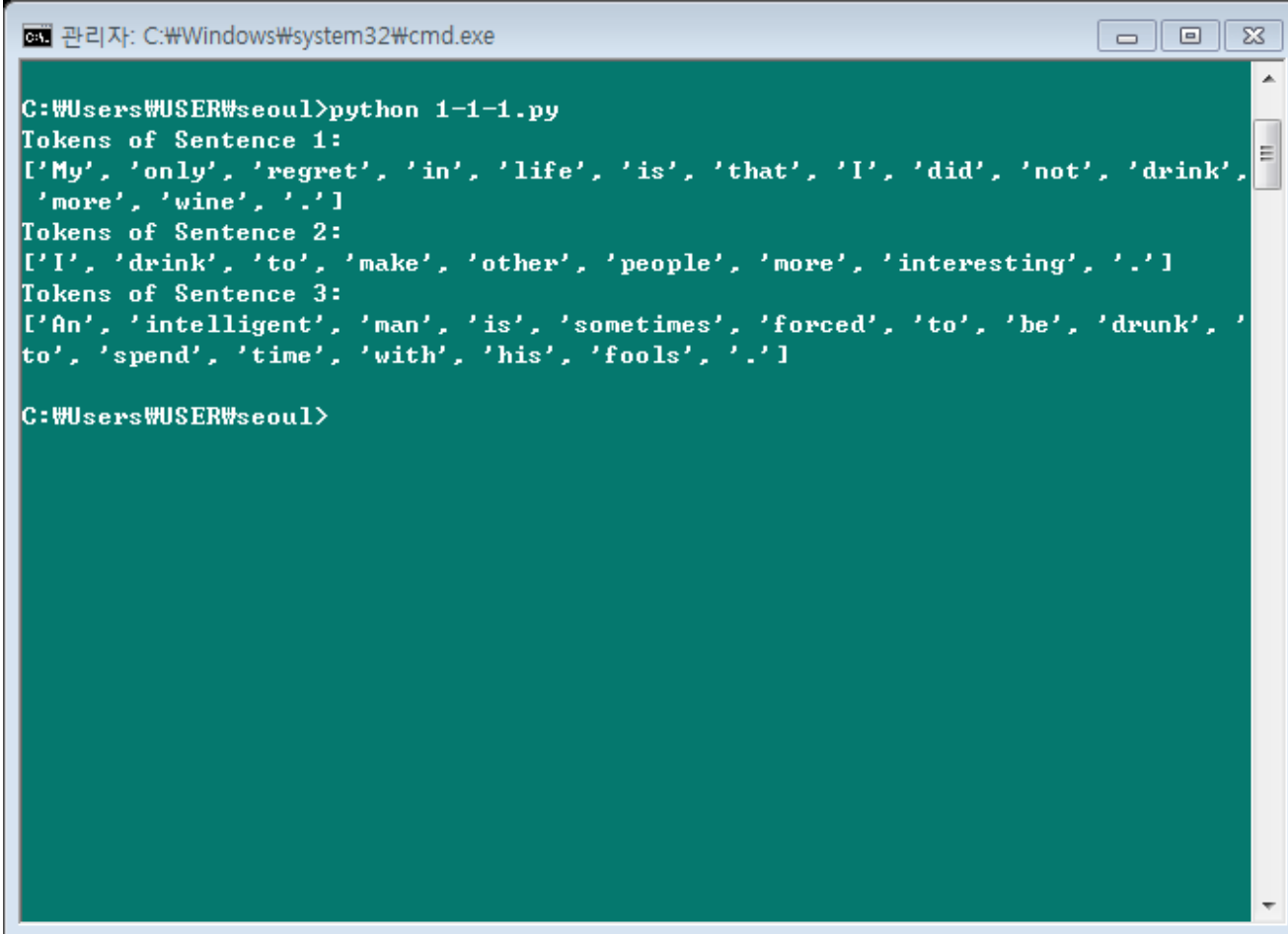
- Sentence 3: “An intelligent man is sometimes forced to be drunk to spend time with his fools.”

- TIPS

- nltk 패키지의 word_tokenize 함수를 사용한다

실습 1-1. 영어 문장 토큰화하기

- 출력 결과



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

C:\Users\WUSER\seoul>python 1-1-1.py
Tokens of Sentence 1:
['My', 'only', 'regret', 'in', 'life', 'is', 'that', 'I', 'did', 'not', 'drink',
'more', 'wine', '.']
Tokens of Sentence 2:
['I', 'drink', 'to', 'make', 'other', 'people', 'more', 'interesting', '.']
Tokens of Sentence 3:
['An', 'intelligent', 'man', 'is', 'sometimes', 'forced', 'to', 'be', 'drunk',
to', 'spend', 'time', 'with', 'his', 'fools', '.']

C:\Users\WUSER\seoul>
```

실습 1-2. 영어 문장 품사 태깅(POS tagging)하기

- 아래 문장들의 품사를 태깅해 출력한다

- Sentence 1: “My only regret in life is that I did not drink more wine.”

- Sentence 2: “I drink to make other people more interesting.”

- Sentence 3: “An intelligent man is sometimes forced to be drunk to spend time with his fools.”

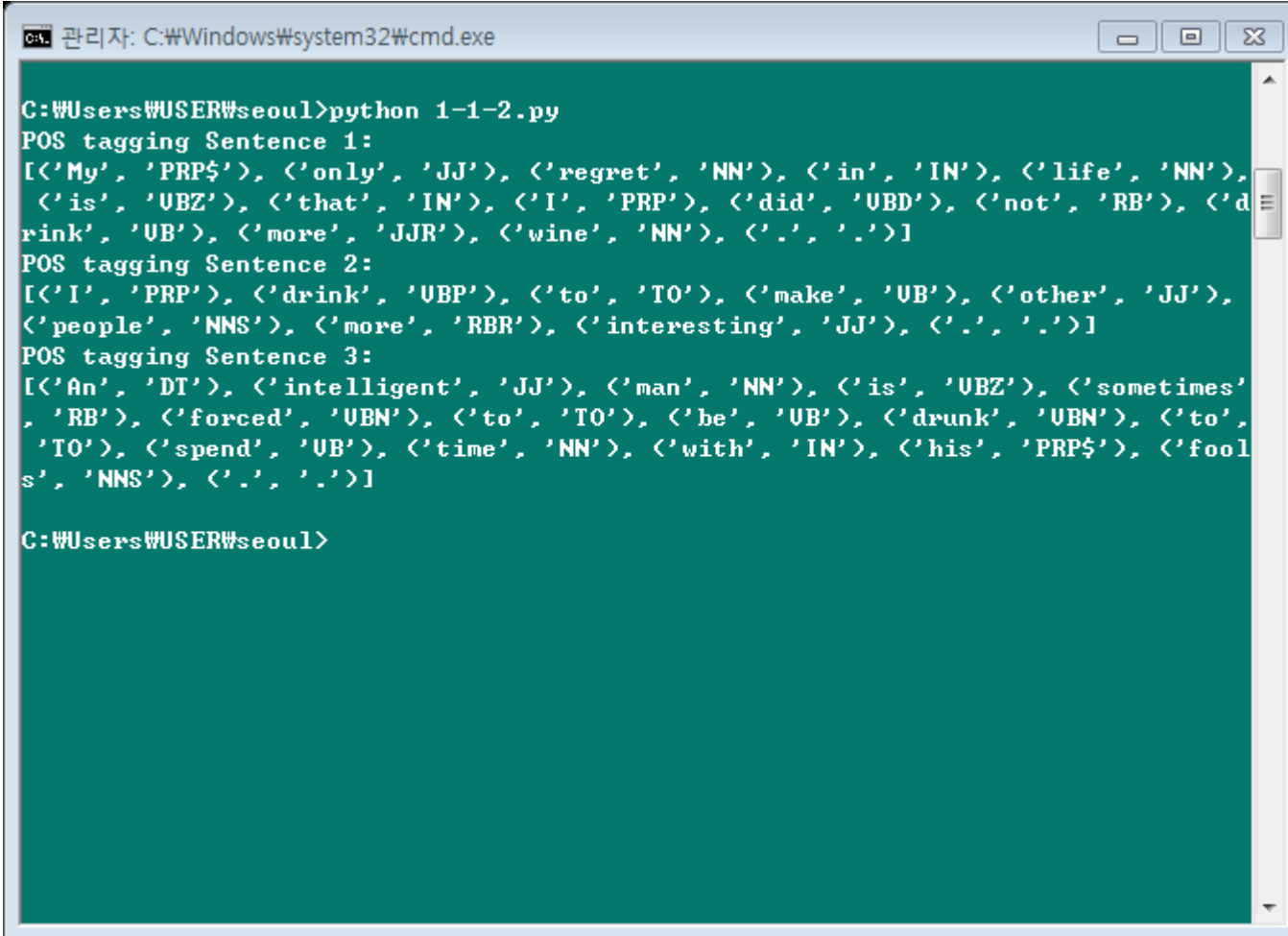
- TIPS

- nltk 패키지의 pos_tag 함수를 사용한다

- 토큰화 된 결과를 활용한다

실습 1-2. 영어 문장 품사 태깅(POS tagging)하기

- 출력 결과



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

C:\Users\USER\seoul>python 1-1-2.py
POS tagging Sentence 1:
[('My', 'PRP$'), ('only', 'JJ'), ('regret', 'NN'), ('in', 'IN'), ('life', 'NN'),
 ('is', 'VBZ'), ('that', 'IN'), ('I', 'PRP'), ('did', 'VBD'), ('not', 'RB'), ('d
rink', 'UB'), ('more', 'JJR'), ('wine', 'NN'), ('.', '.')]
POS tagging Sentence 2:
[('I', 'PRP'), ('drink', 'UBP'), ('to', 'TO'), ('make', 'UB'), ('other', 'JJ'),
 ('people', 'NNS'), ('more', 'RBR'), ('interesting', 'JJ'), ('.', '.')]
POS tagging Sentence 3:
[('An', 'DT'), ('intelligent', 'JJ'), ('man', 'NN'), ('is', 'VBZ'), ('sometimes'
, 'RB'), ('forced', 'UBN'), ('to', 'TO'), ('be', 'UB'), ('drunk', 'UBN'), ('to',
 'TO'), ('spend', 'UB'), ('time', 'NN'), ('with', 'IN'), ('his', 'PRP$'), ('fool
s', 'NNS'), ('.', '.')]

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

실습 1-2. 영어 문장 품사 태깅(POS tagging)하기

· 참고: nltk의 POS 태그 리스트 (source: <https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/>)

태그	품사	번역
CC	Coordinating conjunction	등위접속사
CD	Cardinal digit	기수
DT	Determiner	한정사
EX	Existential	There (is), there (exists)
FW	Foreign word	외래어
IN	Preposition/subordinating conjunction	전치사/종속 접속사
JJ	Adjective	형용사
JJR	Adjective (comparative)	형용사(비교급)
JJS	Adjective (superlative)	형용사(최상급)
LS	List marker	리스트 마커(1), 2), 3) 등)
MD	Modal	조동사
NN	Noun (singular)	명사(단수)
NNS	Noun (plural)	명사(복수)
NNP	Proper noun (singular)	고유 명사(단수)
NNPS	Proper noun (plural)	고유 명사(복수)
PDT	Predeterminer	전치 한정사
POS	Possessive ending	소유격 's

실습 1-2. 영어 문장 품사 태깅(POS tagging)하기

· 참고: nltk의 POS 태그 리스트 – 계속 (source: <https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/>)

태그	품사	번역
PRP	Personal noun	인칭대명사
PRP\$	Possessive pronoun	소유격
RB	Adverb	부사
RBR	Adverb (comparative)	부사(비교급)
RBS	Adverb (superlative)	부사(최상급)
RB	Particle	불변화사
TO	To	To부정사
UH	Interjection	감탄사
VB	Verb (Base form)	동사(기본형)
VBD	Verb (Past tense)	동사(과거형)
VCN	Verb (Past participle)	동사(과거 분사)
VBP	Verb (Singular, present)	동사(현재 단수형)
VBZ	Verb (3 rd person, singular, present)	동사(3인칭 현재 단수형)
WDT	Wh-determiner	한정사(which)
WP	Wh-pronoun	관계대명사(who, what)
WRB	Wh-adverb	관계부사(where, when)

실습 1-3. 단어의 기본형 찾기(Lemmatization)

· Lemmatization: 단어의 기본형 찾기(source: <https://en.wikipedia.org/wiki/Lemmatisation#Description>)

- 한국어와는 달리 영어는 단어의 기본형을 찾기가 상대적으로 용이하다
- Lemmatization은 단어의 어근에 기반하여 단어의 기본형(lemma)를 찾아준다
- 예시
 - play, played, playing → play
 - known, knew, knowing → know
 - apples → apple

실습 1-3. 단어의 기본형 찾기(Lemmatization)

- 아래 세 문장을 lemmatize해 그 결과를 출력해 본다

- Sentence 1: “My only regret in life is that I did not drink more wine.”

- Sentence 2: “I drink to make other people more interesting.”

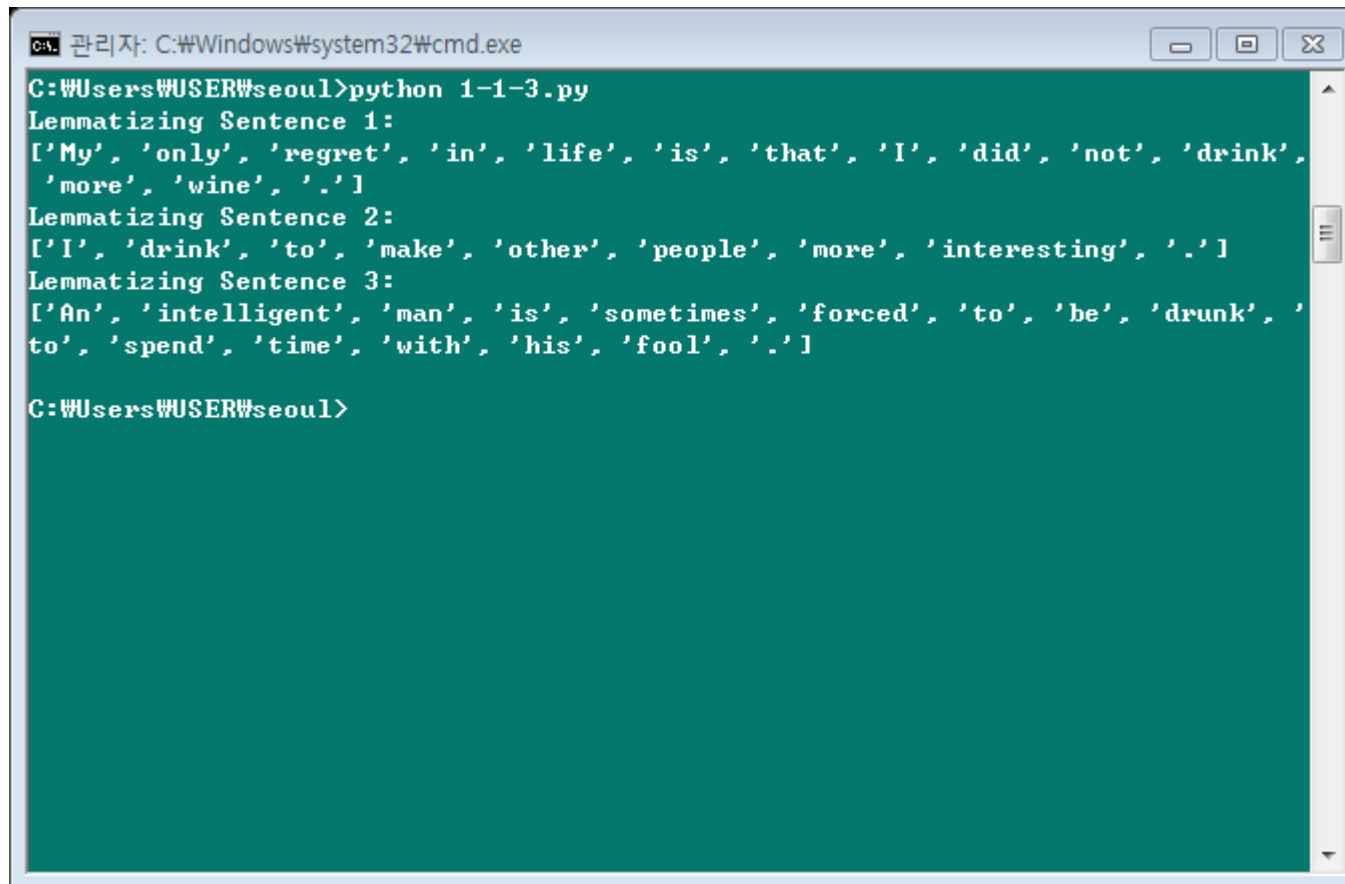
- Sentence 3: “An intelligent man is sometimes forced to be drunk to spend time with his fools.”

- TIPS

- nltk 패키지의 wordnet.WordNetLemmatizer 함수를 활용한다

실습 1-3. 단어의 기본형 찾기(Lemmatization)

- 출력 결과



```
관리자: C:\Windows\system32\cmd.exe
C:\Users\USER\seoul>python 1-1-3.py
Lemmatizing Sentence 1:
['My', 'only', 'regret', 'in', 'life', 'is', 'that', 'I', 'did', 'not', 'drink',
 'more', 'wine', '.']
Lemmatizing Sentence 2:
['I', 'drink', 'to', 'make', 'other', 'people', 'more', 'interesting', '.']
Lemmatizing Sentence 3:
['An', 'intelligent', 'man', 'is', 'sometimes', 'forced', 'to', 'be', 'drunk',
 'to', 'spend', 'time', 'with', 'his', 'fool', '.']

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

실습 1-4. Stopwords 제거하기

- Stopwords: 문맥상 유의하지 않아 전처리 과정에서 흔히 분석의 효율성을 위해 필터링하는 단어들
 - 대표적인 영어의 stopwords로는 the, is, at, which, on 등이 있음
 - nltk 패키지의 stopwords로는 다음을 참고한다: <http://www.nltk.org/book/ch02.html>

실습 1-4. Stopwords 제거하기

- 아래 세 문장에서 stopwords를 제거한 결과를 출력해 본다

- Sentence 1: “My only regret in life is that I did not drink more wine.”

- Sentence 2: “I drink to make other people more interesting.”

- Sentence 3: “An intelligent man is sometimes forced to be drunk to spend time with his fools.”

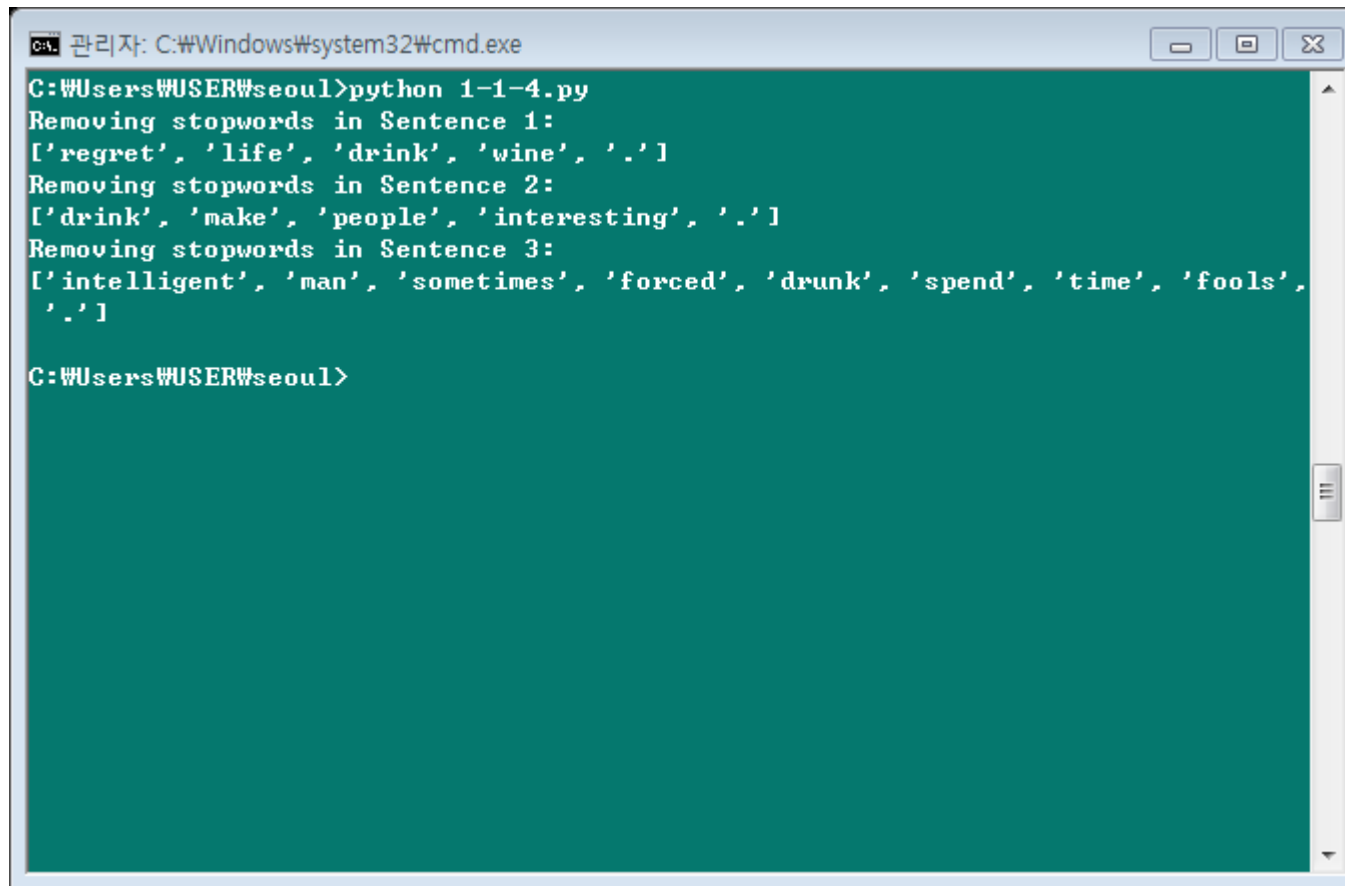
- TIPS

- nltk.corpus의 stopwords를 불러온다

- nltk 패키지의 stopwords list는 모두 소문자로 이루어져 있음을 유의한다

실습 1-4. Stopwords 제거하기





- 출력 결과



```
C:\> 관리자: C:\Windows\system32\cmd.exe
C:\Users\WUSER\seoul>python 1-1-4.py
Removing stopwords in Sentence 1:
['regret', 'life', 'drink', 'wine', '.']
Removing stopwords in Sentence 2:
['drink', 'make', 'people', 'interesting', '.']
Removing stopwords in Sentence 3:
['intelligent', 'man', 'sometimes', 'forced', 'drunk', 'spend', 'time', 'fools', '.']

C:\Users\WUSER\seoul>
```

실습 2 – IMDb 리뷰 데이터 전처리하기

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. 첫 번째 리뷰 전처리하기 (1)

- 실습 1-3-4에서 수집하였던 영화 ‘다크 나이트(The Dark Knight)’ 리뷰를 불러와 그 중 첫 번째 리뷰 텍스트를 토큰화하고 품사 태깅을 해 그 결과를 출력한다

A screenshot of a Windows File Explorer window. The address bar shows the path 'C:\Google Drive\2017-1\워드 데이터 사이언스 연구소 교육실습\src\[Session1] Web Crawling\Result-1-3-4.txt - Notepad++ [Administrator]'. The main pane shows a single file named 'result-1-3-4.txt'. The file is opened in Notepad++, which is visible at the bottom of the window. The text in the Notepad++ window is a long, repetitive paragraph about Christopher Nolan's movie 'The Dark Knight'. The text is repeated multiple times with minor variations. The status bar at the bottom of the Notepad++ window shows 'Normal text file', 'length: 242,563', 'lines: 101', 'Ln: 1', 'Col: 1', 'Sel: 0 | 0', 'Windows (CR LF)', 'UTF-8', and 'INS'.

실습 2-1. 첫 번째 리뷰 전처리하기 (1)

- TIPS

- readlines() 함수를 활용해 리뷰 데이터를 리스트로 받아온다
- 파일을 열 때 인코딩 설정을 꼭 'utf-8'으로 한다

실습 2-1. 첫 번째 리뷰 전처리하기 (1)

출력 결과

```
C:\Windows\system32\cmd.exe
C:\Users\WUSER\seoul>python 2-2-1.py
Tokens for first review:
['Christopher', 'Nolan', 's', 'second', 'bundle', 'of', 'joy', 'The', 'Dark', 'Knight', 'EXCEEDED', 'all', 'of', 'my', 'expectations', 'With', 'the', 'success', 'of', '2005', 's', 'reboot', 'of', 'the', 'Batman', 'franchise', 'they', 'took', 'what', 'was', 'already', 'established', 'and', 'expanded', 'it', 'amped', 'it', 'up', 'and', 'gave', 'a', 'deeper', 'darker', 'and', 'brooding', 'story', 'that', 'is', 'more', 'gripping', 'and', 'the', 'suspense', 'is', 'likely', 'to', 'catch', 'you', 'of', 'guard', 'several', 'times', 'throughout', 'Christian', 'Bale', 'delves', 'more', 'deeper', 'into', 'Batman', 'sworn', 'to', 'fight', 'evil', 'and', 'injustice', 'though', 'also', 'quite', 'reluctant', 'and', 'uncertain', 'if', 'his', 'crusade', 'can', 'ever', 'end', 'and', 'cleanse', 'his', 'inner', 'turmoil', 'from', 'his', 'fractured', 'soul', 'due', 'to', 'the', 'murder', 'of', 'his', 'beloved', 'parents', 'But', 'with', 'the', 'help', 'of', 'his', 'trusted', 'butler/ally', 'Alfred', 'played', 'superbly', 'by', 'Michael', 'Cane', 'grounds', 'him', 'gives', 'him', 'moral', 'support', 'and', 'keeps', 'him', 'in', 'check', 'But', 'the', 'real', 'star', 'of', 'the', 'show', 'is', 'Heath', 'Ledger', 'as', 'Batman', 's', 'most', 'deadly', 'enemy', 'The', 'Joker', 'I', 'can', 'HONESTLY', 'tell', 'you', 'that', 'as', 'good', 'as', 'Jack', 'Nicholson', 'was', 'in', 'Batman 89', 'he', 'is', 'CHILD', 'S', 'PLAY', 'compared', 'to', 'this', 'Joker', 'He', 'is', 'sadistic', 'psychotic', 'and', 'downright', 'SCARIER', 'and', 'PSYCHOLOGICALLY', 'disturbing', 'than', 'the', 'previous', 'incarnation', 'of', 'The', 'Clown', 'Prince', 'of', 'Crime', 'and', 'Ledger', 'gives', 'it', 'his', 'all', 'to', 'do', 'him', 'justice', 'along', 'with', 'the', 'original', 'cast', 'comes', 'some', 'fresh', 'faces', 'such', 'as', 'Aaron', 'Eckhart', 'Maggie', 'Gyllenhaal', 'and', 'more', 'I', 'must', 'say', 'though', 'I', 'liked', 'Katie', 'Holmes', 'Gyllenhaal', 'gives', 'a', 'much', 'better', 'performance', 'and', 'is', 'a', 'fan', 'cry', 'from', 'the', 'damsel-in-distress', 'stereotype', 'though', 'there', 's', 'a', 'little', 'of', 'it', 'THANKFULLY', 'that', 's', 'common', 'in', 'films', 'Bale', 'and', 'Gyllenhaal', 'have', 'MUCH', 'better', 'chemistry', 'this', 'time', 'around', 'more', 'so', 'than', 'Holmes', 'Even', 'better', 'the', 'fight', 'sequences', 'are', 'vastly', 'improved', 'and', 'feature', 'more', 'brutal', 'and', 'bone', 'crushing', 'combat', 'than', 'Begins', 'in', 'addition', 'to', 'new', 'technology', 'at', 'Batman', 's', 'disposal', 'Also', 'worth', 'mentioning', 'is', 'screenwriter', 'Jonathan', 'Nolan', 'who', 'gives', 'the', 'film',
```

<토큰>

```
C:\Windows\system32\cmd.exe
POS tags for first review:
[('Christopher', 'NNP'), ('Nolan', 'NNP'), ('s', 'POS'), ('second', 'JJ'), ('bundle', 'NN'), ('of', 'IN'), ('joy', 'NN'), ('The', 'DT'), ('Dark', 'NNP'), ('Knight', 'NNP'), ('EXCEEDED', 'NNP'), ('all', 'DT'), ('of', 'IN'), ('my', 'PRP$'), ('expectations', 'NNS'), ('With', 'IN'), ('the', 'DT'), ('success', 'NN'), ('of', 'IN'), ('2005', 'CD'), ('s', 'POS'), ('reboot', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Batman', 'NNP'), ('franchise', 'NN'), ('they', 'PRP'), ('took', 'VBD'), ('what', 'WP'), ('was', 'VBD'), ('already', 'RB'), ('established', 'VBN'), ('and', 'CC'), ('expanded', 'VBN'), ('it', 'PRP'), ('amped', 'VBD'), ('it', 'PRP'), ('up', 'RP'), ('and', 'CC'), ('gave', 'VBD'), ('a', 'DT'), ('deeper', 'NN'), ('darker', 'NN'), ('and', 'CC'), ('brooding', 'VBG'), ('story', 'NN'), ('that', 'WDT'), ('is', 'VBZ'), ('more', 'JJR'), ('gripping', 'JJ'), ('and', 'CC'), ('the', 'DT'), ('suspense', 'NN'), ('is', 'VBZ'), ('likely', 'JJ'), ('to', 'TO'), ('catch', 'VB'), ('you', 'PRP'), ('of', 'IN'), ('guard', 'JJ'), ('several', 'JJ'), ('times', 'NNS'), ('throughout', 'IN'), ('Christian', 'JJ'), ('Bale', 'NNP'), ('delves', 'VBZ'), ('more', 'JJR'), ('deeper', 'JJ'), ('into', 'IN'), ('Batman', 'NNP'), ('sworn', 'VBN'), ('to', 'TO'), ('fight', 'VB'), ('evil', 'NN'), ('and', 'CC'), ('injustice', 'NN'), ('though', 'RB'), ('also', 'RB'), ('quite', 'RB'), ('reluctant', 'JJ'), ('and', 'CC'), ('uncertain', 'JJ'), ('if', 'IN'), ('his', 'PRP$'), ('crusade', 'NN'), ('can', 'MD'), ('ever', 'RB'), ('end', 'VB'), ('and', 'CC'), ('cleanse', 'VB'), ('his', 'PRP$'), ('fractured', 'JJ'), ('soul', 'NN'), ('due', 'JJ'), ('to', 'TO'), ('the', 'DT'), ('murder', 'NN'), ('of', 'IN'), ('his', 'PRP$'), ('beloved', 'JJ'), ('parents', 'NNS'), ('But', 'CC'), ('with', 'IN'), ('the', 'DT'), ('help', 'NN'), ('of', 'IN'), ('his', 'PRP$'), ('trusted', 'VBN'), ('butler/ally', 'RB'), ('Alfred', 'NNP'), ('played', 'VBN'), ('superbly', 'RB'), ('by', 'IN'), ('Michael', 'NNP'), ('Cane', 'NNP'), ('grounds', 'NN'), ('him', 'PRP'), ('gives', 'VBZ'), ('him', 'PRP'), ('moral', 'JJ'), ('support', 'NN'), ('and', 'CC'), ('keeps', 'VBZ'), ('him', 'PRP'), ('in', 'IN'), ('check', 'NN'), ('But', 'CC'), ('the', 'DT'), ('real', 'JJ'), ('star', 'NN'), ('of', 'IN'), ('the', 'DT'), ('show', 'NN'), ('is', 'VBZ'), ('Heath', 'NNP'), ('Ledger', 'NNP'), ('as', 'IN'), ('Batman', 'NNP'), ('s', 'POS'), ('most', 'JJ'), ('deadly', 'RB'), ('enemy', 'NN'), ('The', 'DT'), ('Joker', 'NNP'), ('I', 'PRP'), ('can', 'MD'), ('HONESTLY', 'VB'), ('tell', 'VB'), ('you', 'PRP'), ('that', 'WDT'), ('as', 'RB'), ('good', 'JJ'), ('as', 'IN'), ('Jack', 'NNP'), ('Nicholson', 'NNP'),
```

<품사 태그>

실습 2-2. 첫 번째 리뷰 전처리하기 (2)

- 실습 1-3-4에서 수집하였던 영화 ‘다크 나이트(The Dark Knight)’ 리뷰를 불러와 그 중 첫 번째 리뷰 텍스트에서 stopwords를 제거하고 lemmatization을 수행해 그 결과를 출력한다

The image shows a Windows desktop environment. At the top, a taskbar displays the active application: "D:\Google Drive\2017-1\독서 데이터 사이언스 연구소 교육활성화실습\src\Web Crawling\result-1-3-4.txt - Notepad++ [Administrator]". Below the taskbar, the Notepad++ window is open, showing a file named "result-1-3-4.txt". The text in the editor is a review of Christopher Nolan's movie "The Dark Knight". The review is repeated multiple times, with some lines being truncated on the right side of the window. The status bar at the bottom of the Notepad++ window shows "Normal text file", "length: 242,563", "lines: 101", "Ln: 1", "Col: 1", "Sel: 0 | 0", "Windows (CR LF)", "UTF-8", and "INS".

실습 2-2. 첫 번째 리뷰 전처리하기 (2)

- TIPS

- readlines() 함수를 활용해 리뷰 데이터를 리스트로 받아온다
- 파일을 열 때 인코딩 설정을 꼭 'utf-8'으로 한다

실습 2-2. 첫 번째 리뷰 전처리하기 (2)

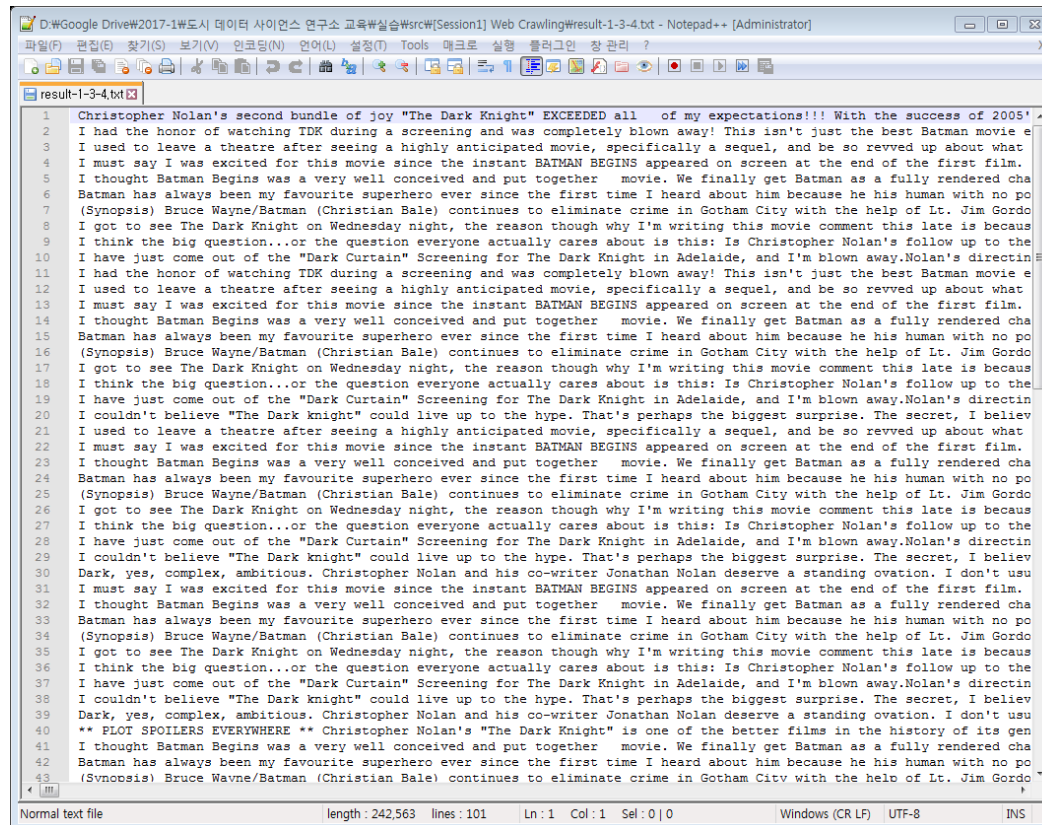
· 출력 결과

```
C:\Windows\system32\cmd.exe
C:\Users\WUSER\python 2-2-2.py
Lemmas of the first review:
['Christopher', 'Nolan', '"s', 'second', 'bundle', 'joy', '"', 'Dark', 'Knight',
, '"', 'EXCEEDED', 'expectation', '!', '!', '!', 'success', '2005', '"s', 'rebo
ot', 'Batman', 'franchise', '!', '!', '!', 'took', 'already', 'established', 'expanded',
, '!', 'amped', '!', '!', 'gave', 'deeper', '!', '!', 'darker', 'brooding', 'story', 'grippin
g', 'suspense', 'likely', 'catch', 'guard', 'several', 'time', 'throughout', '!',
, 'Christian', 'Bale', 'delf', 'deeper', 'Batman', '!', '!', 'sworn', 'fight', 'evil',
, 'injustice', '!', '!', 'though', 'also', 'quite', 'reluctant', 'uncertain', 'crusad
e', 'ever', 'end', 'cleanse', 'inner', 'turmoil', 'fractured', 'soul', 'due', 'm
urder', 'beloved', 'parent', '!', '!', 'help', 'trusted', 'butler/ally', 'Alfred', '<
', 'played', 'superbly', 'Michael', 'Cane', '>', 'ground', '!', '!', 'give', 'moral',
, 'support', '!', '!', 'keep', 'check', '!', '!', 'real', 'star', 'show', 'Heath', 'Ledger',
, 'Batman', '"s', 'deadly', 'enemy', '!', '!', 'Joker', '!', '!', 'HONESTLY', 'tell', ':',
, 'good', 'Jack', 'Nicholson', 'Batman'89', 'CHILD', '"$', 'PLAY', 'compared', 'J
oker', '!', '!', 'sadistic', '!', '!', 'psychotic', '!', '!', 'downright', 'SCARIER', 'PSYCHOLO
GICALLY', 'disturbing', 'previous', 'incarnation', 'Clown', 'Prince', 'Crime', '
Ledger', 'give', 'justice', '!', '!', 'Along', 'original', 'cast', 'come', 'fresh', '
face', 'Aaron', 'Eckhart', '!', '!', 'Maggie', 'Gyllenhaal', '!', '!', 'must', 'say', 'tho
ugh', 'liked', 'Katie', 'Holmes', '!', '!', 'Gyllenhaal', 'give', 'much', 'better', '
performance', 'far', 'cry', '"', '!', 'damsel-in-distress', '"', '!', 'stereotype', '<
, 'though', '"s', 'little', '!', '!', 'THANKFULLY', '>', '"s', 'common', 'film', '!',
, 'Bale', 'Gyllenhall', 'MUCH', 'better', 'chemistry', 'time', 'around', 'Holmes',
, '!', 'Even', 'better', '!', '!', 'fight', 'sequence', 'vastly', 'improved', 'feature
', 'brutal', 'bone', 'crushing', 'combat', '"', '!', 'Begins', '"', '!', 'addition', 'ne
w', 'technology', 'Batman', '"s', 'disposal', '!', '!', 'Also', 'worth', 'mentioning',
, 'screenwriter', 'Jonathan', 'Nolan', '!', '!', 'give', 'film', 'added', 'frosting',
, 'already', 'delicious', 'cake', '!', '!', 'Simply', 'put', '!', '!', 'Dark', 'Knight', 't
otally', 'bad', 'as', '"', '!', 'Begins', '!', '!', '!', 'action', 'great', '!', '!', 'plot',
, 'deeper', 'engrossing', '!', '!', 'applaud', 'Christopher', 'Nolan', '!', '!', 'Christian
', 'Bale', '!', '!', 'especially', 'Heath', 'Ledger', '<', 'sadly', 'passed', 'away',
, 'earlier', 'year', '>', 'aboard', 'believing', 'Mr.', 'Nolan', '"s', 'talent',
, 'second', 'installment', '!', '!', 'Although', 'may', 'feel', 'bit', 'melancholy', 'L
edger', '"s', 'death', '!', '!', 'final', 'note', 'say', 'sincerely', 'heart', ':',
, 'Remember', 'Heath', 'Ledger', 'honor', 'mind', 'heart', 'performance', '!', '!', 'hum
an', 'father', 'daughter', 'Matilda', 'Ledger', '!', '!', 'May', 'issue', 'best', 'wi
sh', 'family', 'friend', 'daughter', 'year', 'come', '!', '!', 'Remember', '!', '!',
, '!', 'Honor', 'role', 'past', 'role', '!', '!', 'incredible', 'individual', 'talented'
```

실습 2-3. 전체 리뷰 데이터 전처리하기(1)

· 실습 1-3-4에서 수집하였던 영화 '다크 나이트(The Dark Knight)' 리뷰를 불러와 각각의 리뷰(총 100개)를 전처리해 새로운 텍스트 파일에 저장한다

- 토큰화와 stopwords 제거와 lemmatization을 한 후 저장한다



```
D:\Google Drive\2017-1부도시 데이터 사이언스 연구소 교육\실습\src\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 창 관리 ?
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'
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 his 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 his 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 his 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 his 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 his human with no po
43 (Synocesis) 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
```

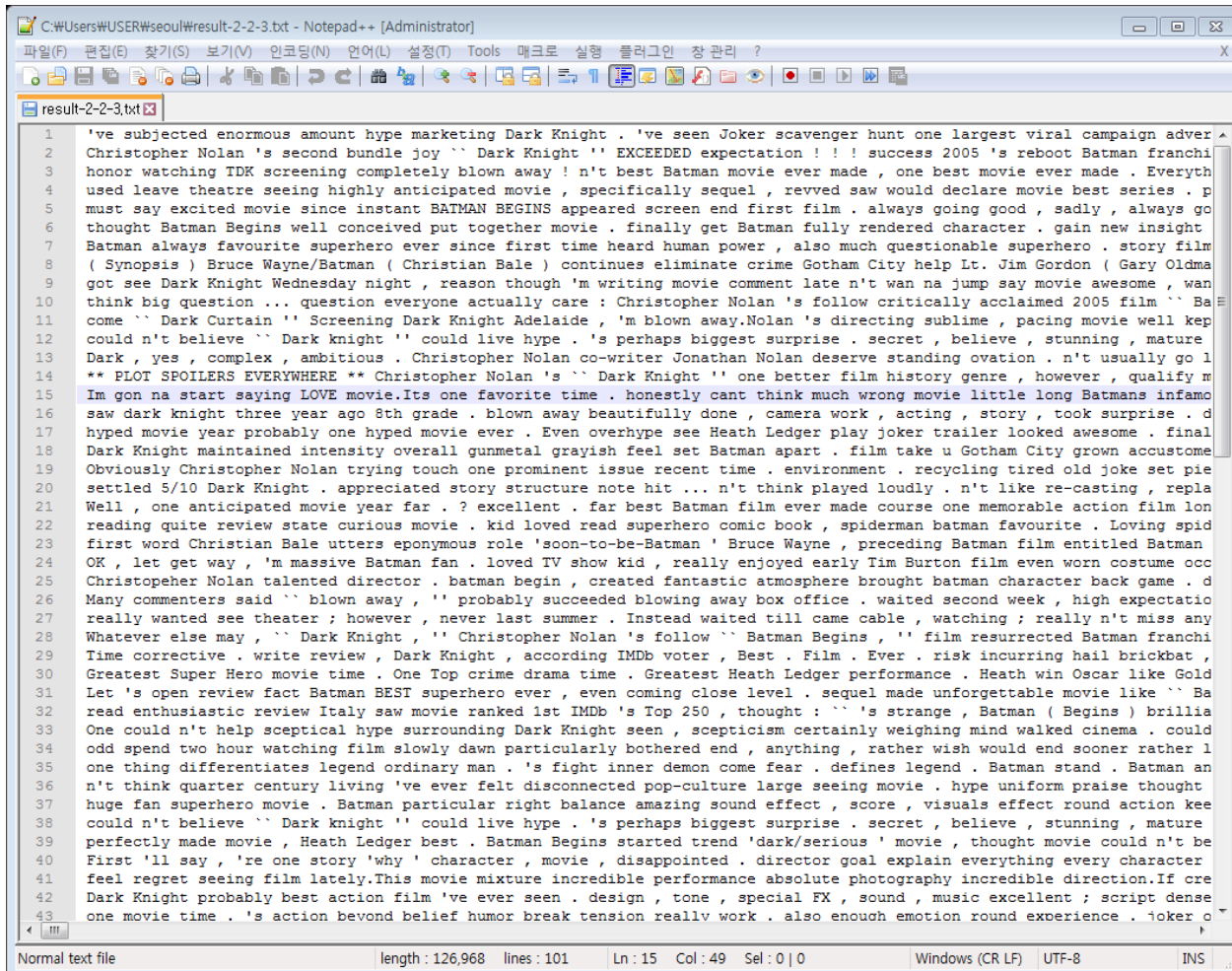
실습 2-3. 전체 리뷰 데이터 전처리하기(1)

- TIPS

- Input 파일(실습 1-3-4의 결과)와 같은 100줄 짜리 텍스트 파일이 결과로 나오지만 각각의 줄을 보면 lemmatize 된 결과가 나와야 한다

실습 2-3. 전체 리뷰 데이터 전처리하기(1)

· 저장 결과



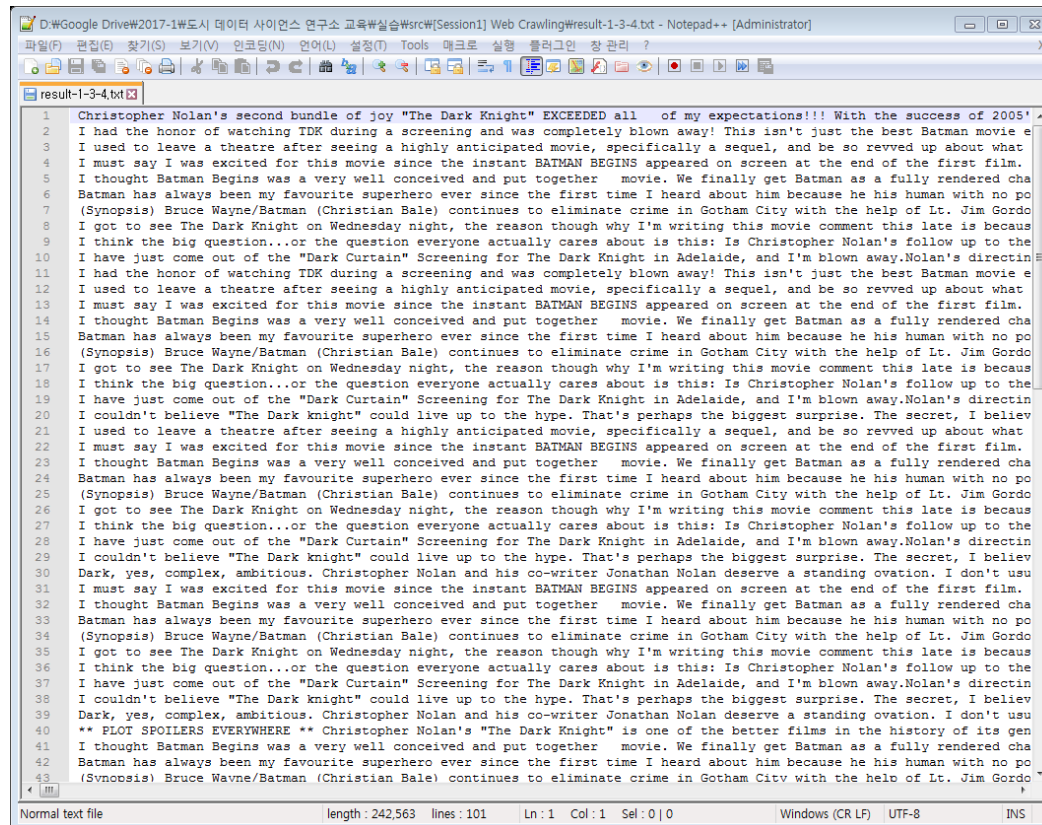
```
C:\Users\USER\seoul\result-2-2-3.txt - Notepad++ [Administrator]
파일(F) 편집(E) 찾기(S) 보기(V) 인코딩(N) 언어(L) 설정(T) Tools 매크로 실행 플러그인 창 관리 ?
result-2-2-3.txt
1 've subjected enormous amount hype marketing Dark Knight . 've seen Joker scavenger hunt one largest viral campaign adver
2 Christopher Nolan 's second bundle joy `` Dark Knight '' EXCEEDED expectation !!! success 2005 's reboot Batman franchi
3 honor watching TDK screening completely blown away ! n't best Batman movie ever made , one best movie ever made . Everyth
4 used leave theatre seeing highly anticipated movie , specifically sequel , revved saw would declare movie best series . p
5 must say excited movie since instant BATMAN BEGINS appeared screen end first film . always going good , sadly , always go
6 thought Batman Begins well conceived put together movie . finally get Batman fully rendered character . gain new insight
7 Batman always favourite superhero ever since first time heard human power , also much questionable superhero . story film
8 ( Synopsis ) Bruce Wayne/Batman ( Christian Bale ) continues eliminate crime Gotham City help Lt. Jim Gordon ( Gary Oldma
9 got see Dark Knight Wednesday night , reason though 'm writing movie comment late n't wan na jump say movie awesome , wan
10 think big question ... question everyone actually care : Christopher Nolan 's follow critically acclaimed 2005 film `` Ba
11 come `` Dark Curtain '' Screening Dark Knight Adelaide , 'm blown away.Nolan 's directing sublime , pacing movie well kep
12 could n't believe `` Dark knight '' could live hype . 's perhaps biggest surprise . secret , believe , stunning , mature
13 Dark , yes , complex , ambitious . Christopher Nolan co-writer Jonathan Nolan deserve standing ovation . n't usually go l
14 ** PLOT SPOILERS EVERYWHERE ** Christopher Nolan 's `` Dark Knight '' one better film history genre , however , qualify m
15 Im gon na start saying LOVE movie.Its one favorite time . honestly cant think much wrong movie little long Batmans infamo
16 saw dark knight three year ago 8th grade . blown away beautifully done , camera work , acting , story , took surprise . d
17 hyped movie year probably one hyped movie ever . Even overhype see Heath Ledger play joker trailer looked awesome . final
18 Dark Knight maintained intensity overall gunmetal grayish feel set Batman apart . film take u Gotham City grown accustom
19 Obviously Christopher Nolan trying touch one prominent issue recent time . environment . recycling tired old joke set pie
20 settled 5/10 Dark Knight . appreciated story structure note hit ... n't think played loudly . n't like re-casting , repla
21 Well , one anticipated movie year far . ? excellent . far best Batman film ever made course one memorable action film lon
22 reading quite review state curious movie . kid loved read superhero comic book , spiderman batman favourite . Loving spid
23 first word Christian Bale utters eponymous role 'soon-to-be-Batman ' Bruce Wayne , preceding Batman film entitled Batman
24 OK , let get way , 'm massive Batman fan . loved TV show kid , really enjoyed early Tim Burton film even worn costume occ
25 Christopher Nolan talented director . batman begin , created fantastic atmosphere brought batman character back game . d
26 Many commenters said `` blown away , '' probably succeeded blowing away box office . waited second week , high expectatio
27 really wanted see theater ; however , never last summer . Instead waited till came cable , watching ; really n't miss any
28 Whatever else may , `` Dark Knight , '' Christopher Nolan 's follow `` Batman Begins , '' film resurrected Batman franchi
29 Time corrective . write review , Dark Knight , according IMDb voter , Best . Film . Ever . risk incurring hail brickbat
30 Greatest Super Hero movie time . One Top crime drama time . Greatest Heath Ledger performance . Heath win Oscar like Gold
31 Let 's open review fact Batman BEST superhero ever , even coming close level . sequel made unforgettable movie like `` Ba
32 read enthusiastic review Italy saw movie ranked 1st IMDb 's Top 250 , thought : `` 's strange , Batman ( Begins ) brillia
33 One could n't help sceptical hype surrounding Dark Knight seen , scepticism certainly weighing mind walked cinema . could
34 odd spend two hour watching film slowly dawn particularly bothered end , anything , rather wish would end sooner rather l
35 one thing differentiates legend ordinary man . 's fight inner demon come fear . defines legend . Batman stand . Batman an
36 n't think quarter century living 've ever felt disconnected pop-culture large seeing movie . hype uniform praise thought
37 huge fan superhero movie . Batman particular right balance amazing sound effect , score , visuals effect round action kee
38 could n't believe `` Dark knight '' could live hype . 's perhaps biggest surprise . secret , believe , stunning , mature
39 perfectly made movie , Heath Ledger best . Batman Begins started trend 'dark/serious ' movie , thought movie could n't be
40 First 'll say , 're one story 'why ' character , movie , disappointed . director goal explain everything every character
41 feel regret seeing film lately.This movie mixture incredible performance absolute photography incredible direction.If cre
42 Dark Knight probably best action film 've ever seen . design , tone , special FX , sound , music excellent ; script dense
43 one movie time . 's action bevond belief humor break tension really work . also enough emotion round experience . joker o
```

Normal text file length: 126,968 lines: 101 Ln: 15 Col: 49 Sel: 0 | 0 Windows (CR LF) UTF-8 INS

실습 2-4. 전체 리뷰 데이터 전처리하기(2)

· 실습 1-3-5에서 수집하였던 영화 리뷰들을 이전 실습과 같이 전처리해 각각 다른 텍스트 파일에 저장해 보자

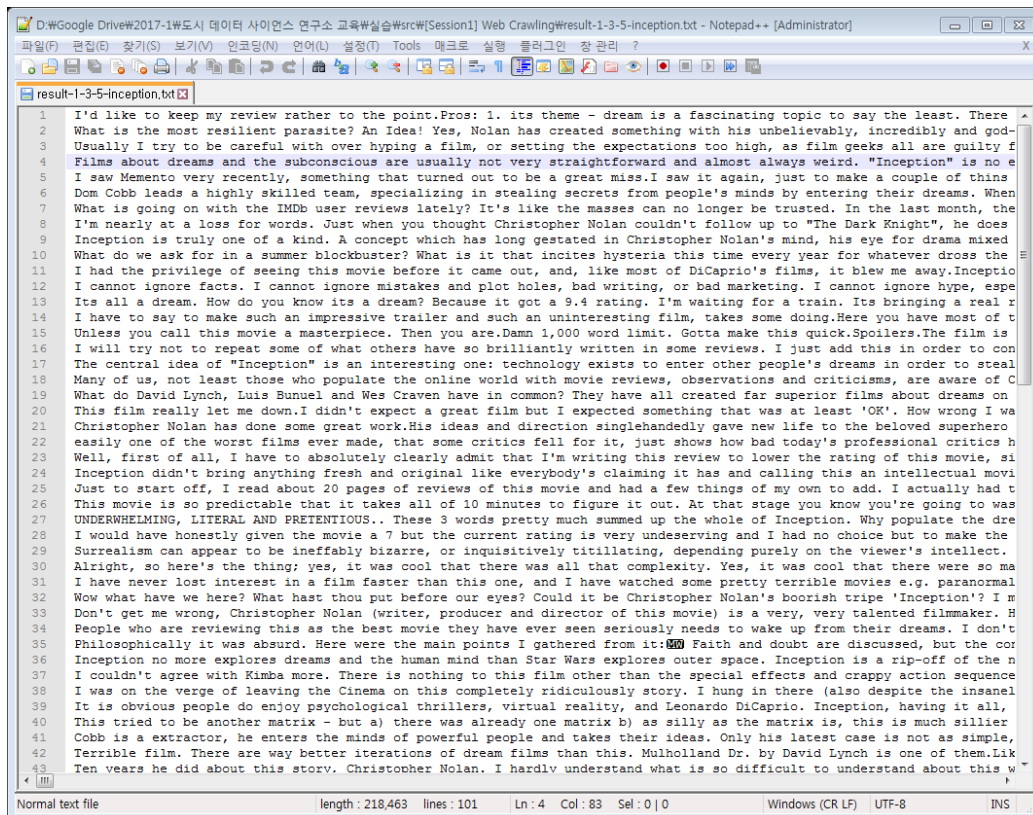
- 토큰화와 stopwords 제거와 lemmatization을 한 후 저장한다



```
D:\Google Drive\2017-1\영화 데이터 사이언스 연구소 교육\실습\src\session1\Web Crawling\result-1-3-4.txt - Notepad++ [Administrator]
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'
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 his 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 his 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 his 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 his 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 his human with no po
43 (Synnoesis) 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 2-4. 전체 리뷰 데이터 전처리하기(2)

· 저장 결과



```
D:\Google Drive\2017-1\도시 데이터 사이언스 연구소 교육\실습\src\Web Crawling\result-1-3-5-inception.txt - Notepad++ [Administrator]
파일(F) 편집(E) 찾기(S) 보기(V) 인코딩(N) 언어(L) 설정(T) Tools 매크로 실행 플러그인 창 관리 ?
result-1-3-5-inception.txt
1 I'd like to keep my review rather to the point. Pros: 1. its theme - dream is a fascinating topic to say the least. There
2 What is the most resilient parasite? An Idea! Yes, Nolan has created something with his unbelievably, incredibly and god-
3 Usually I try to be careful with over hyping a film, or setting the expectations too high, as film geeks all are guilty f
4 Films about dreams and the subconscious are usually not very straightforward and almost always weird. "Inception" is no e
5 I saw Memento very recently, something that turned out to be a great miss. I saw it again, just to make a couple of thins
6 Dom Cobb leads a highly skilled team, specializing in stealing secrets from people's minds by entering their dreams. When
7 What is going on with the IMDB user reviews lately? It's like the masses can no longer be trusted. In the last month, the
8 I'm nearly at a loss for words. Just when you thought Christopher Nolan couldn't follow up to "The Dark Knight", he does
9 Inception is truly one of a kind. A concept which has long gestated in Christopher Nolan's mind, his eye for drama mixed
10 What do we ask for in a summer blockbuster? What is it that incites hysteria this time every year for whatever dross the
11 I had the privilege of seeing this movie before it came out, and, like most of DiCaprio's films, it blew me away. Inceptio
12 I cannot ignore facts. I cannot ignore mistakes and plot holes, bad writing, or bad marketing. I cannot ignore hype, espe
13 Its all a dream. How do you know its a dream? Because it got a 9.4 rating. I'm waiting for a train. Its bringing a real r
14 I have to say to make such an impressive trailer and such an uninteresting film, takes some doing. Here you have most of t
15 Unless you call this movie a masterpiece. Then you are. Damn 1,000 word limit. Gotta make this quick. Spoilers. The film is
16 I will try not to repeat some of what others have so brilliantly written in some reviews. I just add this in order to con
17 The central idea of "Inception" is an interesting one: technology exists to enter other people's dreams in order to steal
18 Many of us, not least those who populate the online world with movie reviews, observations and criticisms, are aware of C
19 What do David Lynch, Luis Bunuel and Wes Craven have in common? They have all created far superior films about dreams on
20 This film really let me down. I didn't expect a great film but I expected something that was at least 'OK'. How wrong I wa
21 Christopher Nolan has done some great work. His ideas and direction singlehandedly gave new life to the beloved superhero
22 easily one of the worst films ever made, that some critics fell for it, just shows how bad today's professional critics h
23 Well, first of all, I have to absolutely clearly admit that I'm writing this review to lower the rating of this movie, si
24 Inception didn't bring anything fresh and original like everybody's claiming it has and calling this an intellectual movi
25 Just to start off, I read about 20 pages of reviews of this movie and had a few things of my own to add. I actually had t
26 This movie is so predictable that it takes all of 10 minutes to figure it out. At that stage you know you're going to was
27 UNDERWHELMING, LITERAL AND PRETENTIOUS.. These 3 words pretty much summed up the whole of Inception. Why populate the dre
28 I would have honestly given the movie a 7 but the current rating is very undeserving and I had no choice but to make the
29 Surrealism can appear to be ineffably bizarre, or inquisitively titillating, depending purely on the viewer's intellect.
30 Alright, so here's the thing; yes, it was cool that there was all that complexity. Yes, it was cool that there were so ma
31 I have never lost interest in a film faster than this one, and I have watched some pretty terrible movies e.g. paranormal
32 Wow what have we here? What hast thou put before our eyes? Could it be Christopher Nolan's boorish tripe 'Inception'? I m
33 Don't get me wrong, Christopher Nolan (writer, producer and director of this movie) is a very, very talented filmmaker. H
34 People who are reviewing this as the best movie they have ever seen seriously needs to wake up from their dreams. I don't
35 Philosophically it was absurd. Here were the main points I gathered from it: a) Faith and doubt are discussed, but the cor
36 Inception no more explores dreams and the human mind than Star Wars explores outer space. Inception is a rip-off of the n
37 I couldn't agree with Kimba more. There is nothing to this film other than the special effects and crappy action sequence
38 I was on the verge of leaving the Cinema on this completely ridiculously story. I hung in there (also despite the insanel
39 It is obvious people do enjoy psychological thrillers, virtual reality, and Leonardo DiCaprio. Inception, having it all,
40 This tried to be another matrix - but a) there was already one matrix b) as silly as the matrix is, this is much sillier
41 Cobb is an extractor, he enters the minds of powerful people and takes their ideas. Only his latest case is not as simple,
42 Terrible film. There are way better iterations of dream films than this. Mulholland Dr. by David Lynch is one of them. Lik
43 Ten years he did about this story. Christopher Nolan. I hardly understand what is so difficult to understand about this w
```

Normal text file length: 218,463 lines: 101 Ln: 4 Col: 83 Sel: 0 | 0 Windows (CR LF) UTF-8 INS

	result-1-3-5-inception	2017-06-22 오후...	텍스트 문서	214KB
	result-1-3-5-old_boy	2017-06-22 오후...	텍스트 문서	155KB
	result-1-3-5-whiplash	2017-06-22 오후...	텍스트 문서	173KB