# **Text Preprocessing**

감성분석

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 2-1-1. 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 2-1-1. 영어 문장 토큰화하기

· 출력 결과

```
. 관리자: C:₩Windows₩system32₩cmd.exe
                                                                       _ 0
C:\Users\USER\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', '.'l
C:\Users\USER\seoul>
```

· 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

#### · 출력 결과

```
한 관리자: C:₩Windows₩system32₩cmd.exe
                                                                       - - X
C:\Users\USER\seoul>python 1-1-2.py
POS tagging Sentence 1:
[('My', 'PRP$'), ('only', 'JJ'), ('regret', 'NN'), ('in', 'IN'), ('life', 'NN'),
('is', 'UBZ'), ('that', 'IN'), ('I', 'PRP'), ('did', 'UBD'), ('not', 'RB'), ('d =
rink', 'UB'), ('more', 'JJR'), ('wine', 'NN'), ('.', '.')]
POS tagging Sentence 2:
[('I'. 'PRP'). ('drink'. 'UBP'). ('to'. 'TO'). ('make'. 'UB'). ('other'. 'JJ').
('people', 'NNS'), ('more', 'RBR'), ('interesting', 'JJ'), ('.', '.')]
POS tagging Sentence 3:
[('An', 'DT'), ('intelligent', 'JJ'), ('man', 'NN'), ('is', 'VBZ'), ('sometimes'
, 'RB'), ('forced', 'UBN'), ('to', 'TO'), ('be', 'UB'), ('drunk', 'UBN'), ('to',
'TO'), ('spend', 'UB'), ('time', 'NN'), ('with', 'IN'), ('his', 'PRP$'), ('fool
s'. 'NNS'). ('.'. ')]
C:\Users\USER\seoul>
```

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

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	

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

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	
ТО	То	To부정사	
UH	Interjection	감탄사	
VB	Verb (Base form)	동사(기본형)	
VBD	Verb (Past tense)	동사(과거형)	
VBN	Verb (Past participle)	동사(과거 분사)	
VBP	Verb (Singular, present)	동사(현재 단수형)	
VBZ	Verb (3 <sup>rd</sup> 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 2-1-3. 단어의 기본형 찾기(Lemmatization)

- · Lemmatization: 단어의 기본형 찾기(source: <a href="https://en.wikipedia.org/wiki/Lemmatisation#Description">https://en.wikipedia.org/wiki/Lemmatisation#Description</a>)
  - 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 2-1-3. 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 2-1-3. 단어의 기본형 찾기(Lemmatization)

· 출력 결과

```
画 관리자: C:₩Windows₩system32₩cmd.exe
                                                                       - © X
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', '.'l
C:\Users\USER\seoul>
```

# 실습 2-1-4. 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 2-1-4. 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."

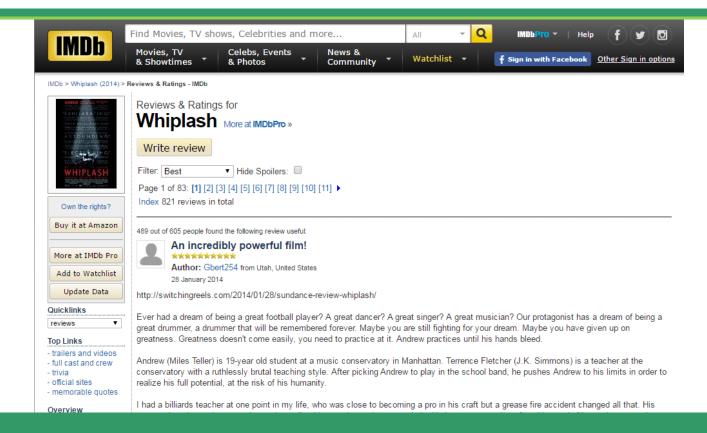
- · TIPS
  - nltk.corpus의 stopwords를 불러온다
  - nltk 패키지의 stopwords list는 모두 소문자로 이루어져 있음을 유의한다

# 실습 2-1-4. Stopwords 제거하기

· 출력 결과

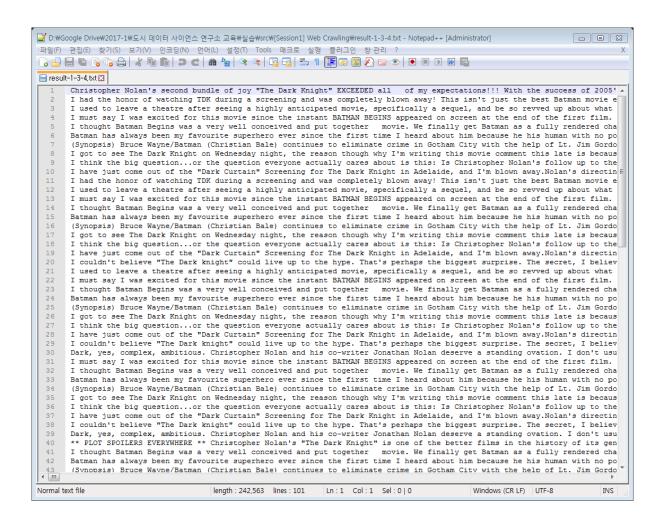
```
. 관리자: C:₩Windows₩system32₩cmd.exe
                                                                      _ @ X
C:\Users\USER\seoul>python 1-1-4.py
Removing stopwords in Sentence 1:
['regret', 'life', 'drink', 'wine', '.']
Removing stopwords in Sentence 2:
['drink', 'make', 'people', 'interesting', '.']
Removing stopwords in Sentence 3:
['intelligent', 'man', 'sometimes', 'forced', 'drunk', 'spend', 'time', 'fools',
2.21
C:\Users\USER\seoul>
```

# 실습 2 - IMDb 리뷰 데이터 전처리하기



# 실습 2-2-1. 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



# 실습 2-2-1. 첫 번째 리뷰 전처리하기 (1)

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

### 실습 2-2-1. 첫 번째 리뷰 전처리하기 (1)

#### · 출력 결과

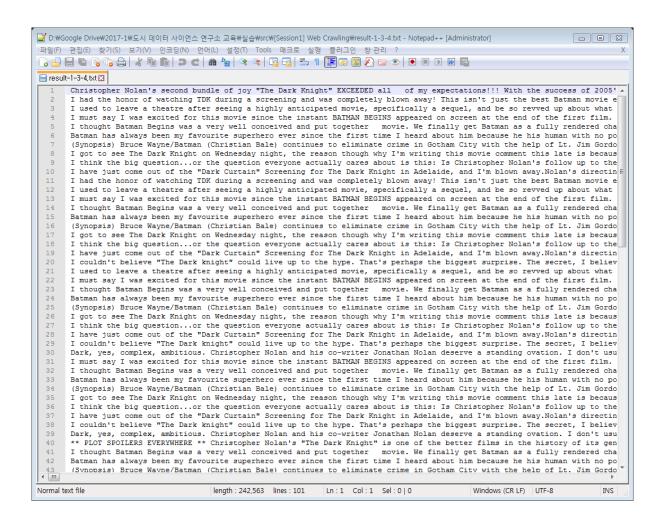
```
● 관리자: C:₩Windows₩system32₩cmd.exe
C:\Users\USER\seoul>python 2-2-1.py
Tokens for first review:
['Christopher', 'Nolan', '''s", 'second', 'bundle', 'of', 'joy', '``', 'The', 'Da
rk', 'Knight', "''", 'EXCEEDED', 'all', 'of', 'my', 'expectations', '!', '!', '!
', 'With', 'the', 'success', 'of', '2005', "'s", 'reboot', 'of', 'the', 'Batman'
'franchise', ',', 'they', 'took', 'what', 'was', 'already', 'established', 'an
d', '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', 'de
eper', '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', 'f
rom', 'his', 'fractured', 'soul', 'due', 'to', 'the', 'murder', 'of', 'his', 'be
loved', 'parents', '.', 'But', 'with', 'the', 'help', 'of', 'his', 'trusted', 'k
utler/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', '
s', 'Heath', 'Ledger', 'as', 'Batman', '''s", 'most', 'deadly', 'enemy', ',', 'Th
   'Joker', '.', 'I', 'can', 'HONESTLY', 'tell', 'you', 'that', ':', 'as', 'goo
d', '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', 'dis
turbing', 'than', 'the', 'previous', 'incarnation', 'of', 'The', 'Clown', 'Princ
e', 'of', 'Crime', 'and', 'Ledger', 'gives', 'it', 'his', 'all', 'to', 'do', 'hi
m', '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', 'Holm
es', ',', 'Gyllenhaal', 'gives', 'a', 'much', 'better', 'performance', 'and', 'i
s', 'a', 'far', 'cry', 'from', 'the', '``', 'damsel-in-distr<u>ess', "''", 'stereot</u>
ype', '(', 'though', 'there', "'s", 'a', 'little', 'of', 'it', ',', 'THANKFULLY
 ')', 'that', "'s", 'common', 'in', 'films', '.', 'Bale', 'and', 'Gyllenhall',
'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', 'crush
ing', 'combat', 'than', '``', 'Begins', "''", 'in', 'addition', 'to', 'new', 'te
chnology', 'at', 'Batman', '''s", 'disposal', '.', 'Also', 'worth', 'mentioning',
'is', 'screenwriter', 'Jonathan', 'Nolan', ',', 'who', 'gives', 'the', 'film',
```

```
- P X
● 관리자: C:₩Windows₩system32₩cmd.exe
POS tags for first review:
[('Chr<sup>i</sup>stopher', 'NNP'), ('Nolan', 'NNP'), ("'s", 'POS'), ('second', 'JJ'), ('bu
ndle', 'NN'), ('of', 'IN'), ('joy', 'NN'), ('``', '``'), ('The', 'DT'), ('Dark',
'NNP'), ('Knight', 'NNP'), ("''", "''"), ('EXCEEDED', 'NNP'), ('all', 'DT'), ('
of', 'IN'), ('my', 'PRP$'), ('expectations', 'NNS'), ('!', '.'), ('!', '.'), ('!
 , '.'), ('With', 'IN'), ('the', 'DT'), ('success', 'NN'), ('of', 'IN'), ('2005'
 'CD'), ("'s", 'POS'), ('reboot', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Batman'
  'NNP'), ('franchise', 'NN'), (',', ','), ('they', 'PRP'), ('took', 'UBD'), ('w
hat', 'WP'), ('was', 'UBD'), ('already', 'RB'), ('established', 'UBN'), ('and',
'CC'), ('expanded', 'UBN'), ('it', 'PRP'), (',', ','), ('amped', 'UBD'), ('it',
'PRP'), ('up', 'RP'), (',', ','), ('and', 'CC'), ('gave', 'VBD'), ('a', 'DT'), (
'deeper', 'NN'), (',', ','), ('darker', 'NN'), ('and', 'CC'), ('brooding', 'UBG'
), ('story', 'NN'), ('that', 'WDT'), ('is', 'UBZ'), ('more', 'JJR'), ('gripping
 'JJ'), ('and', 'CC'), ('the', 'DT'), ('suspense', 'NN'), ('is', 'UBZ'), ('like
ly', 'JJ'), ('to', 'TO'), ('catch', 'UB'), ('you', 'PRP'), ('of', 'IN'), ('guard
 , 'JJ'), ('several', 'JJ'), ('times', 'NNS'), ('throughout', 'IN'), ('.', '.')
('Christian', 'JJ'), ('Bale', 'NNP'), ('delves', 'UBZ'), ('more', 'JJR'), ('dee
per', 'JJ'), ('into', 'IN'), ('Batman', 'NNP'), (',', ','), ('sworn', 'UBN'), ('
to', 'TO'), ('fight', 'UB'), ('evil', 'NN'), ('and', 'CC'), ('injustice', 'NN'),
(',', ','), ('though', 'RB'), ('also', 'RB'), ('quite', 'RB'), ('reluctant', 'J
J'), ('and', 'CC'), ('uncertain', 'JJ'), ('if', 'IN'), ('his', 'PRP$'), ('crusad
 ', 'NN'), ('can', 'MD'), ('ever', 'RB'), ('end', 'UB'), ('and', 'CC'), ('cleans
e', 'VB'), ('his', 'PRP$'), ('inner', 'JJ'), ('turmoil', 'NN'), ('from', 'IN'),
('his', 'PRP$'), ('fractured', 'JJ'), ('soul', 'NN'), ('due', 'JJ'), ('to', 'T0'
), ('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', 'UBN'), ('butler/
ally', 'RB'), ('Alfred', 'NNP'), ('(', '('), ('played', 'UBN'), ('superbly', 'RB
'), ('by', 'IN'), ('Michael', 'NNP'), ('Cane', 'NNP'), (')', ')'), ('grounds',
UBZ'), ('him', 'PRP'), (',', ','), ('gives', 'UBZ'), ('him', 'PRP'), ('moral',
JJ'), ('support', 'NN'), (',', ','), ('and', 'CC'), ('keeps', 'UBZ'), ('him',
RP'), ('in', 'IN'), ('check', 'NN'), ('.', '.'), ('But', 'GC'), ('the', 'DT'),
'real', 'JJ'), ('star', 'NN'), ('of', 'IN'), ('the', 'DT'), ('show', 'NN'), ('is
 , 'UBZ'), ('Heath', 'NNP'), ('Ledger', 'NNP'), ('as', 'IN'), ('Batman', 'NNP'),
("'s", 'POS'), ('most', 'JJS'), ('deadly', 'RB'), ('enemy', 'NN'), (',', ','),
('The', 'DT'), ('Joker', 'NNP'), ('.', '.'), ('I', 'PRP'), ('can', 'MD'), ('HONE
STLY', 'UB'), ('tell', 'UB'), ('you', 'PRP'), ('that', 'WDT'), (':', ':'), ('as'
, 'RB'), ('good', 'JJ'), ('as', 'IN'), ('Jack', 'NNP'), ('Nicholson', 'NNP'), ('
```

<토큰> <품사 태그>

# 실습 2-2-2. 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 lemmtization을 수행해 그 결과를 출력한다



# 실습 2-2-2. 첫 번째 리뷰 전처리하기 (2)

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

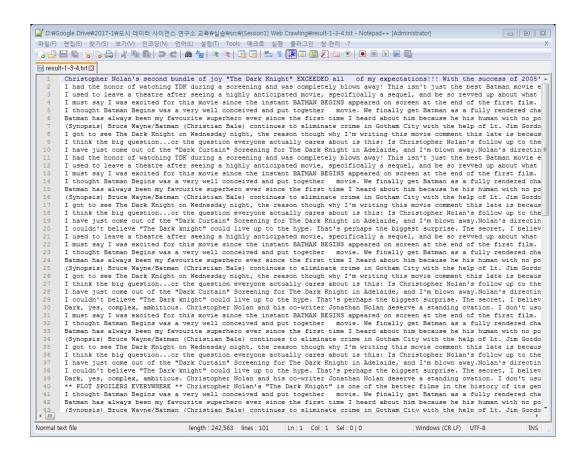
# 실습 2-2-2. 첫 번째 리뷰 전처리하기 (2)

#### · 출력 결과

```
관리자: C:₩Windows₩system32₩cmd.exe
                                                                      - O X
C:\Users\USER\seoul>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
  , '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', "'S", 'PLAY', 'compared', 'G
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', '''<u>", 'stereotype', '('</u>
'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', 'I
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 2-2-3. 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



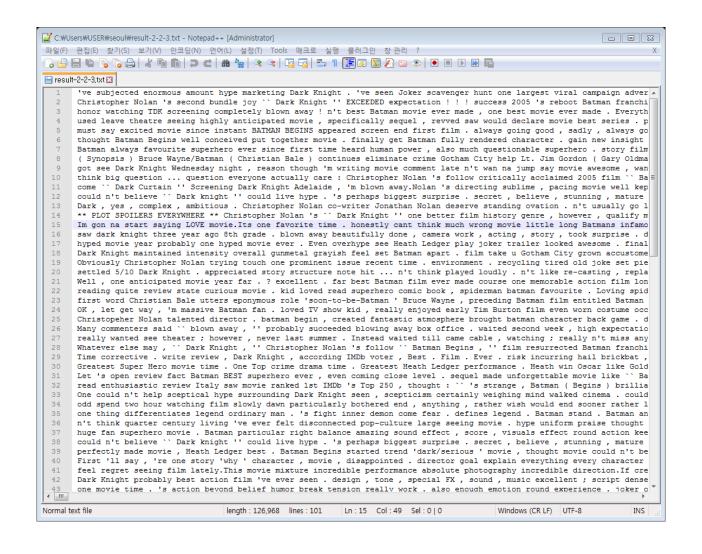
# 실습 2-2-3. 전체 리뷰 데이터 전처리하기(1)

#### · TIPS

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

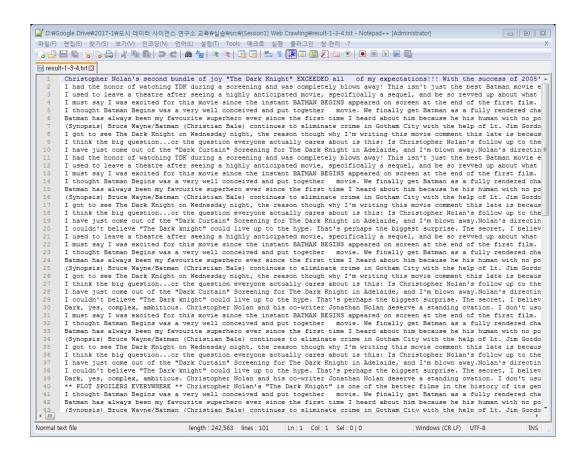
### 실습 2-2-3. 전체 리뷰 데이터 전처리하기(1)

#### ㆍ 저장 결과



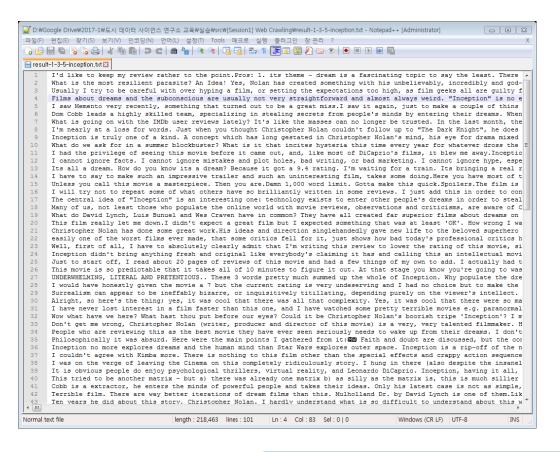
### 실습 2-2-4. 전체 리뷰 데이터 전처리하기(2)

- · 실습 1-3-5에서 수집하였던 영화 리뷰들을 이전 실습과 같이 전처리해 각각 다른 텍스트 파일에 저장해 보자
  - 토큰화와 stopwords 제거와 lemmatization을 한 후 저장한다



# 실습 2-2-4. 전체 리뷰 데이터 전처리하기(2)

#### ㆍ 저장 결과



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