H.S. Kim
Christina

Natural Language Processing

- 영어 텍스트 분석 및 시각화

영어 분석 처리를 위한 환경 셋팅 (nltk 패키지 설치)

1. nltk(Natural Language Took Kit) 패키지를 설치

필요시 conda를 최신 버전으로 업데이트 후...

```
Microsoft Windows [Version 10.0.17763.1518]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\tina>conda install nltk
Collecting package metadata (current_repodata.json): done
Solving environment: done
## Package Plan ##
  environment location: C:\Users\tina\anaconda3
  added / updated specs:
    - nltk
The following packages will be downloaded:
    package                    |            build
    ---------------------------|-----------------
    conda-4.9.1                |   py38haa95532_0       2.9 MB
    ------------------------------------------------------------
                                           Total:       2.9 MB
The following packages will be UPDATED:
  conda                                4.8.5-py38_0 --> 4.9.1-py38haa95532_0
Proceed ([y]/n)? y
Downloading and Extracting Packages
conda-4.9.1          | 2.9 MB    |
################################################################################ | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
C:\Users\tina>
```

영어 텍스트 분석 및 시각화 (기본 문법 실습)

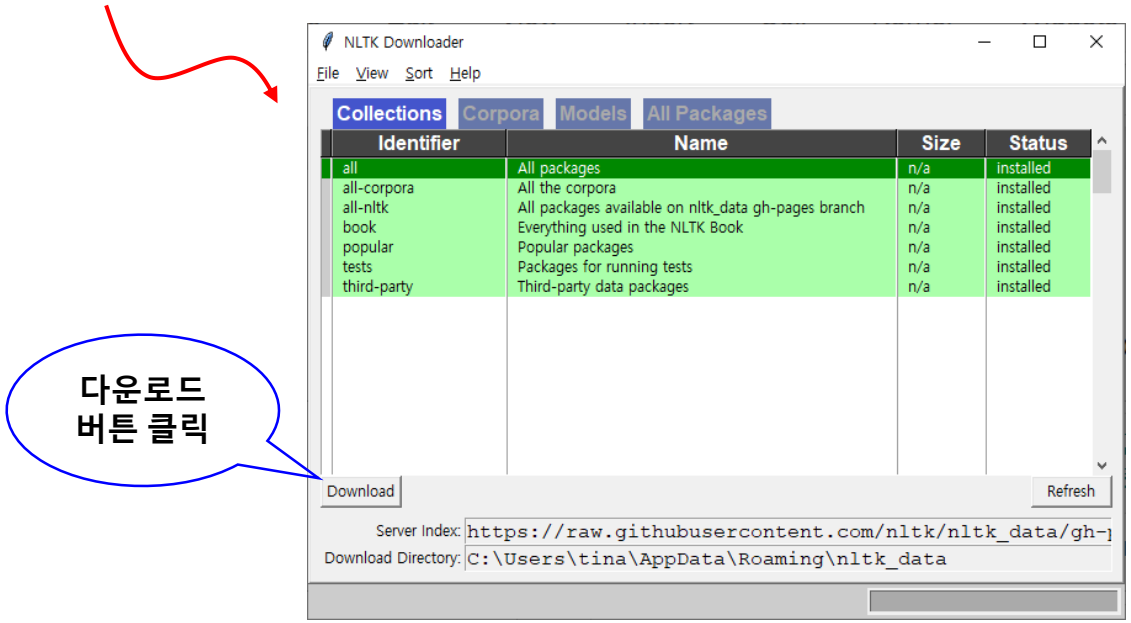
실습 노트 참고

- 12_ Natural Language Processing - 영어 분석 (1).ipynb

nltk 모듈 활용

nltk(Natural Language Took Kit) 모듈을 임포트하고, nltk.download( )메소드를 호출함

```
1  import nltk
2  nltk.download( )
```



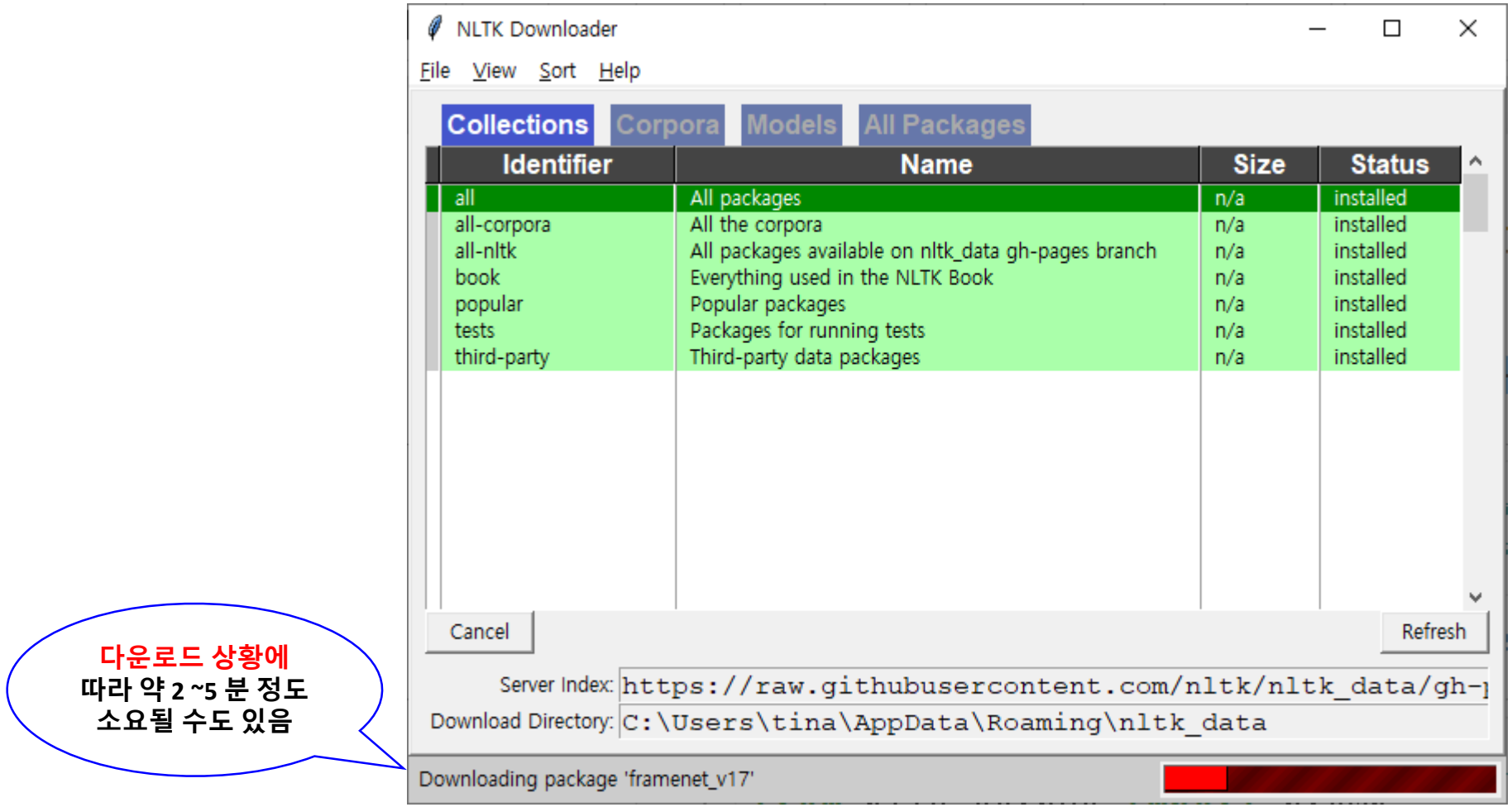
nltk 모듈 활용

자연어 처리

- 영어 텍스트 분석 및 시각화

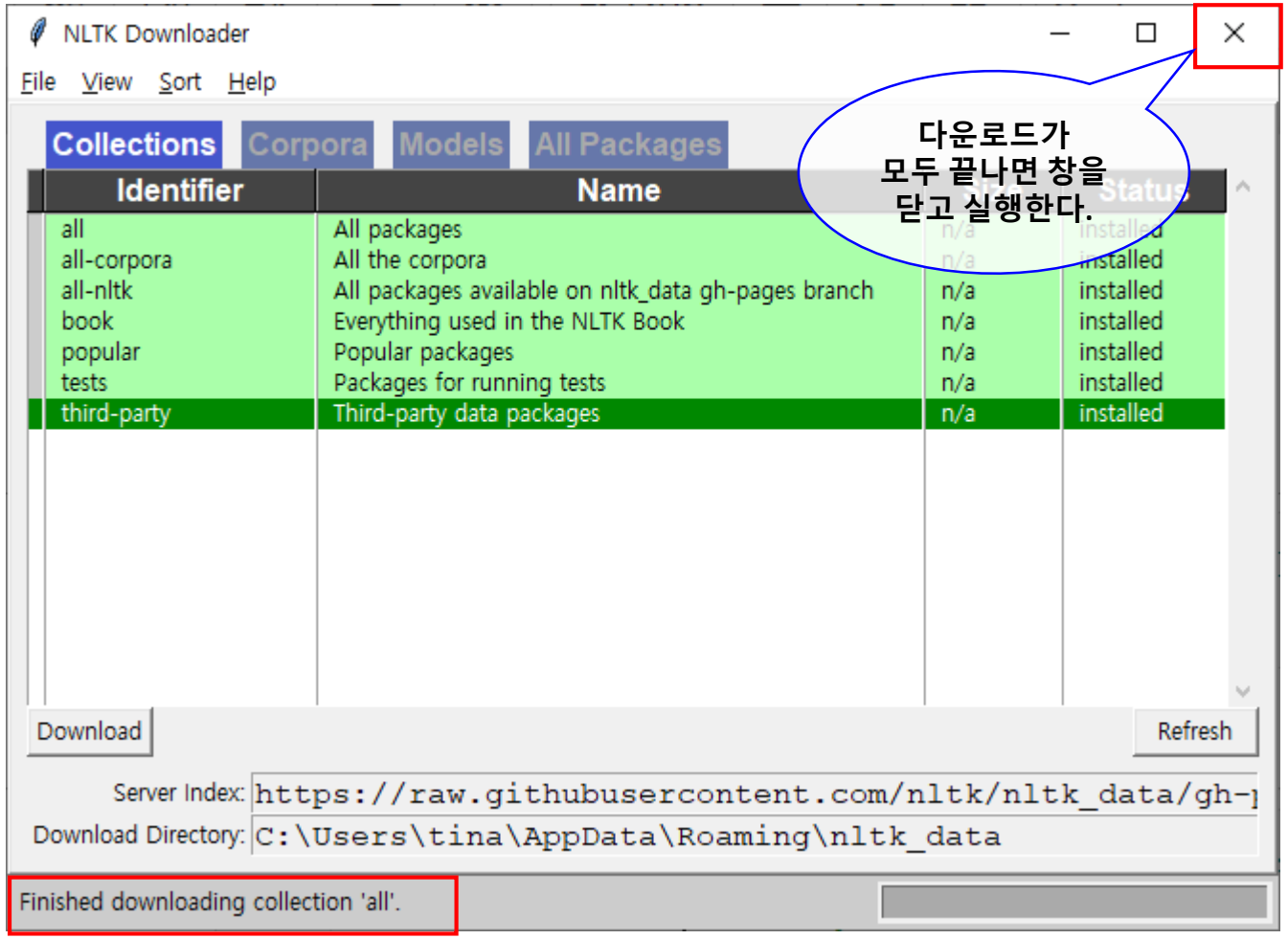
1. 영어 분석을 위한 nltk 패키지 설치
2. 영어 텍스트 분석 및 시각화 (기본 문법 실습)
3. **nltk 모듈 활용**
4. nltk 소개
5. WordNet 소개
6. 문장 토큰화
7. 문장 태깅
8. 펜 트리뱅크 태그셋
9. 영어 텍스트 분석 및 시각화 (트럼프 연설문을 활용한 실습)

- Exercise



다운로드 상황에 따라 약 2 ~5 분 정도 소요될 수도 있음

nltk 모듈 활용



nltk 소개

http://www.nltk.org/

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

- NLTK는 인간 언어 데이터로 작업 할 Python 프로그램을 빌드하기위한 선도적인 플랫폼
- 분류, **토큰화**, 형태소 분석, **태깅**, 파싱 및 의미론적 추론을위한 텍스트 처리 라이브러리와 50 개가 넘는 corpora(말뭉치) 및 WordNet(워드넷)과 같은 어휘 자원에 대한 사용하기 쉬운 인터페이스를 제공

WordNet 소개

https://ko.wikipedia.org/wiki/%EC%9B%8C%EB%93%9C%EB%84%B7

워드넷(WordNet)은 영어의 의미 어휘목록이다. 워드넷은 영어 단어를 'synset'이라는 유의어 집단으로 분류하여 간략하고 일반적인 정의를 제공하고, 이러한 어휘목록 사이의 다양한 의미 관계를 기록한다. 그 목적은 두가지이다. 하나는 사전(단어집)과 시소러스(유의어·반의어 사전)의 배합을 만들어, 보다 직관적으로 사용할 수 있고 자동화된 본문 분석과 인공 지능 응용을 뒷받침하려는 것이다.

데이터베이스와 프로그램 툴은 BSD 형태의 라이선스로 배포되었고, 다운로드 받아 자유롭게 사용할 수 있다. 데이터베이스는 온라인으로도 검색할 수 있다.

워드넷은 심리학 교수인 조지 A. 밀러가 지도하는 프린스턴 대학의 인지 과학 연구소에 의해 만들어졌고 유지되고 있다. 개발은 1985년에 시작되었다. 수 년에 걸쳐, 프로젝트는 3백만 달러의 기금을 모았는데, 주로 기계 번역에 관심이 있는 정부 기관에 의한 것이었다. 최근 몇 년간은, 크리스티안 펠바움(Christiane Fellbaum) 박사가 워드넷의 개발을 살피고 있다.

문장 토큰화(tokenization)

http://www.nltk.org/

Tokenize and tag some text

```
1 sentence = """At eight o'clock on Thursday morning
2 ... Arthur didn't feel very good."""
```

```
1 sentence
```

```
"At eight o'clock on Thursday morning₩nArthur didn't feel very good."
```

```
1 type(sentence)
```

```
str
```

```
1 tokens = nltk.word_tokenize(sentence)
```

```
1 print(tokens)
```

```
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning', 'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
```

문장 태깅(tagging)

http://www.nltk.org/

자연어 처리

- 영어 텍스트 분석 및 시각화

1. 영어 분석을 위한 nltk 패키지 설치
2. 영어 텍스트 분석 및 시각화 (기본 문법 실습)
3. nltk 모듈 활용
4. nltk 소개
5. WordNet 소개
6. 문장 토큰화
7. **문장 태깅**
8. 펜 트리뱅크 태그셋
9. 영어 텍스트 분석 및 시각화 (트럼프 연설문을 활용한 실습)

- Exercise

Tokenize and tag some text

```
1 tagged = nltk.pos_tag(tokens)  #토큰화되어있는 데이터를 태깅 처리
```

```
1 print(tagged)
```

```
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'NN'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'),
('Arthur', 'NNP'), ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.',
'.')]
```

```
1 tagged[0:6]
```

```
[('At', 'IN'),
 ('eight', 'CD'),
 ("o'clock", 'NN'),
 ('on', 'IN'),
 ('Thursday', 'NNP'),
 ('morning', 'NN')]
```

펜 트리뱅크 태그셋(Penn Treebank Tagset)을 활용한 태깅

https://sites.google.com/site/partofspeechhelp/

https://bluebreeze.co.kr/1357

펜 트리뱅크 태그세트(Penn Treebank Tagset)

NKTK는 텍스트에 태그를 붙일 때 널리 쓰이는, 펜실베니아 대학의 펜 트리뱅크를 기본적으로 사용한다.

태그	원어	한국어
CC	coordinating conjunction	등위 접속사(and, or, but 같은 접속사)
CD	cardinal number	기수(순서의 뜸가 없이 수량만 나타내는 수)
DT	determiner	한정사(명사 앞에 붙는 the, some, my 같은 말들)
EX	existential "there"	장소가 아니라 존재는 나타내는 there(There is always some madness in love)
FW	foreign word	외래어
IN	perposition, subordinating conjunction	전치사 종속 접속사
JJ	adjective	형용사
JJR	adjective, comparative	비교급 형용사(My house is larger than hers.)
JJS	adjective, superlative	최상급 형용사(My house is the largest one in our neighborhood.)
LS	list item marker	목록임을 나타내는 문자
MD	modal	법조동사(can, must, may 등)
NN	noun, singular or mass	명사, 단수 또는 복수

펜 트리뱅크 태그셋(Penn Treebank Tagset)을 활용한 태깅

https://sites.google.com/site/partofspeechhelp/

https://bluebreeze.co.kr/1357

펜 트리뱅크 태그세트(Penn Treebank Tagset)

NNS	noun, plural	복수형 명사
NNP	proper noun, singular	단수형 고유명사
NNPS	proper noun, pluar	복수형 고유명사
PDT	predeterminer	선행 한정사(all, both, half 등)
POS	possessive ending	소유격 문자(어포스트로피 및 's)
PRP	personal pronoun	인칭 대명사(I, you, he, she)
PRP$	possessive pronoun	소유격 대명사(The dog is mine)
RB	adberb	부사
RBR	adberb, comparative	비교급 부사(Jim works harder than his brother.)
RBS	adberb, superlative	최상급 부사(Everyoun in the race ran fast, but John ran the fasters of all.)
RP	Particle	불변화사(동사와 함께 쓰이는 부사나 전치사, She tore up the letter.)
SYM	symbol	기호
to	"to"	to
UH	iInterjection	감탄사
VB	verb, base form	동사 원형

펜 트리뱅크 태그셋(Penn Treebank Tagset)을 활용한 태깅

https://sites.google.com/site/partofspeechhelp/

https://bluebreeze.co.kr/1357

펜 트리뱅크 태그세트(Penn Treebank Tagset)

VBD	verb, past tense	과거형 동사
VBG	verb, gerund or present	동명사 또는 현재진행형(~ing)
VBN	verb, past participle	과거분사(I have seen six deer.)
VBP	verb, non-third person singular present	3인칭이 아닌 현재형 동사
VBZ	verb, third person singular persent	3인칭 현재형 동사(s로 끝남)
WDT	wh-determiner	wh로 시작하는 한정사(문장 맨 앞에 등장하지 않는 what, shich)
WP	wh-pronoun	wh로 시작하는 대명사(what, which, who, whoever)
WP$	possessive wh-pronoun	wh로 시작하는 소유격 대명사(whom, whose)
WRP	wh-adverb	wh로 시작하는 부사(when, where, why, how)

영어 텍스트 분석 및 시각화 (트럼프 연설문을 활용한 실습)

실습 노트 참고

- 12_ Natural Language Processing - 영어 분석 - 트럼프 연설문 (2)

Exercise

1) **웹 텍스트 크롤러**를 기반으로 관심있는 분야의 자연어(한글 **or** 영어)를 수집한 후,
주요 **키워드**를 추출하고 **빈도수**를 출력한다.

2) 위의 출력 결과를 딕셔너리 타입의 데이터로 완성한 후,
주요 키워드의 빈도수를 **그래프**와 **워드크라우드**로 시각화하시오.