NLP(Natural Language Processing)

영어 처리

Dr. Heesuk Kim

nltk 소개

http://www.nltk.org/

nltk(Natural Language Took Kit)

NLTK

Search

NLTK Documentation
API Reference
Example Usage
Module Index
Wiki
FAQ

Installation
Installing NLTK
Installing NLTK Data

More
Release Notes
Contributing to NLTK
NLTK Team

Documentation

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**.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at **https://www.nltk.org/book_1ed**.)

nltk 소개

http://www.nltk.org/

- NLTK는 자연어 데이터 처리를 위한 Python 기반의 선도적인 플랫폼
- 무료 오픈 소스 커뮤니티
- Classification, Tokenization, Stemming, Tagging, Parsing, and Semantic Reasoning을 위한 다양한 텍스트 처리 라이브러리와 corpora(말뭉치) 및 WordNet(워드넷)과 같은 어휘 자원 접근을 위한 쉬운 인터페이스 제공
- 언어학자, 엔지니어, 학생, 교육자, 연구원 모두에게 적합한 패키지

WordNet 소개

https://wordnet.princeton.edu/

- WordNet(워드넷)은 무료로 공개적인 다운로드가 가능한 대규모 영어 어휘 데이터베이스이다.
- 심리학 교수인 '조지 아미티지 밀러(George Armitage Miller)'가 지도하는 프린스턴 대학의 인지 과학 연구소에 의해 구축되어 유지되고 있다.
- 영어 단어를 'synsets' 이라는 유의어 집단으로 분류하여 간략하고 일반적인 정의를 제공한다.
- 유의어 사전에 해당하는 '시소러스(thesaurus)'의 배합을 만들어 어휘 목록 사이의 다양한 의미 관계를 기록한다.
- 자동화된 자연어 분석과 인공 지능 응용을 뒷받침할 수 있다.

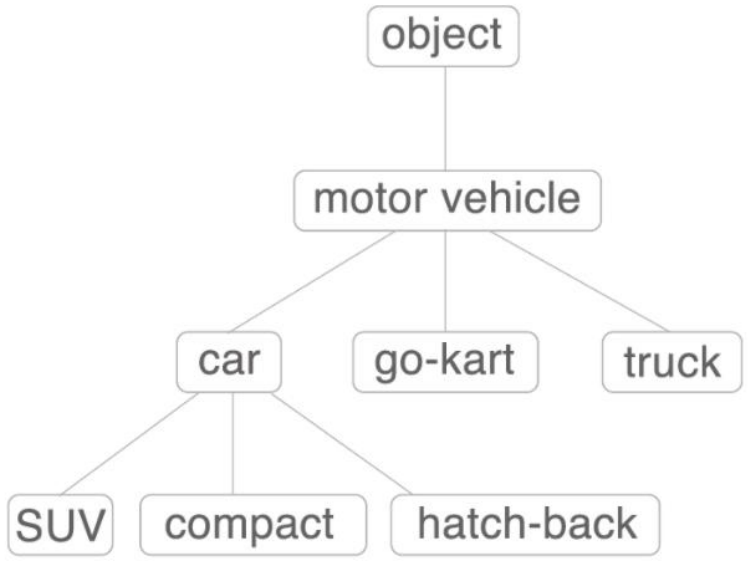
WordNet 소개

https://wordnet.princeton.edu/

유의어

car = auto automobile machine motorcar

단어의 상위 및 하위 관계를 네트워크로 표현



- 자연어 처리 분야에서 가장 유명한 '시소러스(thesaurus)' 는 WordNet이다.
- WordNet을 사용하면 유의어를 얻거나 '단어 네트워크'를 이용할 수 있다.

WordNet 소개

<참고>

한국어를 위한 워드넷

- KorLex

  부산대학교 http://korlex.pusan.ac.kr/

- Korean WordNet(KWN)

  KAIST http://wordnet.kaist.ac.kr/

nltk 설치

nltk(Natural Language Took Kit) 패키지 설치

```
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>
```

필요시 conda를 최신 버전으로 업데이트 후...

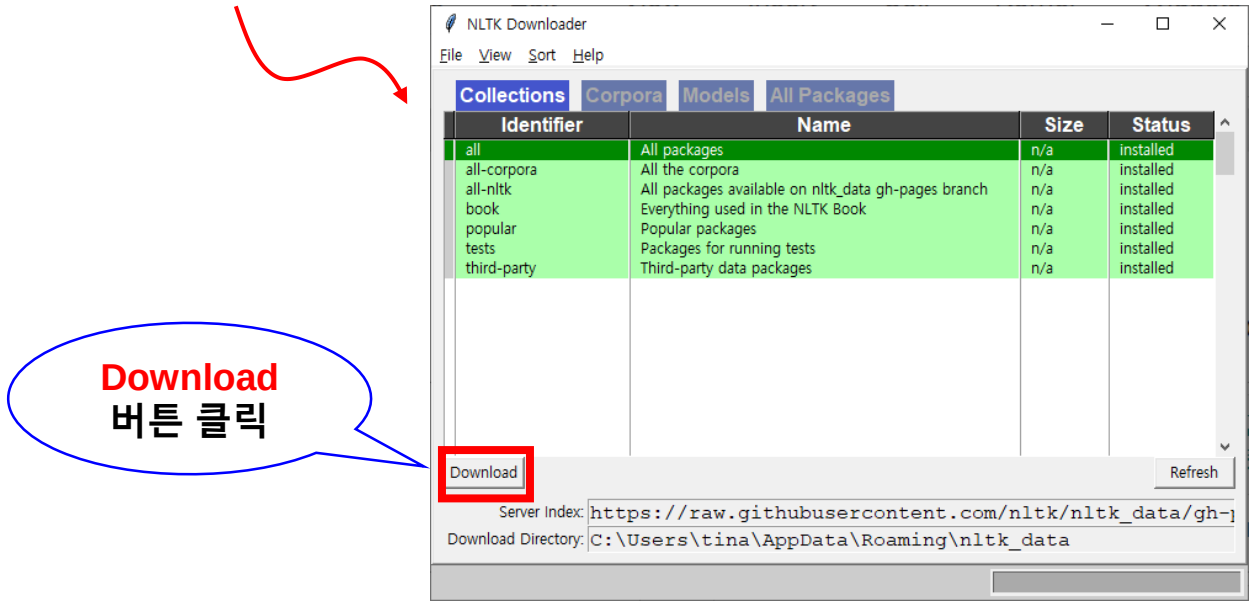
nltk 기본 문법 실습

실습 노트 참고

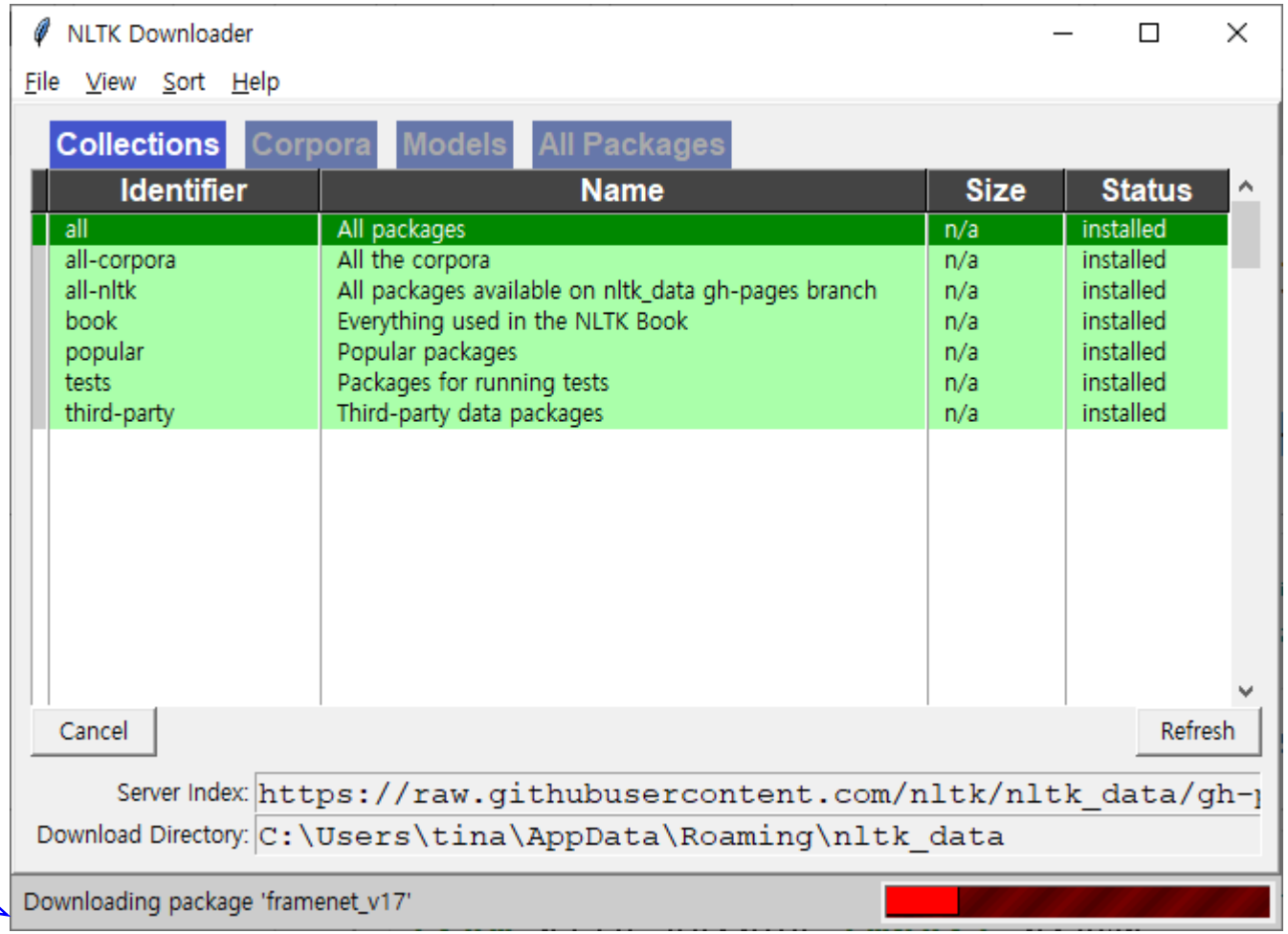
- 11_(8 page) 강의용 NLP_영어 분석용 기본 함수 활용법.ipynb

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

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

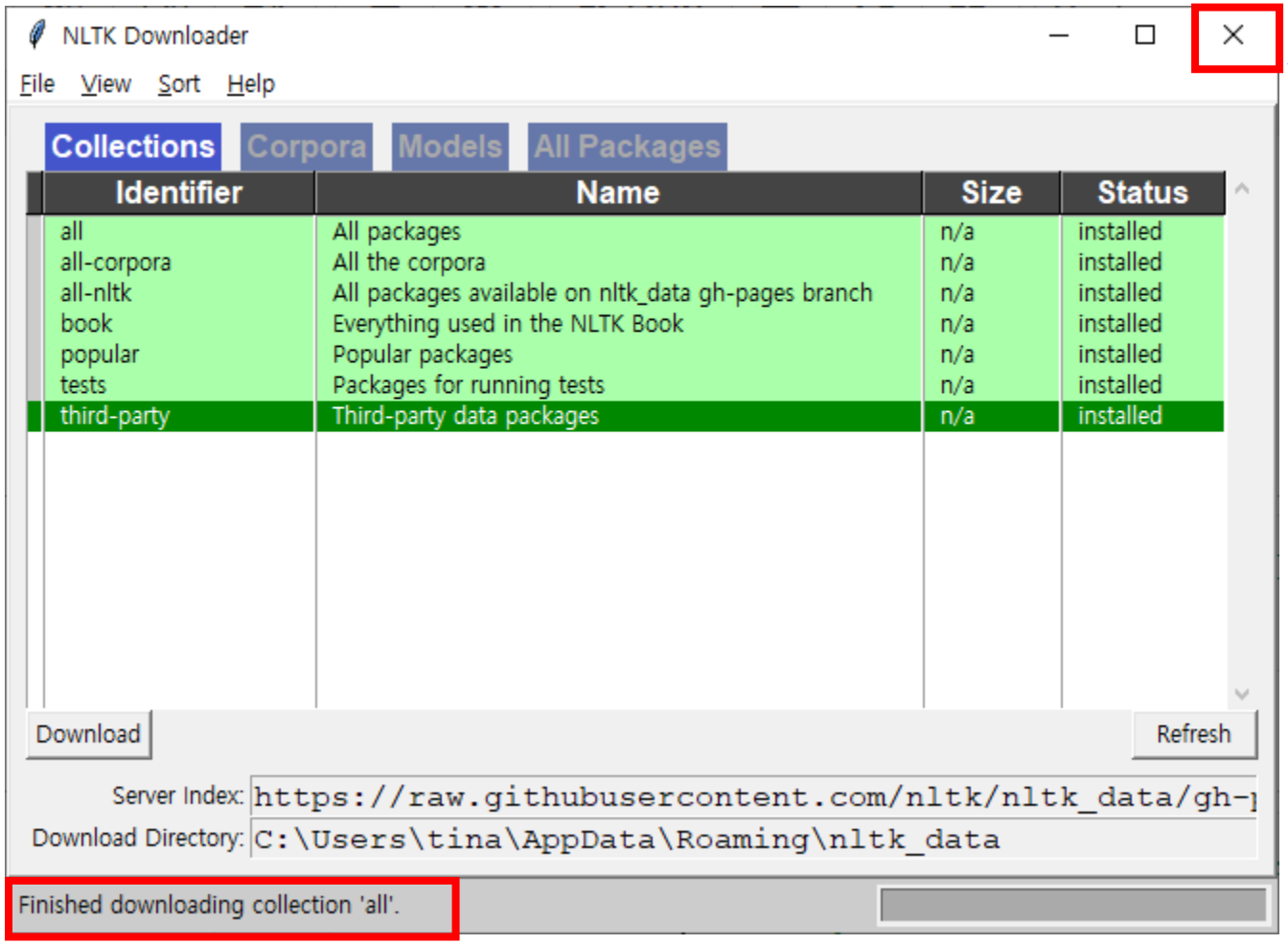


nltk 기본 문법 실습



Download 상황에 따라 약 2 ~5 분 정도 소요될 수도 있음

nltk 기본 문법 실습



다운로드가
모두 끝나면
창을 닫고 실행한다.

문장 토큰화(tokenization)

Coding

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)

Coding

http://www.nltk.org/

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')]
```

Penn Treebank 기반의 영어 태그 확인

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

https://bluebreeze.co.kr/1357

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 읨가 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

Penn Treebank 기반의 영어 태그 확인

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

https://bluebreeze.co.kr/1357

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

Penn Treebank 기반의 영어 태그 확인

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

https://bluebreeze.co.kr/1357

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)

참고

- Biden 대통령 취임사 분석 결과

참고

- Trump 대통령 취임사 분석.ipynb

영어 텍스트 분석 및 시각화 (트럼프 연설문을 활용한 실습)

Nation (Biden 국가관?)

- 보통 국가를 떠올렸을 때 마음속에서 나타나는 개념이다.
- **역사적, 문화적, 언어적, 영토상**의 공유성을 가지는 사람들의 집합이다.
- **신문, 저술, 문헌 등에서 일반적인 국가**를 의미하게 된다.

Country (Trump 국가관?)

- **정치지리학적** 그리고 **법적인 행정구역상** 의미를 가진다.
- 신생 독립국의 국가의 의미는 바로 country 로 구별되어진다.
- **정부의 행정력이 도달하는 범위로 법적의 경계선**이라 할 수 있다.

참고)
미국의 작가 윌리엄 멕과이어 "빌" 브라이슨(William McGuire "Bill" Bryson, 1951년 12월 8일 ~)에 의하면,
“미국은 가장 부유한 국가(nations), 가장 큰 나라(countries) 중 하나”라는 표현을 사용했다.