

NLTK Tutorial

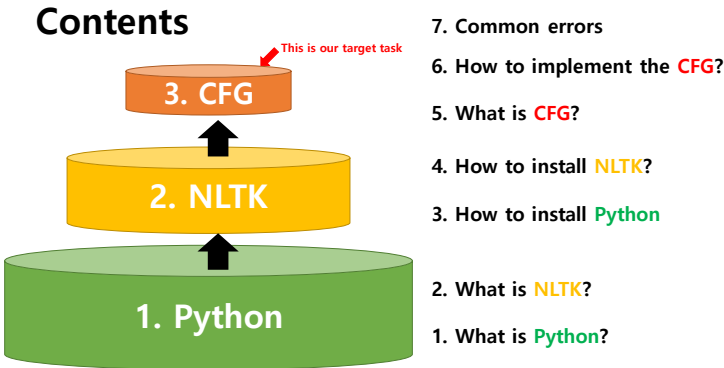
Ki Woong Moon

kiwoongmoon@hufs.ac.kr

Hee Jeong Na

920313@naver.com

Computer & Linguistics, Spring 2021



1. What is Python?

- High-level Programming Language
- From 1991 by Guido van Rossum (named after the comedy show)
- Useful tool for language processing
- Python Library : The package of functions, corpora and variables
ex) NLTK, NUMPY, PANDAS, MATPLOTT ...

3

2. What is NLTK?

- Natural Language ToolKit (Python Library)
 - Powerful toolkit for natural language processing
 - Lots of data (Corpora)
 - NLP Tools (word/sentence tokenizer, POS tagger, Syntactic parser and so on...)
 - FREEware

4

3. How to install Python?

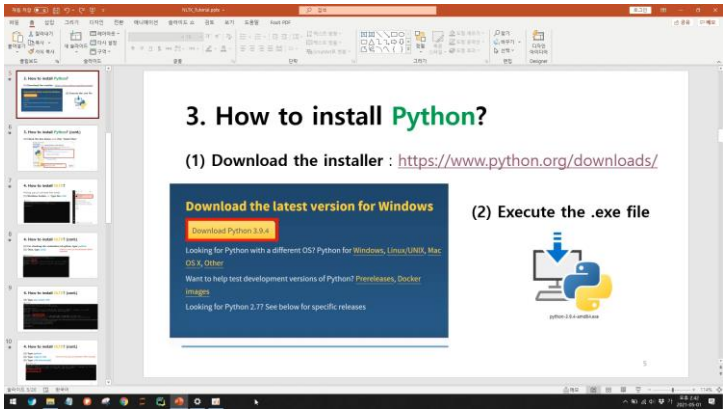
(1) Download the installer : <https://www.python.org/downloads/>



(2) Execute the .exe file

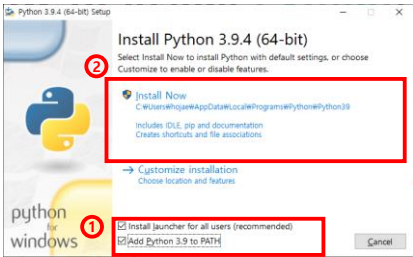


5



3. How to install Python? (cont.)

(3) Check the box below and click "Install Now"

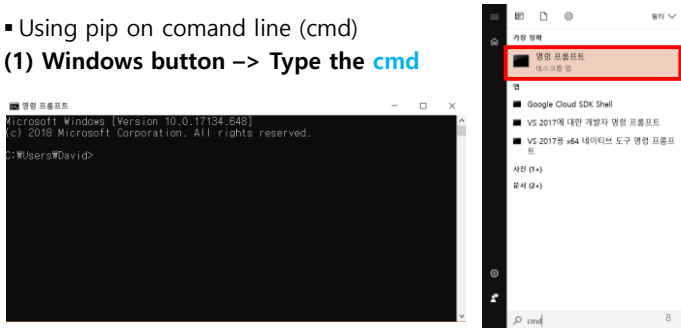


6

4. How to install NLTK?

■ Using pip on comand line (cmd)

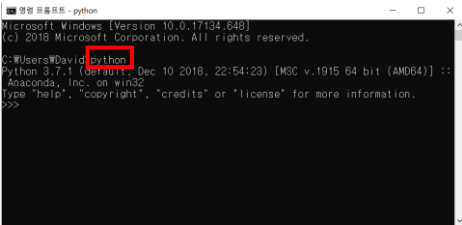
(1) Windows button -> Type the cmd



4. How to install NLTK? (cont.)

- (2) For checking the installation of python, type `python`
- (3) Then, type `exit()`

If there is no error, you have downloaded "Python" successfully

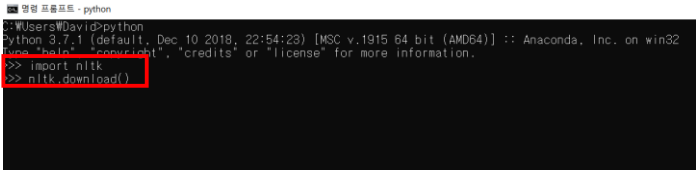


9

4. How to install NLTK? (cont.)

- (3) Type `python`
- (4) Type `import nltk`
- (5) Type `nltk.download()`

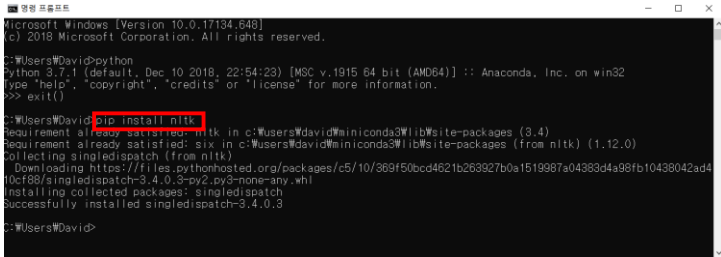
If there is no error, you have downloaded "NLTK" successfully



11

4. How to install NLTK? (cont.)

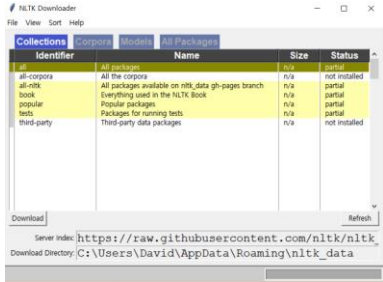
- (4) Type `pip install nltk`



10

4. How to install NLTK? (cont.)

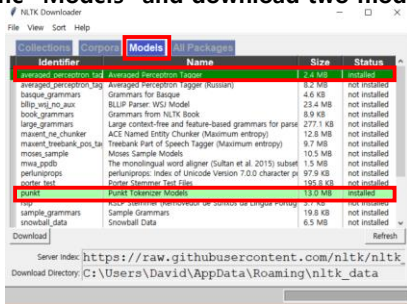
- (6) The NLTK Downloader will pop up



12

4. How to install NLTK? (cont.)

(7) Click the “Models” and download two models below



Identifier	Name	Size	Status
averaged_perceptron_tagger	Averaged Perceptron Tagger	2.4 KB	not installed
averaged_perceptron_tagger	Averaged Perceptron Tagger (Russian)	2.4 KB	not installed
bagbank_grammars	Grammars for bagbank	4.6 KB	not installed
blip_wsj_pos_tag	BLIP Parser: WSJ Model	23.4 MB	not installed
book_grammars	Grammars from NLTK Book	8.9 KB	not installed
large_grammars	Large context-free and feature-based grammars for parsing	277.1 KB	not installed
maxent_ne_chunker	ACE Named Entity Chunker (Maximum entropy)	12.8 MB	not installed
maxent_treelbank_pos_tag	Treelbank Part of Speech Tagger (Maximum entropy)	9.7 MB	not installed
moses_sample	Moses Sample Models	10.5 MB	not installed
mva_gzdb	The monolingual word aligner (Sultan et al. 2015) subset	1.5 MB	not installed
perforcepp	Index of Unicode Version 7.0.0 character properties	97.9 KB	not installed
pos_tagger	Poser Decision Tree Files	195.8 KB	not installed
punkt	Punkt Tokenizer Models	15.0 MB	not installed
sample_grammars	Sample Grammars	19.8 KB	not installed
snowball_data	Snowball Data	6.5 MB	not installed

averaged_perceptron tagger

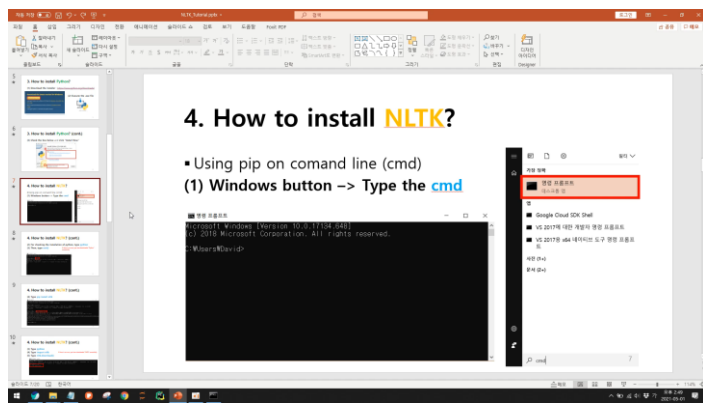
punkt

13

5. What is CFG?

- Context-Free Grammar
 - “context-free” b/c production rules are applied regardless of the context
- Terminal Symbols (Words) & Non-terminal Symbols (Phrases)
- No limit of the number of child nodes (unlike Chomsky Normal Forms)

15



4. How to install NLTK?

- Using pip on comand line (cmd)
- (1) Windows button -> Type the cmd

```
Microsoft Windows [Version 10.0.17134.500]  
© 2018 Microsoft Corporation. All rights reserved.  
C:\Users\David> pip install nltk
```

5. What is CFG? (cont.)

Sentence : I saw the man with a telescope

<Non-terminal Symbols>

S -> NP VP

NP -> Pron | Det N | NP PP

VP -> V NP | VP PP

PP -> P NP | : pipe symbol (shift + W)

<Terminal Symbols>

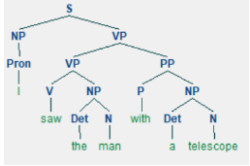
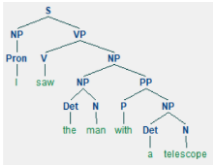
Pron -> ‘I’

V -> ‘saw’

Det -> ‘the’ *Single Quotations!*

N -> ‘man’ | ‘telescope’

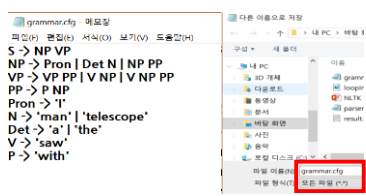
P -> ‘with’



16

6. How to implement CFG?

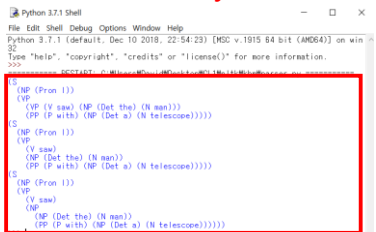
- Write down production rules of your own
- Save the rules under the name 'grammar.cfg' in the **same** directory as 'parser.py'



17

6. How to implement CFG? (cont.)

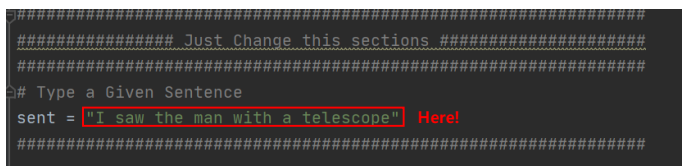
- Click 'F5' (Run the scripts)
 - See the Results
- (If there is **ERROR**, you will see a red message)



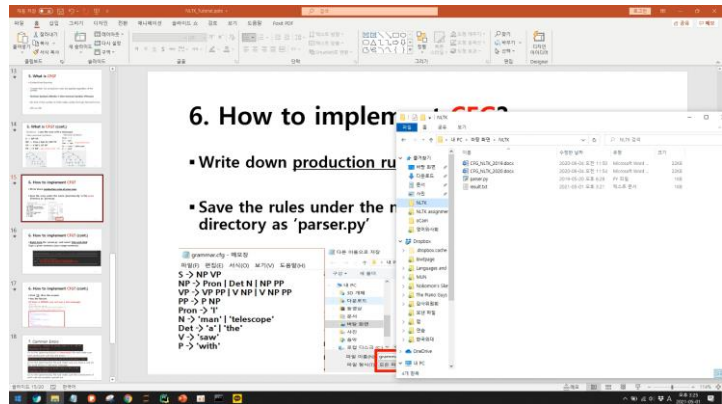
19

6. How to implement CFG? (cont.)

- Right Click the 'parser.py' and select 'Edit with IDLE'
- Type a given sentence (your target sentence)



18



7. Common Errors

```
ValueError: Unable to parse line 1: S - NP VP
Expected an arrow
```

Go to the 'grammar.cfg'(all in lowercase) file and make sure each production rule has the arrow '->'

```
ValueError: Grammar does not cover some of the input words: "saw".
```

Go to the 'grammar.cfg' file and make sure you have a rule for the word shown in the error message

```
ValueError: Unable to parse line 4: PP-> P NP
Expected an arrow
```

Go to the 'grammar.cfg' file and make sure the constituents of each rule are properly spaced out

21

Thank you!

- Below is our contact information.
Contact me if you have questions

Ki Woong Moon	Hee Jeong Na
kiwoongmoon@hufs.ac.kr	920313@naver.com

Computer & Linguistics, Spring 2021

7. Common Errors

```
LookupError:
*****
Resource C: not found.
Please use the NLTK Downloader to obtain the resource:

>>> import nltk
>>> nltk.download('C:')

For more information see: https://www.nltk.org/data.html

Attempted to load /C:/Users/oian/Downloads/nltk/nltk/grammar.cfg
```

Make sure you have saved the rule file under the name 'grammar.cfg' in the same folder as 'parser.py'

22