Part of Speech Tagging

Ko, Youngjoong

Sungkyunkwan University

nlp.skku.edu, nlplab.skku.edu

Contents

- ❖ Morphological Analysis
  - Natural Language Processing
  - Morpheme
  - Morphological analysis
  - Part-Of-Speech Tag
- ❖ Setting environment to develop
  - Google, ‘Colab’
  - Install NLTK
- ❖ Homework

Natural Language Processing

❖ What is Natural Language Processing?

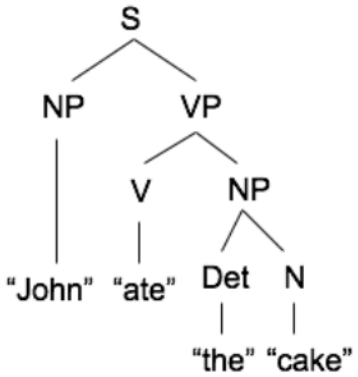
NLP

A branch of artificial intelligence
that helps computers understand,
interpret and manipulate human language.

Natural Language Processing

❖ Levels of Natural Language Processing

Morphology → **Syntax** → **Semantics** → **Discourse**

Morphology	Syntax	Semantics	Discourse
The study of sub-word unit meaning.	The study of structural relationship between words.	The study of literal meaning and semantic relationship between words.	The study of structure of conversation
NNP DT Jone act the cake VBP NN		Jone act the cake NP V NP ((action INGEST) ; syntactic verb (actor JOHN-01) ; syntactic subj (object FOOD)) ; syntactic obj	**Tom** : What are you talking about Hal? : Ask ref

Morpheme

❖ What are Morphemes?

Morpheme The smallest units of meaning in a language

Morphological Analysis

❖ What is Morphological analysis?

Morphological analysis

The identification of the structure of morphemes and other linguistic units, such as root words, affixes, or parts of speech.

❖ What is POS (part-of-speech) tagging?

POS tagging

The process of marking up morphemes in a phrase, based on their definitions and contexts.

Morphological Analysis

❖ Example of Part-Of-Speech Tagging

"John ate the cake"

⇩

POS tagging

⇩

"John/NNP + ate/VNP + the/DT + cake/NN"

Morphological Analysis

❖ Example of NLTK

- Python ‘NLTK’

```
[6]  1 from nltk.tokenize import word_tokenize
     2
     3 tokens = word_tokenize("John ate the cake")
     4 tagged_tokens = nltk.pos_tag(tokens)
     5
     6 print(tagged_tokens)
```

```
[('John', 'NNP'), ('ate', 'VBP'), ('the', 'DT'), ('cake', 'NN')]
```

Part-Of-Speech Tag

❖ Part-Of-Speech Tag

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	big
JJR	adjective, comparative	bigger
JJS	adjective, superlative	biggest
LS	list marker	1)
MD	Modal	could, will
NN	noun, singular or mass	Door
NNS	noun plural	Doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	Predeterminer	*both* the boys
POS	possessive ending	friend*'s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	Adverb	however, usually
RBR	adverb, comparative	Better
RBS	adverb, superlative	Best
RP	Particle	give *up*
TO	To	*to* go, *to* him
UH	Interjection	Uhhuhhuhh
VB	verb, base form	Take
VBD	verb, past tense	Took
VBG	verb, gerund/present participle	Taking
VBN	verb, past participle	Taken
VBP	verb, sing. present, non-3d	Take
VBZ	verb, 3rd person sing. Present	Takes
WDT	wh-determiner	Which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	Whose
WRB	wh-abverb	where, when

http://nltk.org/

Build Environment

❖ Google, ‘Colab’

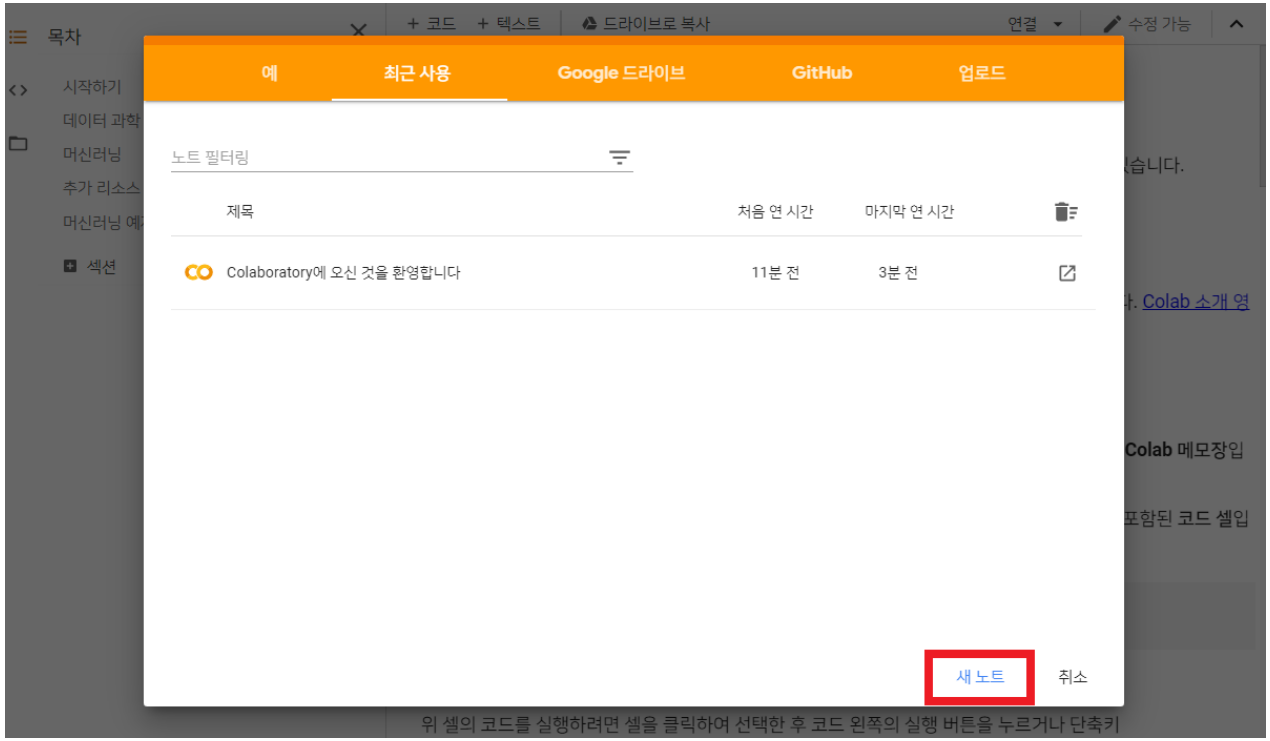
- Google Cloud Development Environment
- This allows you to access a free GPU for up to 12 hours at a time.
- Need a personal Google account

(https://colab.research.google.com/)

Build Environment

❖ Google, ‘Colab’

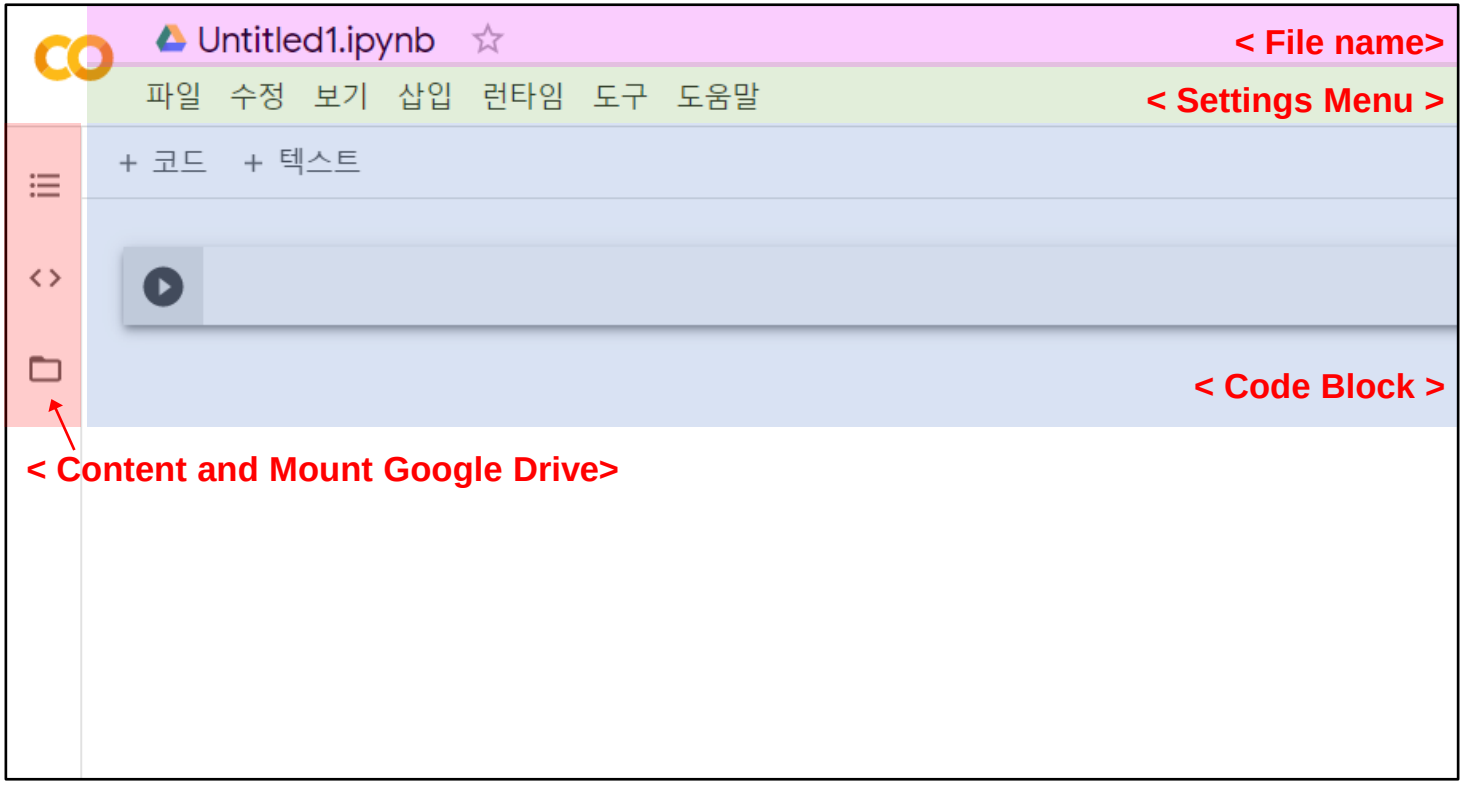
- Create New Notebook



Build Environment

❖ Google, ‘Colab’

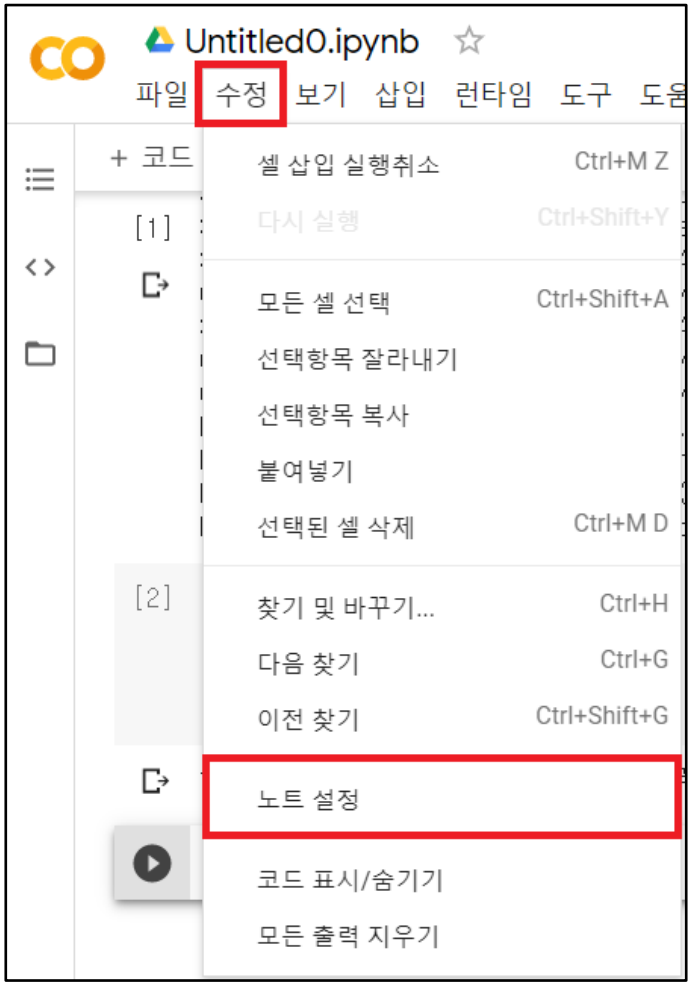
- New Notebook



Build Environment

❖ Google, ‘Colab’

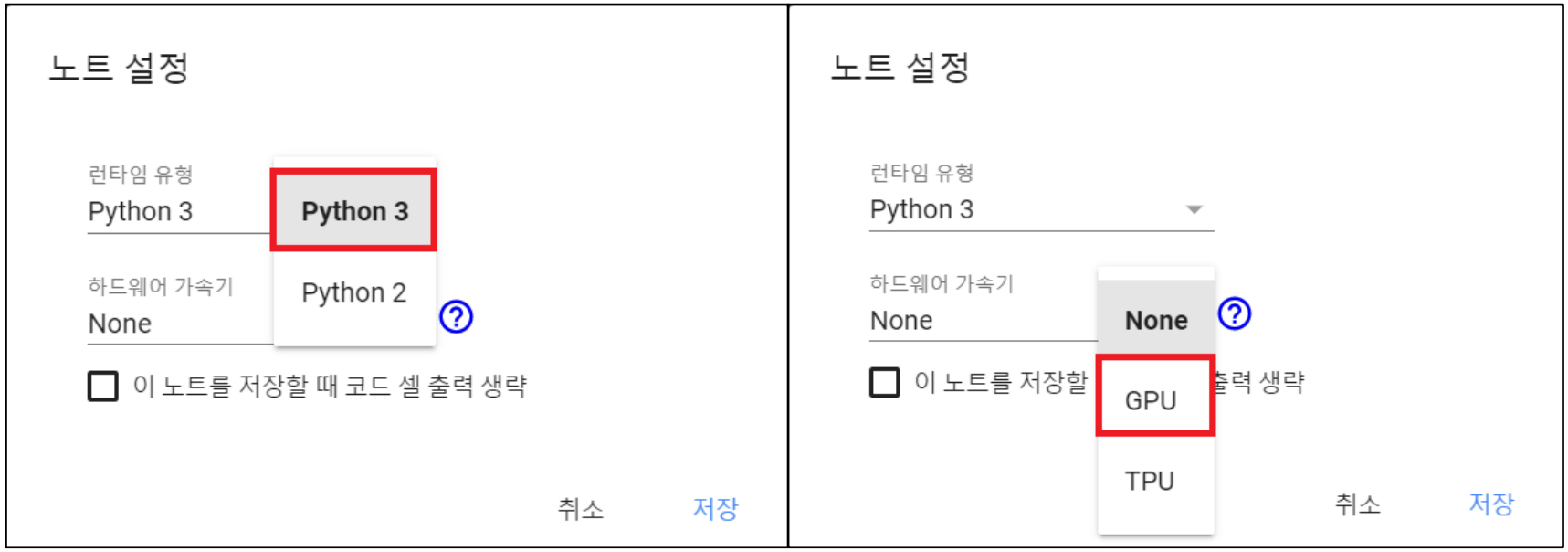
- Notebook Settings



Build Environment

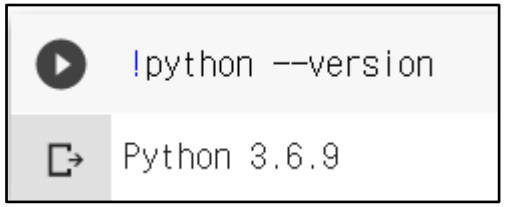
❖ Google, ‘Colab’

- Notebook Settings
  - Python 3
  - GPU



Build Environment

- Google, ‘Colab’
  - Check Notebook settings
    - Python



    - GPU

```
[4]  !nvidia-smi

Tue Mar 17 04:40:14 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.59       Driver Version: 418.67       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   38C    P0    26W / 250W |      0MiB / 16280MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Build Environment

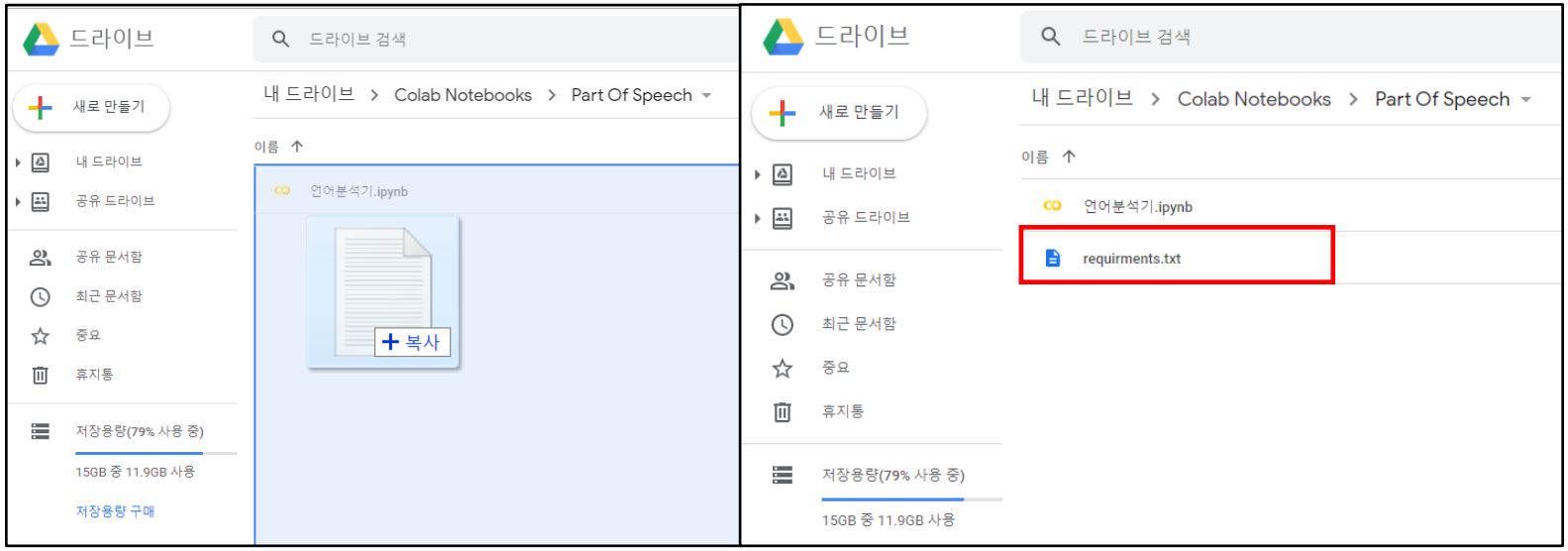
❖ Google, ‘Colab’

- Google Drive Mount

Build Environment

❖ Google, ‘Colab’

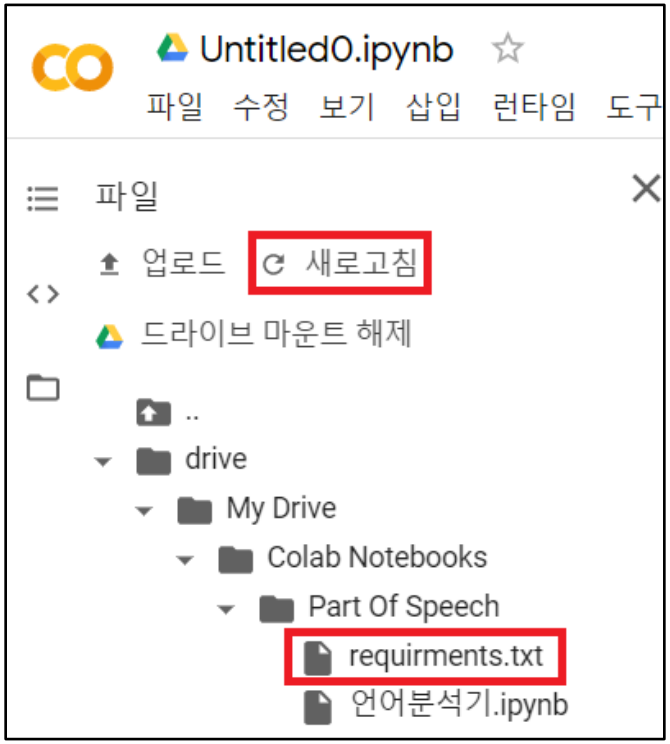
- Upload data to google drive



Build Environment

❖ Google, ‘Colab’

- After refresh, check the drive update history



Build Environment

❖ How to install NLTK in Python

- Install NLTK

```
[7]  !pip install nltk

     Requirement already satisfied: nltk in /usr/local/lib/python3.6/dist-packages (3.2.5)
     Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from nltk) (1.12.0)
```

Build Environment

❖ Example of NLTK

```
[15]  import nltk
      from nltk.tokenize import word_tokenize

      # To use the nltk library
      # You need to write the code below, when you first run the Colab
      nltk.download('punkt')
      nltk.download('averaged_perceptron_tagger')

      tokens = word_tokenize("The sky is blue")
      tagged_tokens = nltk.pos_tag(tokens)

      print(tagged_tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[('The', 'DT'), ('sky', 'NN'), ('is', 'VBZ'), ('blue', 'JJ')]
```

Build Environment

❖ Example of NLTK

```
[16]  import nltk
      from nltk.tokenize import word_tokenize

      # To use the nltk library
      # You need to write the code below, when you first run the Colab
      nltk.download('punkt')
      nltk.download('averaged_perceptron_tagger')

      tokens = word_tokenize("The sky is blue")
      tagged_tokens = nltk.pos_tag(tokens)


      # If you want to get only tokens with specific POS tag
      extract_tokens = list()
      for token, pos in tagged_tokens:
        if pos in ['NN', 'VBZ']:
          extract_tokens.append((token,pos))
      print(extract_tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[('sky', 'NN'), ('is', 'VBZ')]
```

Assignment

❖ Assignment

1) The attached JSON file consists of 300 documents with multiple sentences.

2) In this assignment, you MUST use Python.

3) You can use external libraries.
But you MUST use NLTK library for Tokenizer and POS tagging.

```
[15]  import nltk
      from nltk.tokenize import word_tokenize
```

4) The result text file must include 300 morphologically analyzed documents.
(please refer to format in the page 24)

Assignment

❖ Submit File

1) Python code file (.py) (python version 3.x)
   - Format: “Student Number_Name.py”.
     - Ex) “2020000000_홍길동.py”

2) TEXT file (.txt)
   - Format: “Student Number_Name.txt”.

Result File Format

Student Number_Name.txt

Original text

Ad sales boost Time Warner profit Quarterly profits at US ….(중략)….. stake in AOL Europe as a loss on the value of that stake.

Result text

Ad/NN+sales/NNS+boost/VBP+Time/NNP+Warner/NNP+profit/VB+Quarterly/JJ+profits/NNS+at/IN+US/NNP….(중략)….+stake/NN+in/IN+AOL/NNP+Europe/NNP+as/IN+a/DT+loss/NN+on/IN+the/DT+value/NN+of/IN+that/DT+stake/NN+./.

Document 1

Original text

Dollar gains on Greenspan speech The dollar has hit its highest level….(중략)….. give blind people access to photographs and other images.

Result text

Dollar/NN+gains/NNS+on/IN+Greenspan/NNP+speech/VBP+The/DT+dollar/NN+has/VBZ+hit/VBN+its/PRP$+highest/JJS+level/NN….(중략)….+close/RB+to/TO+half/PDT+a/DT+trillion/CD+dollars/NNS+./.

Document 2

.
.
.
(중략)
.
.

Blind student 'hears in colour' A blind ….(중략)….. stake in AOL Europe as a loss on the value of that stake.

Blind/NNP+student/NN+'hears/NNS+in/IN+colour/NN+'/''+A/DT+blind/JJ….(중략)….+give/VB+blind/NN+people/NNS+access/NN+to/TO+photographs/VB+and/CC+other/JJ+images/NNS+./.

Document 300