Corpus Statistics, Language Model using NLTK

Indian Institute of Technology Kharagpur

Subject :
NLP for E-learning

Anaconda Installation

1. Windows:
   - Download Anaconda Python 3 installer : ▸ Link
   - Follow the default instruction
2. Linux:
   - Download Anaconda Python 3 installer : ▸ Link
   - Go to downloaded directory from terminal and type : bash ./Anaconda3-2019.10-Linux-x86_64.sh
3. macOS :
   - Download graphical installer: ▸ Link
4. Installation Document Link : ▸ Link

Tutorial Requirements

1. First download Wikipedia dump from the ▸ Link
2. Install Packages:
   - Open Anaconda Prompt/terminal
   - Install NLTK : conda install nltk
   - Similarly install packages: gensim, simplejson, smart_open, matplotlib etc.
   - Installing NLTK Data:

```
import nltk
nltk.download()
```

cont..

- □ Convert downloaded dump from .xml-p10*.bz2 to *.json.gz by using single command from bash terminal or Anaconda terminal:

```
python -m gensim.scripts.segment_wiki -i -f
enwiki-latest-pages-articles.xml.bz2 -o enwiki-latest.json.gz
```

- □ Open file wikipedia_preprocessing.ipynb from Jupyter Editor and run to get all the Wikipedia articles in folder wikipedia

Data Structures in Python

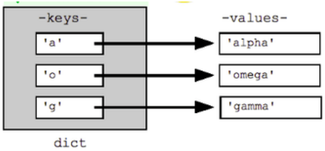
Python Dictionary Basics:

1. To create dictionary in python write:

```
var_name={}
OR
var_name=dict()
```

2. To insert (key, value) pairs in dictionary use command:

```
var_name[key]=value
```



Working with dict data type in python

Contd..

3. To access the (key, value) pairs iterate over the dictionary using loop OR directly use key, if known to get the value

```
var_name[key]
```

4. If key not available in dictionary it will return "None"
5. In order to get default value if key is not available in dictionary, use the syntax:

```
a=var_name.get(key,0)
```

6. In above case if key is not available, "a" is assigned with default value i.e. zero.

Python list Basics

1. To create list type:

   ```
   var_name=[]
   ```

   OR

   ```
   var_name=list()
   ```

2. To insert any element into list:

   ```
   var_name.append(element)
   ```

   OR

   ```
   var_name[index]=element
   ```

3. In order to access element from list, either iterate over list using loops or use list index:

   ```
   val=var_name[index]
   ```

Contd..

4. In order to access fixed number of elements from start (i.e 5 elements)

   ```
   val=var_name[:5]
   ```

5. To access fixed number of elements from last (i.e 5 elements)

   ```
   val=var_name[-5:]
   ```

6. To access elements between given index i to j:

   ```
   val=var_name[i:j]
   # the element at j index is not included in the output
   (excluded)
   #It will output elements from index i to (j-1)
   ```

Resources Link

- □ NLTK Documentation link: ▸ Link
- □ Natural Language Toolkit: Probability and Statistics : ▸ Link
- □ Convert Wikipedia dump to Json documentation ▸ Link
- □ Dictionary Tutorial ▸ Link

POS Tagging and Dependency Parsing using Spacy

Installation :

- □ conda install -c conda-forge spacy
- □ Download spacy models using command :

```
python3 –m spacy download en_core_web_sm
```

- □ Spacy Installation Detail: ▸ Link

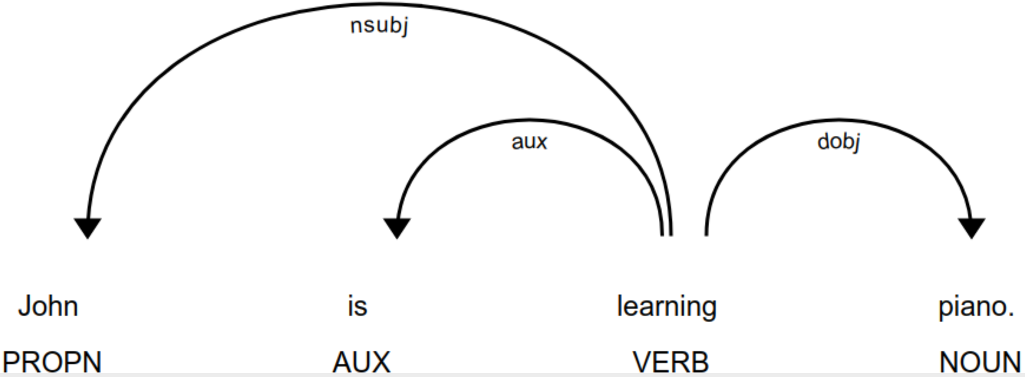
Parts of Speech Tagging

□ List of tags: ▸ Link

POS	Description
ADJ	Adjective
ADV	Adverb
AUX	Auxiliary
CONJ	Conjunction
DET	Determiner
NOUN	Noun

Dependency Parsing

- □ Spacy Dependency parsing: ▸ Link
- □ Input: John is learning piano


nsubj
aux
dobj
John
PROPN
is
AUX
learning
VERB
piano.
NOUN