Shaina Mehta
7CSE 4Y
A2305219268

Assignment 5.1

Q1) List the various libraries and the functions related to the various steps of NLP.

Ans) Libraries are used for:

(a) Lexical Analysis:
1. Nltk
2. Spacy
3. Re

(b) Syntax Analysis:
1. NLTK
2. Spacy

(c) Semantic Analysis:
1. NLTK
2. Spacy
3. Scikit Learn

(d) Discourse Integration:
1. NLTK
2. Spacy
3. Scikit Learn
4. Keras
5. PyTorch

(e) Pragmatic Analysis:
1. NLTK
2. Spacy
3. Scikit Learn
4. Keras
5. PyTorch

Q2) List the various applications of NLP. Explain any two in detail.

Ans) There are the following applications of NLP which are given below as:

1. **Question Answering:** Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.

2. **Spam Detection:** Spam detection is used to detect unwanted e-mails getting to a user's inbox.

3. **Sentiment Analysis:** Sentiment Analysis is also known as opinion mining. It is used on the web to analyze the attitude, behavior, and emotional state of the sender. This application is implemented through a combination of NLP (Natural Language Processing) and statistics by assigning the values to the text (positive, negative, or natural), and identifying the mood of the context (happy, sad, angry, etc.)

4**. Machine Translation:** Machine translation is used to translate text or speech from one natural language to another natural language. Example: Google Translator

5. **Spelling Correction:** Microsoft Corporation provides word processor software like MS-word and PowerPoint for spelling correction.

Shaina Mehta
7CSE 4Y
A2305219268

6. **Speech Recognition:** Speech recognition is used for converting spoken words into text. It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.

7**. Chatbot:** Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.

8. **Information Extraction:** Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

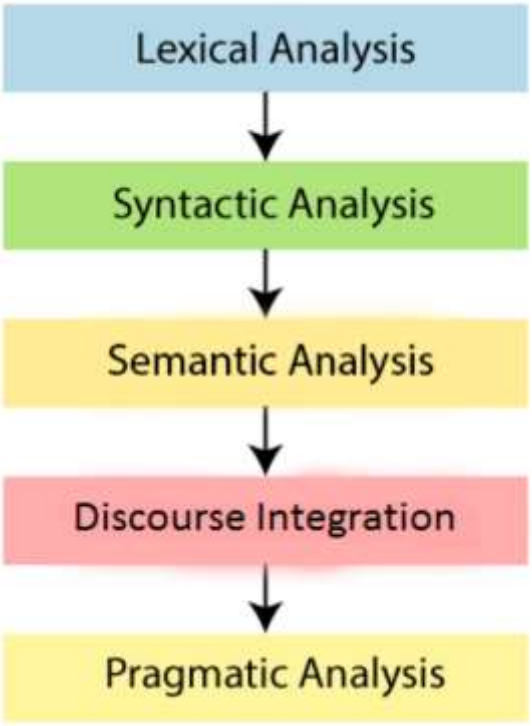
9. **Natural Language Understanding (NLU):** It converts a large set of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate notations of the natural language processing.

Q3) Describe the different stages of NLP.

Ans) Following stages of NLP are given below:

1. **Morphological Analysis/ Lexical Analysis:** Morphological or Lexical Analysis deals with text at the individual word level. It looks for morphemes, the smallest unit of a word. For example, irrationally can be broken into ir (prefix), rational (root), and -ly (suffix). Lexical Analysis finds the relation between these morphemes and converts the word into its root form. A lexical analyzer also assigns the possible Part-Of-Speech (POS) to the word. It takes into consideration the dictionary of the language. For example, the word “character” can be used as a noun or a verb.
2. **Syntax Analysis:** Syntax Analysis ensures that a given piece of text is the correct structure. It tries to parse the sentence to check correct grammar at the sentence level. Given the possible POS generated from the previous step, a syntax analyzer assigns POS tags based on the sentence structure.
   For example:
   Correct Syntax: Sun rises in the east.
   Incorrect Syntax: Rise in sun the east.

Shaina Mehta
7CSE 4Y
A2305219268



3. **Semantic Analysis:** Consider the sentence: "The apple ate a banana". Although the sentence is syntactically correct, it doesn't make sense because apples can't eat. The semantic analysis looks for meaning in the given sentence. It also deals with combining words into phrases.
   For example, "red apple" provides information regarding one object; hence we treat it as a single phrase. Similarly, we can group names referring to the same category, person, object, or organization. "Robert Hill" refers to the same person and not two separate names – "Robert" and "Hill".
4. **Discourse Integration:** Discourse deals with the effect of a previous sentence on the sentence in consideration. In the text, "Jack is a bright student. He spends most of the time in the library." Here, discourse assigns "he" to refer to "Jack".
5. **Pragmatic Analysis:** In the final stage of NLP, Pragmatics interprets the given text using information from the previous steps. Given a sentence, "Turn off the lights" is an order or request to switch off the lights.

Q4) Consider the following passage:

"NLP is a subfield of computer science and artificial intelligence concerned with interactions between computers and human (natural) languages. It is used to apply machine learning algorithms to text and speech.

For example, we can use NLP to create systems like speech recognition, document summarization, machine translation, spam detection, named entity recognition, question answering, autocomplete, predictive typing, and so on.

Nowadays, most of us have smartphones that have speech recognition. These smartphones use NLP to understand what is said. Also, many people use laptops whose operating system has built-in speech recognition."

Shaina Mehta
7CSE 4Y
A2305219268

Carry out the following tasks:

1. **Tokenization**
2. **Vectorization**
3. **Visualization**

Ans) The answer is in the next page.

Shaina Mehta
7CSE 4Y
A2305219268

In [34]:
```
import nltk
```

In [35]:
```
text = "NLP is a subfield of computer science and artificial intelligence concerr
```

In [36]:
```
print('Original Text:',text)
```

```
Original Text: NLP is a subfield of computer science and artificial intelligenc
e concerned with interactions between computers and human (natural) languages.
It is used to apply machine learning algorithms to text and speech. For exampl
e, we can use NLP to create systems like speech recognition, document summariza
tion, machine translation, spam detection, named entity recognition, question a
nswering, autocomplete, predictive typing, and so on. Nowadays, most of us have
smartphones that have speech recognition. These smartphones use NLP to understa
nd what is said. Also, many people use laptops whose operating system has built
-in speech recognition.
```

1. Tokenization

In [38]:
```
punc = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
for ele in text:
    if ele in punc:
        text = text.replace(ele, "")
print("The text after punctuation filter : " + text)
```

```
The text after punctuation filter : NLP is a subfield of computer science and a
rtificial intelligence concerned with interactions between computers and human
natural languages It is used to apply machine learning algorithms to text and s
peech For example we can use NLP to create systems like speech recognition docu
ment summarization machine translation spam detection named entity recognition
question answering autocomplete predictive typing and so on Nowadays most of us
have smartphones that have speech recognition These smartphones use NLP to unde
rstand what is said Also many people use laptops whose operating system has bui
ltin speech recognition
```

In [39]:
```
def tokenization(text):
    tweet = text.lower()
    tokens = nltk.word_tokenize(tweet)
    return tokens
```

Shaina Mehta
7CSE 4Y
A2305219268

In [40]:
```
tokens = tokenization(text)
print('Tokens:',tokens)
```

```
Tokens: ['nlp', 'is', 'a', 'subfield', 'of', 'computer', 'science', 'and', 'art
ificial', 'intelligence', 'concerned', 'with', 'interactions', 'between', 'comp
uters', 'and', 'human', 'natural', 'languages', 'it', 'is', 'used', 'to', 'appl
y', 'machine', 'learning', 'algorithms', 'to', 'text', 'and', 'speech', 'for',
'example', 'we', 'can', 'use', 'nlp', 'to', 'create', 'systems', 'like', 'speec
h', 'recognition', 'document', 'summarization', 'machine', 'translation', 'spa
m', 'detection', 'named', 'entity', 'recognition', 'question', 'answering', 'au
tocomplete', 'predictive', 'typing', 'and', 'so', 'on', 'nowadays', 'most', 'o
f', 'us', 'have', 'smartphones', 'that', 'have', 'speech', 'recognition', 'thes
e', 'smartphones', 'use', 'nlp', 'to', 'understand', 'what', 'is', 'said', 'als
o', 'many', 'people', 'use', 'laptops', 'whose', 'operating', 'system', 'has',
'builtin', 'speech', 'recognition']
```

2. Vectorization

In [41]:
```
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
```

In [42]:
```
def stopwordsRemoval(tokens):
    sw = stopwords.words('english')
    sws=set(sw)
    remsw = [tok for tok in tokens if tok not in sws]
    return remsw
def lemmatization(remsw):
    wnl = WordNetLemmatizer()
    lemms = []
    for rsw in remsw:
        lemms.append(wnl.lemmatize(rsw))
    return lemms
```

In [43]:
```
remsw = stopwordsRemoval(tokens)
print('Tokens which are not stopwords:',remsw)
```

```
Tokens which are not stopwords: ['nlp', 'subfield', 'computer', 'science', 'art
ificial', 'intelligence', 'concerned', 'interactions', 'computers', 'human', 'n
atural', 'languages', 'used', 'apply', 'machine', 'learning', 'algorithms', 'te
xt', 'speech', 'example', 'use', 'nlp', 'create', 'systems', 'like', 'speech',
'recognition', 'document', 'summarization', 'machine', 'translation', 'spam',
'detection', 'named', 'entity', 'recognition', 'question', 'answering', 'autoco
mplete', 'predictive', 'typing', 'nowadays', 'us', 'smartphones', 'speech', 're
cognition', 'smartphones', 'use', 'nlp', 'understand', 'said', 'also', 'many',
'people', 'use', 'laptops', 'whose', 'operating', 'system', 'builtin', 'speec
h', 'recognition']
```

Shaina Mehta
7CSE 4Y
A2305219268

In [44]:
```
lemms = lemmatization(remsw)
print('Tokens after lemmatization:',lemms)
```

```
Tokens after lemmatization: ['nlp', 'subfield', 'computer', 'science', 'artific
ial', 'intelligence', 'concerned', 'interaction', 'computer', 'human', 'natura
l', 'language', 'used', 'apply', 'machine', 'learning', 'algorithm', 'text', 's
peech', 'example', 'use', 'nlp', 'create', 'system', 'like', 'speech', 'recogni
tion', 'document', 'summarization', 'machine', 'translation', 'spam', 'detectio
n', 'named', 'entity', 'recognition', 'question', 'answering', 'autocomplete',
'predictive', 'typing', 'nowadays', 'u', 'smartphones', 'speech', 'recognitio
n', 'smartphones', 'use', 'nlp', 'understand', 'said', 'also', 'many', 'peopl
e', 'use', 'laptop', 'whose', 'operating', 'system', 'builtin', 'speech', 'reco
gnition']
```

In [45]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [46]:
```
count_vector = CountVectorizer(stop_words='english')
x = count_vector.fit_transform(lemms)
```

Shaina Mehta
7CSE 4Y
A2305219268

In [47]:
```
print('Count Vectors:')
print(x)
```

```
Count Vectors:
  (0, 23)       1
  (1, 35)       1
  (2, 6)        1
  (3, 31)       1
  (4, 3)        1
  (5, 14)       1
  (6, 7)        1
  (7, 15)       1
  (8, 6)        1
  (9, 13)       1
  (10, 22)      1
  (11, 16)      1
  (12, 42)      1
  (13, 2)       1
  (14, 20)      1
  (15, 18)      1
  (16, 0)       1
  (17, 37)      1
  (18, 34)      1
  (19, 12)      1
  (20, 41)      1
  (21, 23)      1
  (22, 8)       1
  (24, 19)      1
  (25, 34)      1
  :     :
  (32, 9)       1
  (33, 21)      1
  (34, 11)      1
  (35, 29)      1
  (36, 28)      1
  (37, 1)       1
  (38, 4)       1
  (39, 27)      1
  (40, 39)      1
  (41, 24)      1
  (43, 32)      1
  (44, 34)      1
  (45, 29)      1
  (46, 32)      1
  (47, 41)      1
  (48, 23)      1
  (49, 40)      1
  (50, 30)      1
  (53, 26)      1
  (54, 41)      1
  (55, 17)      1
  (57, 25)      1
  (59, 5)       1
  (60, 34)      1
  (61, 29)      1
```

3. Visualization

Shaina Mehta
7CSE 4Y
A2305219268

In [48]:
```
from collections import Counter
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

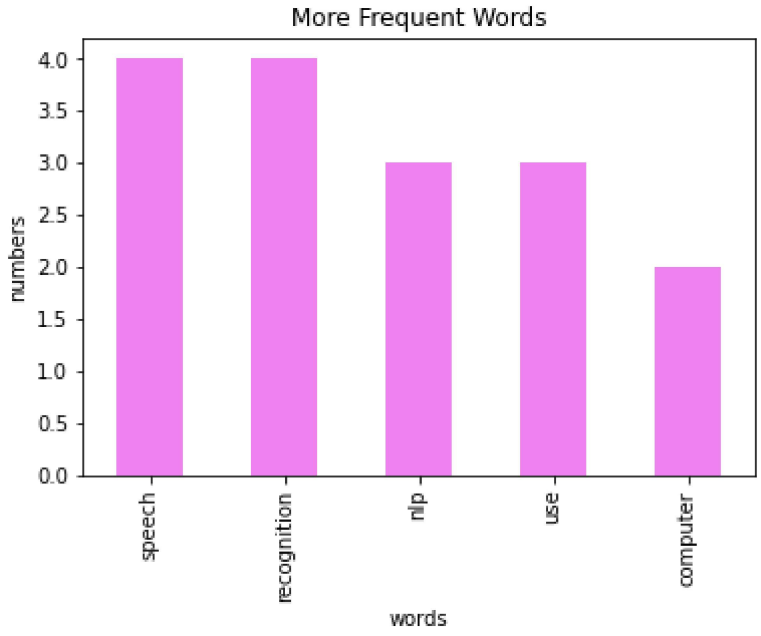
In [49]:
```
count = Counter(lemms).most_common(5)
df = pd.DataFrame.from_dict(count)
df = df.rename(columns={0:"common words",1:"count"})
```

In [50]:
```
df
```

Out[50]:

	common words	count
0	speech	4
1	recognition	4
2	nlp	3
3	use	3
4	computer	2

In [51]:
```
df.plot.bar(legend=False,color='violet')
y_pos = np.arange(len(df['common words']))
plt.xticks(y_pos,df['common words'])
plt.title('More Frequent Words')
plt.xlabel('words')
plt.ylabel('numbers')
plt.show()
```



Shaina Mehta
7CSE 4Y
A2305219268