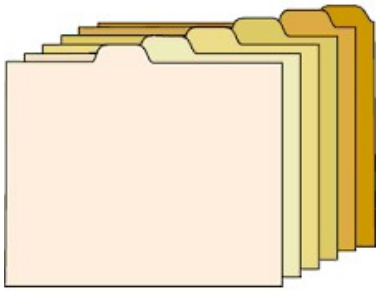
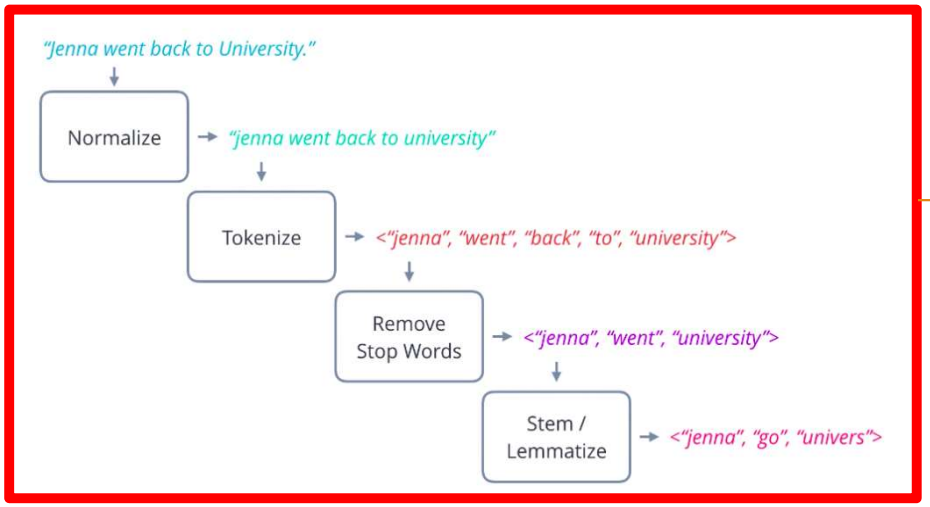
Pre-Processing Textual Data
Lab Session III(a)

Dr. JASMEET SINGH
ASSISTANT PROFESSOR, CSED
TIET, PATIALA

Corpus to Features





Bag-of-Words
Or
Weighted
tf-idf
Matrix

Corpus

Pre-Processing

Feature Vector

Loading your own corpus

Using files:

File_object=open(r"File_Name","Access_Mode")

Access Modes:

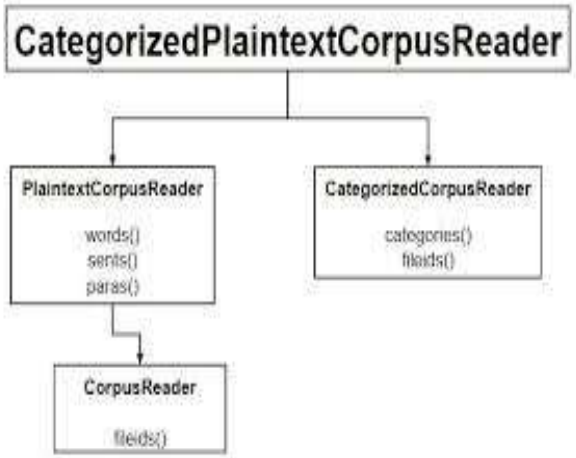
- **Read Only ('r')**
- **Read and Write ('r+')**
- **Write Only ('w')**
- **Write and Read ('w+')**
- **Append Only ('a')**
- **Append and Read ('a+')**

Example

```
import os
filenames=os.listdir('path')
contents=[]
for i  in range(len(filenames)):
    f=open(path+filenames[i],'r')
    text=f.read()
    contents.append(text)
    f.close()
```

Loading your own Corpus Contd….

Using PlaintextCorpusReader:



Example

```
from nltk.corpus import PlaintextCorpusReader
dataset=PlaintextCorpusReader(path ,'.*')
```

Pre-Processing Step 1: Normalization

- Normalization in text includes following steps:

1. Converting the text into same case (lower, upper, or proper
2. Removing numbers, special symbols, urls from text.

Example:

```
corpus = ['Data Science is an important field of science .’,  'This is an important data science course’,  'The cars are driven on the roads .',
 'The trucks are driven on the highways’]
lower=[]
for i in corpus:
    lower.append(' '.join([word.lower() for word in i.split()]))
alpha=[]
for i in lower:
    alpha.append(' '.join([word for word in i.split() if word.isalpha()]))
```

Pre-Processing Step 2: Tokenization

1. Using word_tokenize nltk.tokenize.**word_tokenize**(s): Tokenize a string to split off punctuation other than periods 2. Using split method of list	**Example1:** tokenize=[] from nltk.tokenize import word_tokenize for i in alpha: tokenize.append(word_tokenize(i)) **Example 2:** tokenize=[] for i in alpha: tokenize.append([word for word in i.split()])

Pre-Processing Step 3: Stopword Removal

A stop word is a commonly used word (such as "the", "a", "an", "in") that does not have any linguistic importance in NLP applications

NLTK(Natural Language Toolkit) in python has a list of stopwords stored in *stopwords corpus* in 16 different languages.

The name of fields is the name of language.

```
import nltk from nltk.corpus
import stopwords
print(stopwords.words('english'))
```

Example:

```
import nltk
from nltk.corpus import *
stopword=nltk.corpus.stopwords.words('english')
no_stop=[]
for i in tokenize:
    no_stop.append([word for word in i if word not in stopword])
```

Pre-Processing Step 4: Stemming

Stemming is a process that maps variant word forms to their base forms (play, plays, playing, played)

NLTK has an algorithm named as "PorterStemmer". This algorithm accepts the list of tokenized word and stems it into root word.

```
from nltk.stem import PorterStemmer

ps =PorterStemmer()

ps.stem(w)
```

Example:

```
final=[] #will contain final pre-processed documents

from nltk.stem import PorterStemmer

ps=PorterStemmer()

for i in no_stop:

    final.append(' '.join([ps.stem(word) for word in i]))
```