NATURAL LANGAUGE PROCESSING (DAY 6)

- STEMMING
- LEMMATIZATION

Stemming and Lemmatization are algorithms that are used in Natural Language Processing (NLP) to normalize text and prepare words and documents for further processing in Machine Learning. In NLP, for example, you may want to acknowledge the fact that the words “like” and “liked” are the same word in different tenses.

(Or)

Stemming and Lemmatization are Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing.

Playing ------------------------> play

plays ------------------------> play

played ------------------------> play

"PLAY" IS COMMON ROOT FROM ABOVE WORDS

am, are, is --------------------> be

car, cars, car's, cars' ------------> car

using above mapping a sentence could be normalized as follows:

```
the boy's cars are different colors  ------------> the boy car be differ
color
```

The above example represented "normalization of text"

Stemming and Lemmatization helps us to achieve the root forms (sometimes called synonyms in search context) of inflected (derived) words.

PART 1 STEMMING

STEMMING usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

LEMMATIZATION:

```
Lemmatization usually refers to doing things properly with the use of a
vocabulary and morphological analysis of words, normally aiming to remove
inflectional endings only and to return the base or dictionary form of a
word, which is known as the lemma .
```

(Or)

lemmatizer , a tool from Natural Language Processing which does full morphological analysis to accurately identify the lemma for each word. Doing full morphological analysis produces at most very modest benefits for retrieval.

Loading [MathJax]/extensions/Safe.js

EXAMPLE:

```
 If confronted with the token saw, stemming might return just s, whereas
lemmatization would attempt to return either see or saw depending on
whether the use of the token was as a verb or a noun.
```

- The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma. Linguistic processing for stemming or lemmatization is often done by an additional plug-in component to the indexing process, and a number of such components exist, both commercial and open-source.

NLTK

PorterStemmer

In [4]:
```python
import nltk
from nltk.stem import PorterStemmer
```

In [6]:
```python
porter_stemmer = PorterStemmer()
```

In [10]:
```python
porter_stemmer.stem('went')
```

Out[10]:
```
'went'
```

In [11]:
```python
porter_stemmer.stem('wins')
```

Out[11]:
```
'win'
```

In [12]:
```python
list1 = ['going','writing','eating','meaning',]
```

In [13]:
```python
[porter_stemmer.stem(l) for l in list1]
```

Out[13]:
```
['go', 'write', 'eat', 'algorithms ', 'mean', 'porterstemmer ']
```

In [14]:
```python
porter_stemmer.stem('algorthims')
```

Out[14]:
```
'algorthim'
```

In [15]:
```python
porter_stemmer.stem('porterstemmer')
```

Out[15]:
```
'porterstemm'
```

In [9]:
```python
list_words = ['university','univese','universal', 'teaching','are', 'eating','working']
[porter_stemmer.stem(words) for words in list_words]
```

Out[9]:
```
['univers', 'unives', 'univers', 'teach', 'are', 'eat', 'work']
```

In [ ]:
```python
```

LANCASTER STEMMING

Loading [MathJax]/extensions/Safe.js

In [16]:

```
from nltk.stem import LancasterStemmer
lancaster_stemming = LancasterStemmer()
```

In [17]:

```
lancaster_stemming.stem('working')
```

Out[17]:

```
'work'
```

In [18]:

```
lancaster_stemming.stem('algorthims')
```

Out[18]:

```
'algorthim'
```

In [19]:

```
lancaster_stemming.stem('lancasterstemming')
```

Out[19]:

```
'lancasterstem'
```

In [20]:

```
list1 = ['going','writing','eating','meaning',]
[lancaster_stemming.stem(word) for word in list1]
```

Out[20]:

```
['going', 'writ', 'eat', 'mean']
```

In [21]:

```
list_words = ['university','univese','universal', 'teaching','are', 'eating','working']
[lancaster_stemming.stem(words) for words in list_words]
```

Out[21]:

```
['univers', 'unives', 'univers', 'teach', 'ar', 'eat', 'work']
```

In [ ]:

```

```

Snowball stemming

In [22]:

```
from nltk.stem import SnowballStemmer
snowball_stemming = SnowballStemmer('english')
```

In [23]:

```
snowball_stemming.stem('working')
```

Out[23]:

```
'work'
```

In [24]:

```
snowball_stemming.stem('algorthims')
```

Out[24]:

```
'algorthim'
```

In [25]:

```
snowball_stemming.stem('lancasterstemming')
```

Out[25]:

```
'lancasterstem'
```

In [26]:

```
list1 = ['going','writing','eating','meaning',]
[snowball_stemming.stem(word) for word in list1]
```

Out[26]:

```
['go', 'write', 'eat', 'mean']
```

In [27]:

```
list_words = ['university','univese','universal', 'teaching','are', 'eating','working']
[snowball_stemming.stem(words) for words in list_words]
```

Out[27]:

```
['univers', 'unives', 'univers', 'teach', 'are', 'eat', 'work']
```

In [28]:

```
SnowballStemmer.languages #snowballstemmer this languages
```

Loading [MathJax]/extensions/Safe.js

```
 'danish',
 'dutch',
 'english',
 'finnish',
 'french',
 'german',
 'hungarian',
 'italian',
 'norwegian',
 'porter',
 'portuguese',
 'romanian',
 'russian',
 'spanish',
 'swedish')
```

```
In [ ]:
```

RegexpStemmer

```
In [32]: from nltk.stem import RegexpStemmer
         regex_stemmer = RegexpStemmer('es')
```

```
In [33]: regex_stemmer.stem('takes ')

Out[33]: 'tak '
```

```
In [34]: regex_stemmer.stem('working')

Out[34]: 'working'
```

```
In [35]: regex_stemmer.stem('systems')

Out[35]: 'systems'
```

```
In [36]: regex_stemmer = RegexpStemmer('s')
```

```
In [39]: regex_stemmer.stem('Expression ')

Out[39]: 'Expreion '
```

```
In [38]: regex_stemmer.stem('helps')

Out[38]: 'help'
```

```
In [40]: regex_stemmer = RegexpStemmer('ing')
```

```
In [41]: regex_stemmer.stem('working')

Out[41]: 'work'
```

```
In [42]: regex_stemmer.stem('systems')

Out[42]: 'systems'
```

```
In [ ]:
```

```
In [50]: snowball_stemming = SnowballStemmer('english')
                           mming = LancasterStemmer()
```

Loading [MathJax]/extensions/Safe.js

```
porter_stemmer = PorterStemmer()
```

In [44]:
```
list_words = ["friend", "friendship", "friends", "friendships","stabil","destabilize","mis
print("{0:20}{1:20}{2:20}{3:20}".format("Word","Porter Stemmer","lancaster Stemmer","snowb
for word in list_words:
    print("{0:20}{1:20}{2:20}{3:20}".format(word,porter_stemmer.stem(word),lancaster_stemm
```

```
Word                Porter Stemmer      lancaster Stemmer   snowball stemmer
friend              friend              friend              friend
friendship          friendship          friend              friendship
friends             friend              friend              friend
friendships         friendship          friend              friendship
stabil              stabil              stabl               stabil
destabilize         destabil            dest                destabil
misunderstanding    misunderstand       misunderstand       misunderstand
railroad            railroad            railroad            railroad
moonlight           moonlight           moonlight           moonlight
football            footbal             footbal             footbal
```

In [ ]:

PART2 LEMMATIZATION

WORDNETLEMMATIZER

In [52]:
```
from nltk.stem import WordNetLemmatizer
word_lemmatizer = WordNetLemmatizer()
```

In [53]:
```
word_lemmatizer.lemmatize('working')
```

Out[53]:
```
'working'
```

In [54]:
```
word_lemmatizer.lemmatize('sentences')
```

Out[54]:
```
'sentence'
```

In [55]:
```
word_lemmatizer.lemmatize('misunderstanding')
```

Out[55]:
```
'misunderstanding'
```

In [57]:
```
list_words = ["friend", "friendship", "friends", "friendships","stabil","destabilize","mis
print("{0:20}{1:20}{2:20}{3:20}{4:20}".format("Word","Porter Stemmer","lancaster Stemmer",

for word in list_words:
    print("{0:20}{1:20}{2:20}{3:20}{4:20}".format(word,porter_stemmer.stem(word),lancaster
```

```
Word                Porter Stemmer      lancaster Stemmer   snowball stemmer    lemmatizat
ion
friend              friend              friend              friend              friend
friendship          friendship          friend              friendship          friendship
friends             friend              friend              friend              friend
friendships         friendship          friend              friendship          friendship
stabil              stabil              stabl               stabil              stabil
destabilize         destabil            dest                destabil            destabiliz
e
misunderstanding    misunderstand       misunderstand       misunderstand       misunderst
anding
railroad            railroad            railroad            railroad            railroad
moonlight           moonlight           moonlight           moonlight           moonlight
football            footbal             footbal             footbal             football
```

Loading [MathJax]/extensions/Safe.js

In [ ]:

In [ ]:

USE OF STEMMING AND LEMMATIZATION (ONE EXAMPLE:)

- Stemming and Lemmatization are widely used in tagging systems, indexing, SEOs, Web search results, and information retrieval. For example, searching for fish on Google will also result in fishes, fishing as fish is the stem of both words.

WHY WE USE STEMMING AND LEMMATIZATION:

- Stemming and Lemmatization are methods that help us in text preprocessing for Natural Language Processing.
- Both of them help to map multiple words to a common root word.
- That way, these words are treated similarly and the model learns that they can be used in similar contexts.

Lemmatization vs stemming:

- Stemming is a faster process than lemmatization as stemming chops off the word irrespective of the context, whereas the latter is context-dependent
- Stemming is a rule-based approach, whereas lemmatization is a canonical dictionary-based approach.
- Lemmatization has higher accuracy than stemming.
- Lemmatization is preferred for context analysis, whereas stemming is recommended when the context is not important.

REFERENCES :

```
- https://www.datacamp.com/tutorial/stemming-lemmatization-python
- https://towardsdatascience.com/stemming-vs-lemmatization-in-nlp-dea008600a0
- https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/
- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
- https://www.datacamp.com/tutorial/stemming-lemmatization-python
- https://www.kaggle.com/code/astraz93/beginners-tokenization-stemming-and-lemmatization
- https://www.kaggle.com/code/hassanamin/nlp-basics-including-stemming-and-lemmatization/notebook
```

Loading [MathJax]/extensions/Safe.js