Date: 24/08/2020.

ASSIGNMENT- 04.

#Title: Stemming and feature selection techniques. using vectors.

#Problem statement: Consider a suitable text. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors. Classify documents and evaluate precision and recall.

#Learning objective:
- Implementation of problem using python.
- Remove stop words, applying stemming and feature selection.

#Learning outcome:
- Understanding the stemming and feature selection process.
- Learn about precision and recall.

#Theory:

1] STOP WORDS

- In computing, stop words are words which are filtered out before or after processing of natural language data (text).
- Though "stop words" usually refers to most common words in a language, there is no universal list of stop words used by all natural

processing tools, and indeed not all tools even use such a list.

- Any group of words can be chosen as the stop words for a given purpose. For some search engine following are the most common, short function words:— the, is, at, which, and, on, etc...

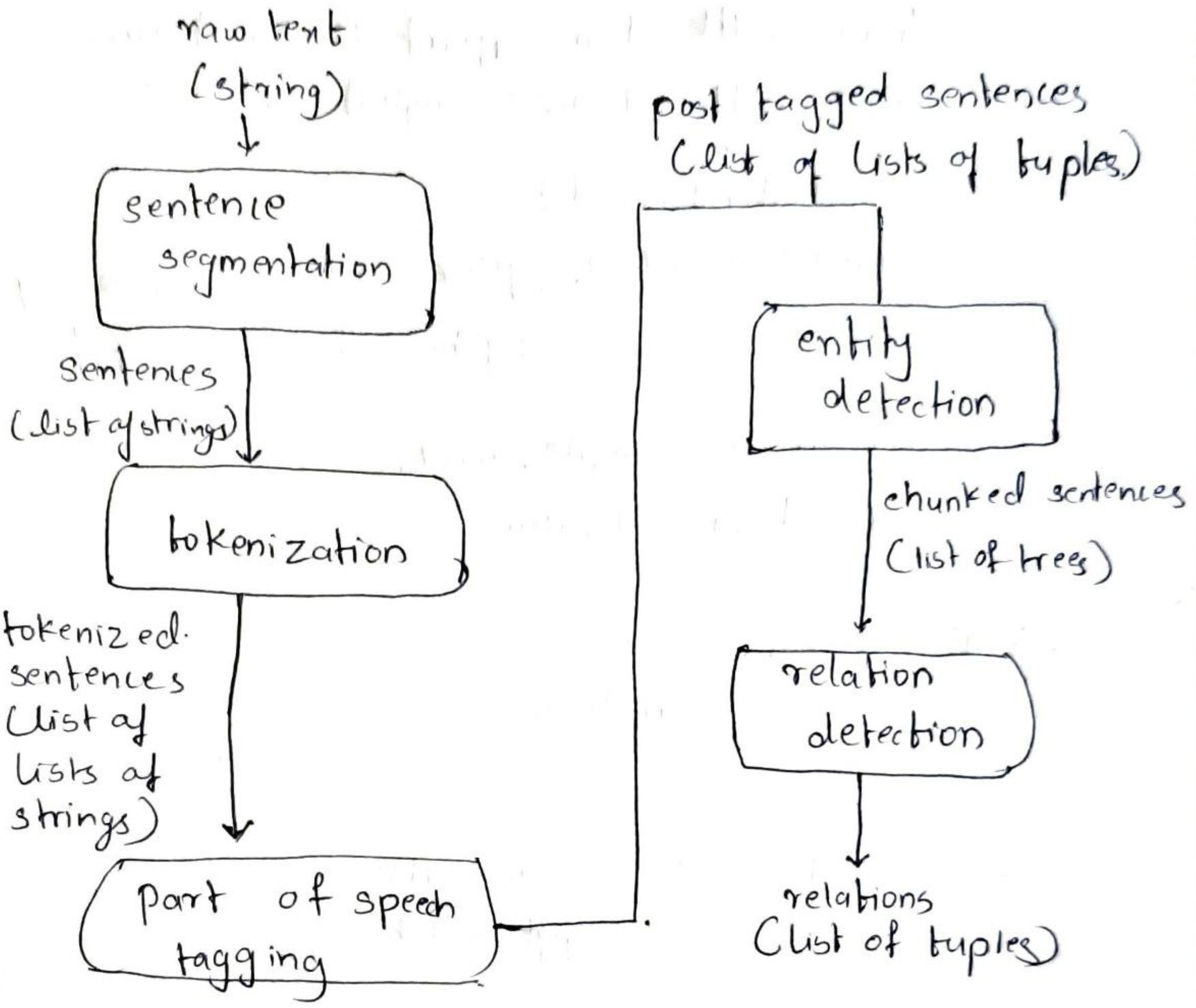
2] STEMMING:-

- Stemming is the process of reducing inflated words to their word stem, base or root form — generally a written word form.
- The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if the stem is not itself a valid root.
- Many search engines treat words with same stems as synonyms as a kind of query expansion, a process called conflation.
- Suffix-stripping algorithm is famous for stemming

3] FEATURE SELECTION

- In machine learning and statistics, Feature selection, also known as variable selection, attribute selection or variable subset selection, is a process of selecting a subset of relavant features (variables, predictors) for use in model construction.

- Feature extraction Architecture



→ Feature selection techniques are used for four reasons :

1. Simplication of models to make them easier to interpret by researchers/users.
2. Shorter training time.
3. To avoid the curse of dimensionality.
4. Enhanced generalization by reducing over fitting (formally, reduction of variance)

→ Stemming with nltk tool in python module.

```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize

ps = PorterStemmer()
example_words = ["python", "pythoner", "pythoning",
                 "pythoned", "pythonly"]

for w in example_words:
    print(ps.stem(w))
```

Test-cases.

Sr.no.	Description	Expected o/p	Actual o/p
1.	Import pandas, os and nltk libraries into jupyter notebook	Success	Success
2.	pre-process the data, and append pos/neg to review and sentiment columns.	Success	Success
3.	Apply suffix-strimming stemming algorithm	Success	Success
4.	Divide data into train and test and obtain accuracy, precision and recall of the model built.	Success	Success

PAGE No.

DATE

Conclusion: Thus, we have studied to remove stop words, apply stemming and feature extraction techniques to represent documents as vectors.