Text Semantics Similarity Machine Learning Group 3

Aneri Sheth(1401072)
Himanshu Budhia (1401039)
Raj Shah (1401050)
Twinkle Vaghela (1401106)

Under the guidance of
Dr. Mehul Raval

May 9, 2017

Introduction

- Given two sentences/phrases, find semantic similarity and classify them.
- Natural Language Processing(NLP) is explored for Semantic Analysis.

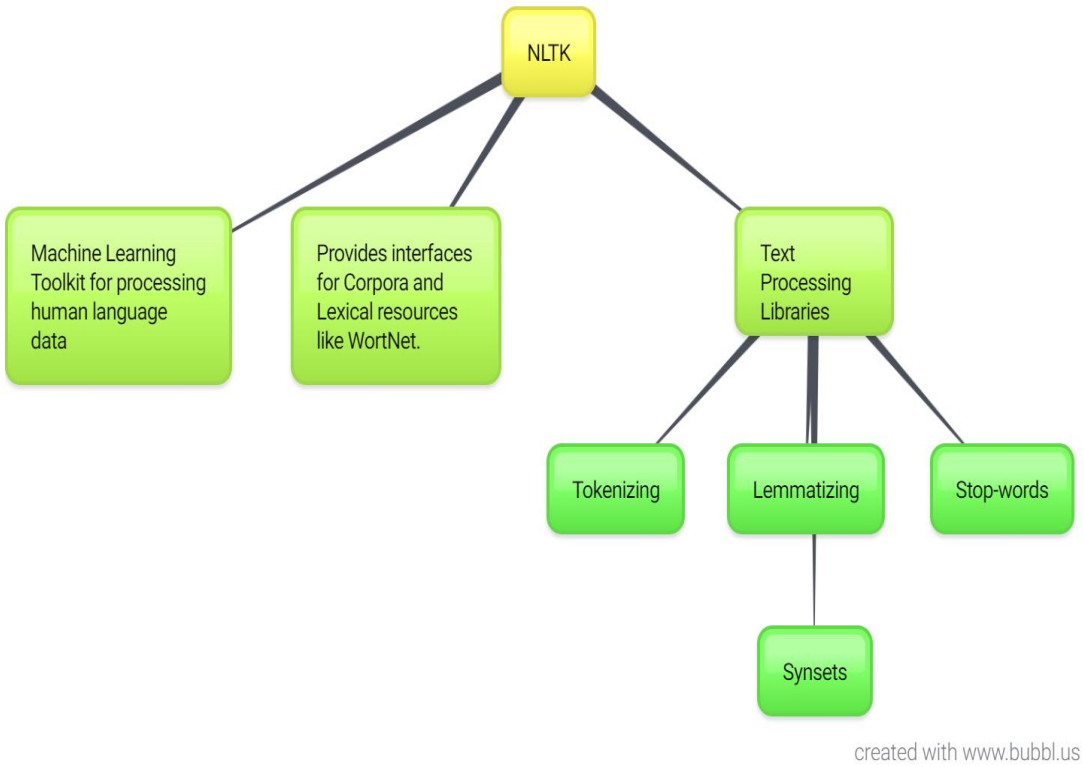

Figure 1 - NLTK

Our Approach

- Supervised Learning Algorithm
- Corpus-based Approach

S1 - A cemetery is a place where dead peoples bodies or their ashes are buried.
S2 - A graveyard is an area of land, sometimes near a church, where dead people are buried.

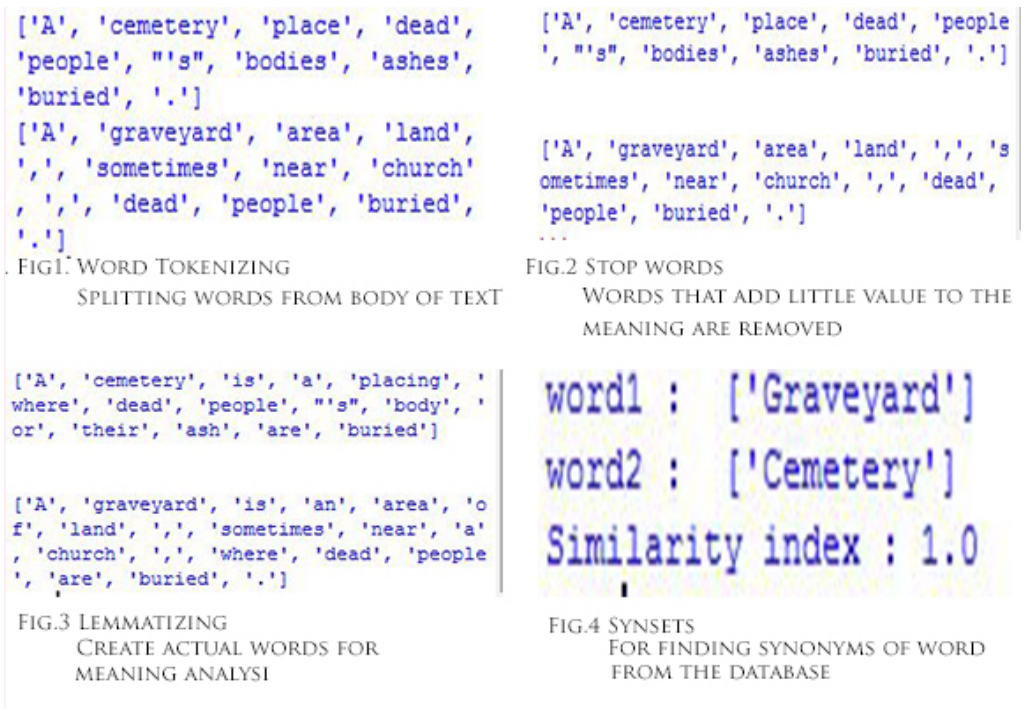

FIG.1 WORD TOKENIZING
SPLITTING WORDS FROM BODY OF TEXT

FIG.2 STOP WORDS
WORDS THAT ADD LITTLE VALUE TO THE MEANING ARE REMOVED

FIG.3 LEMMATIZING
CREATE ACTUAL WORDS FOR MEANING ANALYSI

FIG.4 SYNSETS
FOR FINDING SYNONYMS OF WORD FROM THE DATABASE

Figure 2 - Steps for performing Semantics Similarity

Results

- Let S1 be "I was given a card by her in the garden"and S2 be "In the garden, she gave me a card."
- After eliminating the special characters and punctuations and then all the stop words are removed and the remaining are lemmatized.
- After lemmatizing, we find the synonyms of the lemmatized words which are called synsets. Then, we compare first word of S1 with all the words of S2 and continue this iteratively and find the similarity index of each word with words in the S2.
- If the similarity index is less than 0.65, the sentences are labeled as 'Not Similar', if it is between 0.65 and 0.8, the sentences are labeled as 'Somewhat Similar'and more than 0.8, the sentences are 'Similar'.

Results

```
Sentence 1:  I was given a card by her in the garden.
Sentence 2:  In the garden, she gave me a card.
Similarity index value :  0.95
Similar
```

Figure 3 - Similar Semantics

```
Sentence 1:  Ballmer has been vocal in the past warning that Linux is a threat to Microsoft.
Sentence 2:  In the memo, Ballmer reiterated the open-source threat to Microsoft.
Similarity index value :  0.64
Somewhat Similar
```

Figure 4 - Somewhat Similar Semantics

```
Sentence 1:  The boy is fetching water from the well.
Sentence 2:  The lion is running in the forest.
Similarity index value :  0.57
Not Similar
```

Figure 5 - Dissimilar Semantics

- Semantics Similarity has been done for sentences and phrases. However, for paragraphs and short texts will need complex algorithms for separating of sentences and finding their semantics similarity.
- We find similarity word by word and thus we may get false positives and negatives.
- Our implementation does not consider spellings. To implement that, Longest Common Subsequence (LCS) Algorithm can be used.

THANK YOU