Vidya Vikas Education Trust's
Universal College of Engineering, Kaman Road, Vasai-401208
Accredited B+ Grade by NAAC

Experiment 3

Name of the Student: - Sanket Belekar
Roll No. 67 Batch: A3 Branch: TE COMPS-A
Date of Practical Performed: - Staff Signature with Date & Marks

Aim: Write a program for stemming and lemmatization of text.

Theory:
Theory: Stemming: Stemming is a rule-based approach that produces variants of a root/base word. In simple words, it reduces a base word to its stem word. This heuristic process is the simpler of the two as the process involves indiscriminate cutting of the ends of the words. Stemming helps to shorten the look-up and normalize the sentences for a better understanding.

The process has two main challenges:

Over-stemming: The inflected word is cut off so much that the resultant stem is nonsensical. Over stemming can also result in different words with different meanings having the same stem. For example, "universal", "university" and "universe" is reduced to "univers". Here, even though these three words are etymologically related, their modern meanings are widely different. Treating them as synonyms in a search engine will lead to inferior search results.

Under-stemming: Here, various inflected words have the same stem despite different meanings. The issue crops up when we have several words that actually are forms of one another. An example of under-stemming in the Porter stemmer is "alumnus" → "alumnu", "alumni" → "alumni", "alumna"/" alumnae" → "alumna". The English word has Latin morphology, and so these near-synonyms are not combined.

Lemmatization: Lemmatization entails reducing a word to its canonical or dictionary form. The root word is called a 'lemma'. The method entails assembling the inflected parts of a word in a way that can be recognized as a single element. The process is similar to stemming but the root words have meaning.

Lemmatization has applications in:
● Biomedicine: Using lemmatization to parse biomedicine literature may increase the efficiency of data retrieval tasks.
● Search engines
● Compact indexing: Lemmatization is an efficient method for storing data in the form of index values.

Code:

Stemming

Code:

```
import nltk
nltk.download('punkt')
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

# Sample text
text = "Stemming is used to reduce words to their root form, such as
stemming, stemmed, stems"

# Tokenize the text
words = word_tokenize(text)

# Initialize PorterStemmer
porter = PorterStemmer()

# Apply stemming
stemmed_words = [porter.stem(word) for word in words]

# Print the results
print("Original text:", text)
print("Stemmed text:", " ".join(stemmed_words))
```

Output:

```
Original text: Stemming is used to reduce words to their root form, such as stemming, stemmed, stems
Stemmed text: stem is use to reduc word to their root form , such as stem , stem , stem
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Lemmatization:

Code:

```
import nltk
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize

# Download WordNet (if not already downloaded)
nltk.download('wordnet') nltk.download('punkt')

# Sample text
#text = "Lemmatization is used to group together different inflected forms
of a word"
#text1 = "She likes to read books in her free time and enjoys writing
stories."
#text1 = "The cats are chasing mice and playing in the garden"
text1 = "He was swimming across the river when suddenly a fish jumped out
of the water."


# Tokenize the text
words = word_tokenize(text1)

# Initialize WordNetLemmatizer
lemmatizer = WordNetLemmatizer()

# Apply lemmatization
lemmatized_words = [lemmatizer.lemmatize(word) for word in words]

# Print the results
print("Original text:", text1)
print("Lemmatized text:", " ".join(lemmatized_words))
```

Output:

```
Original text: He was swimming across the river when suddenly a fish jumped out of the water.
Lemmatized text: He wa swimming across the river when suddenly a fish jumped out of the water .
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Conclusion:

Thus, we have successfully studied and performed the concept of lemmatization and stemming of text.