Department of Computer Engineering

Academic Year: 2023-24 Semester: VII Class / Branch: BE COMP

Subject: Natural Language Processing

EXPERIMENT 3
PREPROCESSING: STEMMING & LEMMATIZATION

Aim: Implementation of: (i) Porter Stemmer (ii) Lancaster Stemmer (iii) Lemmatization

Theory:

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.

Languages we speak and write are made up of several words often derived from one another. When a language contains words that are derived from another word as their use in the speech changes is called Inflected Language.

In grammar, inflection is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood. An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change. The degree of inflection may be higher or lower in a language. As you have read the definition of inflection with respect to grammar, you can understand that an inflected word(s) will have a common root form. Few examples:

Playing, Plays, Played -> Play (Common root form 'Play')
Am, Are, Is -> Be
Car, Cars, Car's, Cars' -> Car

Using the above mapping a sentence can be normalized as follows:
the boy's cars are different colors -> the boy car be differe color

Stemming and Lemmatization helps us to achieve the root forms (sometimes called synonyms in search context) of inflected (derived) words. Stemming is different to Lemmatization in the approach it uses to produce root forms of words and the word produced.

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.

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.
Stem (root) is the part of the word to which you add inflectional (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis). So stemming a word or sentence may result in words that are not actual words. Stems are created by removing the suffixes or prefixes used with a word.
Information: Removing suffixes from a word is called Suffix Stripping.

Stemming Algorithms and Code

There are English and Non-English Stemmers available in nltk package. A computer program or subroutine that stems word may be called a stemming program, stemming algorithm, or stemmer.
For the English language, you can choose between Porter Stemmer or Lancaster Stemmer, PorterStemmer being the oldest one originally developed in 1979. LancasterStemmer was developed in 1990 and uses a more aggressive approach than Porter Stemming Algorithm.
PorterStemmer uses Suffix Stripping to produce stems. PorterStemmer algorithm does not follow linguistics rather a set of 05 rules for different cases that are applied in phases (step by step) to generate stems. This is the reason why PorterStemmer does not often generate stems that are actual English words. It does not keep a lookup table for actual stems of the word but applies algorithmic rules to generate stems. It uses the rules to decide whether it is wise to strip a suffix. One can generate its own set of rules for any language. PorterStemmer is known for its simplicity and speed.
The LancasterStemmer (Paice-Husk stemmer) is an iterative algorithm with rules saved externally. One table containing about 120 rules indexed by the last letter of a suffix. On each iteration, it tries to find an applicable rule by the last character of the word. Each rule specifies either a deletion or replacement of an ending. If there is no such rule, it terminates. It also terminates if a word starts with a vowel and there are only two letters left or if a word starts with a consonant and there are only three characters left. Otherwise, the rule is applied, and the process repeats.
LancasterStemmer is simple, but heavy stemming due to iterations and over-stemming may occur. Over-stemming causes the stems to be not linguistic, or they may have no meaning.
For example, in above code destabilized is stemmed to dest in LancasterStemmer whereas, using PorterStemmer destabl. LancasterStemmer produces an even shorter stem than porter because of iterations and over-stemming is occurred.

Lemmatization with Python nltk package

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. A lemma (plural lemmas or

lemmata) is the canonical form, dictionary form, or citation form of a set of words.

For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words. Because lemmatization returns an actual word of the language, it is used where it is necessary to get valid words.
Python NLTK provides WordNet Lemmatizer that uses the WordNet Database to lookup lemmas of words.

Conclusion:
Stemming and Lemmatization both generate the root form of the inflected words. The difference is that stem might not be an actual word whereas, lemma is an actual language word.
Stemming follows an algorithm with steps to perform on the words which makes it faster. Whereas, in lemmatization, you used WordNet corpus and a corpus for stop words as well to produce lemma which makes it slower than stemming. You also had to define a parts-of-speech to obtain the correct lemma.

Department of Computer Engineering

Academic Year: 2023-24 Semester: VII Class / Branch: BE COMP

Subject: Natural Language Processing

Sample Code:

Department of Computer Engineering

Academic Year: 2023-24 Semester: VII Class / Branch: BE COMP

Subject: Natural Language Processing

Output:

```
Porter Stemmer
Give: give
giving: give
given: given
Given: given
giver: giver
gives: give
gave: gave
regives: regiv

Stemming a sentence
['He', 'wa', 'run', 'and', 'eat', 'at', 'same', 'time', '.', 'He', 'ha', 'bad', 'habit', 'of', 'swim', 'after', 'play', 'long', 'hour', 'in', 'the', 'sun', '.']
================================
Lancaster Stemmer
Give:giv
giving:giv
given:giv
Given:giv
giver:giv
gives:giv
gave:gav
regives:reg
Stemming a sentence
['he', 'was', 'run', 'and', 'eat', 'at', 'sam', 'tim', '.', 'he', 'has', 'bad', 'habit', 'of', 'swim', 'aft', 'play', 'long', 'hour', 'in', 'the', 'sun', '.']
================================
WordNet Lemmatizer
giving : giving
given : given
Given : Given
giver : giver
gives : give
gave : gave
regives : regives
================================
WordNet Lemmatizer on a sentence
['He', 'was', 'running', 'and', 'eating', 'at', 'same', 'time', '.', 'He', 'has', 'bad', 'habit', 'of', 'swimming', 'after', 'playing', 'long', 'hours', 'in', 'the', 'Sun', '.']
He wa running and eating at same time . He ha bad habit of swimming after playing long hour in the Sun.
```

Department of Computer Engineering

Academic Year: 2023-24 Semester: VII Class / Branch: BE COMP

Subject: Natural Language Processing

Assignments:

1. Take the input from the user as list of words to find the root words form, using Porter, Lancaster stemmer and lemmatizer. Write the processed text in the out file called: processedtext.txt
2. Design a menu driven program to take input from the user to perform porter stemming or lancaster stemming or lemmatization. Read input from the file. Display the processed text to the user.