Assignment Part 1

M.Nikhil(EE17B138),Anirudh(EE17B043)

February 28, 2020

Question 1

The most obvious top down version to solve sentence segmentation is to split a sentence at end of full-stop or question mark when it is followed by space(‘. ’ or ‘? ’).

Question 2

No,For the following sentence takes from cranfield docs

> observations include.. flow visualization, spark- schlieren pictures of the fluctuations of the free shear layer, and studies of the diffusion of heat from sources placed in the separated region .

our rule for breaking incorrectly breaks at include.. . The above method also fails when grammar rules are not followed strictly

Question 3

Punkt tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences. It must be trained on a large collection of plain text in the target language before it can be used.

The NLTK data package includes a pre-trained Punkt tokenizer for English.

Question 4

part a

No example for now to say naive is better than punkt tokenizer.

part b

Example mentioned in question 2 and ”a. ferri’s vortical layer is brought into evidence .” In this sentence if we follow naive approach we have to split at a. which is wrong.

Question 5

Simplest way to tokenize words is to split the words when they are separated by white spaces or commas

Question 6

The Treebank tokenizer uses regular expressions to tokenize text as in Penn Treebank. This implementation is a port of the tokenizer sed script written by Robert McIntyre and available at `http://www.cis.upenn.edu/~treebank/tokenizer.sed`.

It is a top down approach based on some fixed rules.

Question 7

part a

No example for now to say naive is better than Treebank tokenizer.

part b

Naive word tokenizer does not consider other punctuation marks so for this example 'how do interference-free longitudinal stability measurements (made using free-flight models) compare with similar measurements made in a low-blockage wind tunnel .' here brackets are ignored by naive model but not by Treebank tokenizer.

Question 8

Stemming is getting the base word after removing suffixes eg: Dancing-danc.This done by using a top down approach using some fixed rules

Where as lemmatization is getting the meaningful root word of the words eg:Dancing-dance,Ate- eat. Lemmatization requires part of speech of the word for better result and works for word for which we prior knowledge of root word.

Question 9

For search engines stemming is better than lemmatization The downsides of using lemmatization over stemming are.

- Lemmatization also gives less no of related words as it focuses on precision.eg: operate, operation and operational even though they are related they give different root words in lemmatization(themselves) while stemming gives same result(oper).This means we can get more related documents while using stemming over lemmatization which is what required for a general search engine.
- without pos tags to the word it is assumed as noun by wordnet lemmatizer.This means it is not accurate for some cases like meeting as as noun vs meeting as a verb.This can be solved by feeding sentences to spacy and lemmatizing during which spacy automatically finds pos tags.

but there are some downsides of using stemming as lemmatization does full morphological analysis to accurately identify the lemma for each word instead of following fixed rules eg: for 'leaves' and 'leaf' stemming gives 'leav' and 'leaf' while lemmatization gives 'leaf' and 'leaf'.Another advantage is we can improve lemmatization performance by updating dictionary and also results are interpretable .

Question 12

Generally,

- Stop words have high frequency of occurrence
- Stop words are general words and not used specifically in a certain domain,hence they occur in documents of every domain.

From this we can gather all words with high number occurrences and high number of documents it occurred from corpora representing domains we are interested.