2017 Fall: COMP-SCI 5590/490 - Special Topics

Python Programming

In Class Exercise 7

**Course Plan**

This lesson focusses on some work on text processing for example unigram, bigram, trigram, tokenization, pose tagging, lemmatization, normalization, entity extraction, language model. Learning these features will help us for more meaningful project as document classification, spelling corrector, document summarization, etc.

**Use case Description**

Given a spelling corrector source code, we are aiming to go even deeper in some concept in natural language processing. In this use-case we will learn how to correct the mis-typing of words in a sentence.

Spelling corrector is about using some of the NLP features we learned during the class then correcting a mis-typed word. Thus, students are able to see the right application of these features in a project.

**In-Class Exercise**

In class, we further work on the tokenization, pos-tagging, entity extraction, biagram and trigram.

For all the exercises import the right module from NLTK. You need to go through the slides to find them.

1. Define a paragraph of random text of three sentences in any text file. Name it as input.txt.
2. Use WordNet module to find the meaning of some words in input.txt.
3. Apply the following on the text and show output:

* Tokenization
* Stemming
* POS
* Lemmatization
* Trigram
* Named Entity Recognition