CS410: Tech Review

Topic: Comparison of NLP Toolkits

Soumya Kanti Dutta (skdutta2@illinois.edu)

What is NLP (Natural Language Processing)

Natural languages are built to communicate among human beings. Human can comprehend the natural Languages because human has common sense to understand the meaning of words,sentences and word phases. But machine or computer does not have the common sense or large historical databases to infer Knowledge to comprehend the word relationships and meaning of the sentences or phases.

Natural Language Processing (NLP) makes it possible for computers to understand the human language(eg. English or Bengali).

Probably, the most popular examples of NLP in action are virtual assistants, like Google Assist, Siri, and Alexa. NLP understands and translates the human language, like *"Hey Siri, where is the nearest gas station?"* into numbers, making it easy for machines to understand.

Now the question arises how these tools or systems are able to comprehend complex human natural Language to machine comprehensive languages.

Natural Language Processing (NLP) applies two techniques to help computers understand text: **syntactic analysis and semantic analysis**. We are not discussing in detail of these techniques here. Because this is out of scope of this topic.

Syntactic Analysis

Some of its main sub-tasks include:

- **Tokenization** consists of breaking up a text into smaller parts called *tokens* (which can be sentences or words) to make text easier to handle.
- **Part of speech tagging (PoS tagging)** labels tokens as *verb, adverb, adjective, noun*, etc. This helps infer the meaning of a word (for example, the word "book" means different things if used as a verb or a noun).
- **Lemmatization & stemming** consist of reducing inflected words to their base form to make them easier to analyze.
- **Stop-word removal** removes frequently occuring words that don't add any semantic value, such as *I, they, have, like, yours*, etc.These words occuer everywhere but do not entitle any value.

Semantic Analysis

- **Word sense disambiguation** tries to identify in which sense a word is being used in a given context.
- **Relationship extraction** attempts to understand how entities (places, persons, organizations, etc) relate to each other in a text.

Popular NLP Tools to handle this task:

A typical information extrction pipeline looks like to the following figure: