Getting Started with Natural Language Processing

Natural Language Processing

Natural language processing (NLP) is concerned with enabling computers to interpret, analyze, and approximate the generation of human speech. Typically, this would refer to tasks such as generating responses to questions, translating languages, identifying languages, summarizing documents, understanding the sentiment of text, spell checking, speech recognition, and many other tasks. The field is at the intersection of linguistics, AI, and computer science.

Language Models

Language models are probabilistic machine models of language used for NLP comprehension tasks. They learn a *probability of word occurrence* over a *sequence* of words and use it to estimate the relative likelihood of different phrases. This is useful in many applications, such as speech recognition, optical character recognition, handwriting recognition, machine translation, spelling correction, and many other applications.
Common language models include:

- Statistical models
    - Bag of words (unigram model)
        - applications include term frequency, topic modeling, and word clouds
    - *n*-gram models
- Neural Language Modeling (NLM).

Natural Language Toolkit

Natural Language Toolkit (NLTK) is a Python library used for building Python programs that work with human language data for applying in statistical natural language processing (NLP).
NLTK contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets for NLP.

Text Similarity in NLP

Text similarity is a facet of NLP concerned with the similarity between texts. Two popular text similarity metrics are *Levenshtein distance* and *cosine similarity*.
Levenshtein distance, also called edit distance, is defined as the minimum number of edit operations (deletions, insertions, or substitutions) required to transform a text into another.
Cosine similarity measures the cosine of the angle between two vectors. To determine cosine similarity, text documents need to be converted into vectors.

Language Prediction in NLP

Language prediction is an application of NLP concerned with predicting language given preceding language. Auto-suggest and suggested replies are common forms of language prediction. Common approaches inlcude:

- *n*-grams using Markov chains,
- *Long Short Term Memory (LSTM)* using a neural network.