**NLP and N-GRAMS**

**Abstract**

NLP, Natural Language Processing is broadly defined as the automatic manipulation of natural language, like speech and text, by software. Natural language refers to the way we, humans, communicate with each other. Namely, speech and text. We are surrounded by text.

Natural Language Processing is the driving force behind the following common applications:

* Language translation applications such as Google Translate
* Word Processors such as Microsoft Word and Grammarly that employ NLP to check grammatical accuracy of texts.
* Interactive Voice Response (IVR) applications used in call centers to respond to certain users’ requests.
* Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.

N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occuring words within a given window and when computing the n-grams you typically move one word forward (although you can move X words forward in more advanced scenarios). For example, for the sentence *"The cow jumps over the moon"*. If N=2 (known as bigrams), then the ngrams would be:

* the cow
* cow jumps
* jumps over
* over the
* the moon

The value of N determines the Ngrams. If N=3 then we may have the ngrams as “The cow jumps”. N can take value from 1 or more.

N-grams are used for a variety of different task. For example, when developing a language model, n-grams are used to develop not just unigram models but also bigram and trigram models. Google and Microsoft have developed web scale n-gram models that can be used in a variety of tasks such as spelling correction, word breaking and text summarization . It can also help make next word predictions. The partial sentence “Please hand over your”. Then it is more likely that the next word is going to be “test” or “assignment” or “paper” than the next word being “school”.

Rahul Dokania

16841A0527