LinkedIn NLP Workflow

* Natural Language Processing (NLP)
* Will require the [NLTK](https://www.nltk.org/api/nltk.html) (Natural Language Toolkit)
* Cross Industry Standard Process for Data Mining (CRISP-DM) workflow [1]

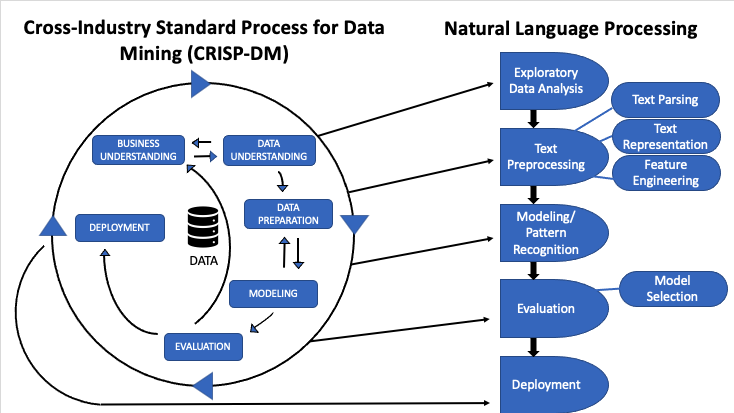


Figure 1: CRISP-DM workflow [1]

**Data Cleaning**

* Vectorize the text to create a Bag of Words (BoW)
  + Count how many times each word appears in a corpus [1]
* Need to clean and tokenize text data first:
  + Remove punctuation and lowercase everything
  + 1. Make lower case
  + 2. Remove punctuation
  + 3. Remove stop words
  + 4. Remove numbers
    - Note, I think numbers may be valuable here, so let’s try not removing them
* Stemming
  + Removing end of words down to root word
* Lemmatization
  + Instead of stemming, can transform the word to its root
  + Seems more useful
  + Used instead of stemming
* Part-of-Speech (POS) Tagging
  + Adds a ‘tag’ to each word like “Noun”, “Adjective”, etc.
  + Package has a lot of fancy tags

**Bag of Words**

* Create a document term matrix, each column is a unique word, each row is a document
* Term Frequency-Inverse Document Frequency (TF-IDF)
  + Scales down the frequencies of tokens to represent how important a word is

References

[1] <https://towardsdatascience.com/natural-language-processing-workflow-1dddf3a48ab5>

[2] <https://towardsdatascience.com/nlp-with-lda-latent-dirichlet-allocation-and-text-clustering-to-improve-classification-97688c23d98>

[3] <https://towardsdatascience.com/topic-modeling-in-nlp-524b4cffeb68>

[4] <https://sanjayasubedi.com.np/nlp/nlp-with-python-document-clustering/>