Evaluating NLP models:

Extract features out of texts (Feature Extraction) => Vectors for each word

* Cosine Similarity
* Synonyms
* Clustering:
  + Cluster texts, words together using either supervised or unsupervised learning
  + => See if the clustering makes sense

Labels data

* by source: Github, Leetcode, Conference,
* by data attribute: Python, Java, C++
* => Train supervised clustering
* Without labels, try cluster base on tag, etc

Next Steps:

* Pick our own interesting data
* Try Data Preprocessing: Stemming, Lemming, clean stop words, etc
* Feature Extraction
* Get Word Embeddings (NLTK, Standford NLP, spacy, TF-iDF etc, …)
* Evaluate (Clustering, word similarity)
* Check **POS tags**: Take a sample 5 rows of data, tag them yourself, then run POS tag on it to see if the model tag it the way you do
* => Conclude how does the model perform, can it predict, spam email?, can it group good reviews together?