**Data Parsing and Pre-process**

Done..scraping.py, tokenStemRemove.py

**Feature Extraction**

So far TfidfVectorizer is used successfully to extract features from all\_plays\_tokenized.txt to form bag-of-words, the extracted features are arrays of similarities among plays, acts within each play and scenes within each play, those arrays are built based on the frequency use of words. The features of each array is limited to 5000 max.

TfidfVectorizer uses a in-memory vocabulary (a python dict) to map the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix. The word frequencies are then reweighted using the Inverse Document Frequency (IDF) vector collected feature-wise over the corpus.

Useful links :

<http://scikit-learn.org/dev/modules/feature_extraction.html#text-feature-extraction>

<http://scikit-learn.org/dev/auto_examples/text/document_clustering.html#example-text-document-clustering-py>

<http://scikit-learn.org/dev/auto_examples/applications/plot_out_of_core_classification.html#example-applications-plot-out-of-core-classification-py>

<https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words>

To-do:

1. Test to apply PCA to further extract features and reduce dimensionality

2. Apply k- means and other clustering methods

3. Somehow there are dtype error using HashingVectorizer and TfidfTransformer these two feature extraction methods—open to fix and try

4. Is it to compare acts or scenes’ bag of words similarities within each play or across play?