**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.

Latent Semantic Analysis is applied then to reduce the dimensionality from 5000 to 2, however the explained variance of the LSA reduced model is about 5% to 10% only.

The extracted features are exported as txt files in folders, so clustering can be done by importing those extracted features arrays, instead waiting for the extraction completion on the raw all\_data.txt fiel, which process can take minutes.

Using K means++ to cluster, the vectorized 5000- feature array got much better clustering resulst than using LSA reduced 2 feature array. Maybe increasing features will get better result?

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 the effect of apply PCA to extract features and reduce dimensionality , and compare with the k clustering results with LSA reduced dataset

2. Add label text files to denote the row in feature files coming from which play, so the Homogeneity, Completeness, V-measure can be calculated between labels of original and km.labels

3. Somehow there are dtype error using HashingVectorizer and TfidfTransformer these two feature extraction methods—open to fix and try if these 2 methods will work better than TfidfVectorizer

4. Cross – validation needed? How to detect suspicious outlier( doubted not Shakespear authored works?)