Text Clustering

Take five different samples of Gutenberg digital books, which are all of five different genres and of five different authors, that are semantically different. Separate and set aside unbiased random partitions for training (no need for test segment necessarily!).

The overall aim is to produce *similar* clusters and compare them; analyze the pros and cons of algorithms, generate and communicate the insights.

Prepare the data: create random samples of 200 documents of each book, representative of the source input.
Preprocess the data; prepare the records of 150 words records for each document,
Label them as a, b, c etc. as per the book they belong to so can later compare with clusters.
Transform to BOW and TF-IDF (also use other features LDA, Word-Embedding).
Evaluation: Calculate *Kappa* against true authors, *Consistency, Coherence* and *Silhouette* .
Perform **Error-Analysis:** Identity what were the characteristics of the instance records that threw the machine off, using the top 10 frequent words and/or top collocations.
Document your steps, explain the results effectively, using graphs.
Verify and validate your programs; Make sure your programs run without syntax or logical errors.

Rubric: (accounts for 20% of the final grade.)

Choose data of your choice, (labelled data) 1%
Preprocessing and Data Cleansing 1%
Feature Engineering 2%
Use K-means, EM, Hierarchical etc. clustering algorithms 3%
Perform Evaluations, 2%
Compare and decide which clustering result is the closest to the human labels 2%
Perform Error Analysis, 3%
Perform Visualizations, Graph the results 1%
presentation 2%
Report, detail explanations 3%