Clustering Wikipedia Articles Through Topic Modeling
CMPE 256 - Large Scale Analytics
Group 4

Team:

Swayam Swaroop Mishra (013725595)
Aashay Mokadam (012724998)
Karthik Munipalle (013854867)

Data Set:

Wikipedia 2019 Dump (Total Word Count - Unprocessed Text: 2,968,288,446):
https://dumps.wikimedia.org/enwiki/20190920/enwiki-20190920-pages-articles.xml.bz2
This dataset shall be imported and processed using Gensim library modules before application. Raw text from Wikipedia page bodies will be used to cluster articles based on semantic and topical similarity, while meta-data (such as the page title and "See also" section) will be used in order to evaluate our model performance.

Goals:

Availability of number of articles in wikipedia is one of the examples of text explosion on the Internet. Due to the presence of large number of wiki articles, many experiments can be performed using them and article categorization is one of the basic experiments that can be done.

Categorizing articles has lots of advantages like searching articles. Standard keyword search are ineffective because they treat the article as a collection of unrelated words without considering the articles structure and meaning as a result many irrelevant articles are returned.

Description of Methodology:

For the purposes of our task, we shall use standard text-preprocessing techniques (cleaning and tokenization, stemming, lemmatization, stop-word removal), followed by **TF-IDF** decomposition for the purposes of document clustering. **K-Means Clustering** will be employed for this task. We shall also investigate the application of Topic Modelling through **Latent Dirichlet Allocation [3] (LDA)**, as well as refined dimensionality reduction using Singular Value Decomposition or Principle Component Analysis.

Reference:

[1]. Paul Thompson, Jacob Carter, John McNaught, Sophia Ananiadou, "Semantically

enhanced search system for historical medical archives” 2015 Digital Heritage
[2]. Dani Gunawan, Amalia Amalia, Indra Charisma, “Clustering articles in bahasa Indonesia using self-organizing map” 2017 ICELTICs
[3] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John (ed.). "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993