Title

Clustering on the 20Newsgroup Dataset

Domain

The goal of the project is to parse the newsgroup messages of 2-3 categories of the 20Newsgroup dataset, apply word2vec technique and attempt to run some form clustering on them.

Data

Starting from the original 20 Newsgroup Dataset
(http://qwone.com/~jason/20Newsgroups/20news-19997.tar.gz)
the following data would be extracted:

Variable Name	**Type**	**Description**
Path	String	File System Path
Category	String	Newsgroup Category
Filename	String	Post number
Raw Text	String	Raw text of the message
Modeling Text	List of Strings	Unigram and Bigram of the text cleaned by header, footer and quoted text

Pipeline:

- Extract 3 categories from 20Newsgroup dataset
- Produce Modeling Text
- Evaluate Clustering through K-Means

Known Unknown

- Dimensionality Reduction