CS 484: Application Research Project Proposal

Clustering Validation on Textual Dataset

Jean Michel Rouly & Joshua Wells

April 15, 2014

1 Problem Statement

One of the most common types of data on the Internet today is human-generated textual data. This data comes with many different internal patterns and possible categorizations. Identifying patterns and subsets within an otherwise unknown large dataset is a difficult problem, addressed namely by the application of clustering algorithms.

There are a great deal of clustering algorithms which can be applied to these datasets, many of which have been implemented using scientific Python libraries. Validating the performance and overall quality of these algorithms can be difficult on large data sets, especially as many of these data sets do not have a connected ground-truth.

Current state-of-the-art techniques include K-Means, Affinity Propagation, Mean-Shift, Spectral Clustering, Hierarchical Clustering (multiple techniques), DBSCAN, and Gaussian Mixtures. Many of these algorithms have pros and cons over different data set structures and patterns. We would like to determine how these algorithm classes apply to large data sets of human text with a great deal of variation in size of clusters and a large number of total clusters.

2 Datasets

URL	Name	Size
archive.org/details/stackexchange	StackExchange Data Dump	184 GB

2.1 StackExchange Data Dump

StackExchange is a major online community where users can post questions within some field, and other users respond and answer. Posts in each field range in complexity from general questions to sophisticated, technical inquiries. There

are several thousand posts in each category (around 20GB data average, ranging up to several hundred) and 230 class labels (*i.e.* different fields).

3 Procedure

- Preprocess the data
  - Parse out questions from data set, generate flat file structure
  - Generate TF-IDF vectors from data points, generate similarity matrix (using `scikit-learn`)
- Perform analytics
  - Run multiple clustering algorithms (those listed above) and keep track of data point cluster membership. `scikit-learn` will be useful in this process
- Evaluate performance
  - Using `matplotlib` generate visualizations
  - Perform accuracy metrics against ground-truth labels from the initial data set

4 Evaluation Methodologies

Because clustering is an inherently unsupervised process, it can be difficult to evaluate performance. Additionally, these algorithms can be applied to a variety of different structures of data with varying results. We will evaluate algorithmic performance over a particular type of data commonly found on the Internet by clustering a data set and analyzing performance accuracy against a ground-truth set of labels.