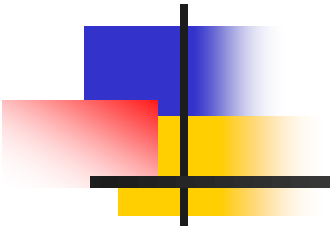

Text Cluster and Text Topic

Dr. Goutam Chakraborty

Outline

- Quick overview of clustering algorithms in SAS EM
- Differences between clusters and topics from a corpus
- Demonstration of clustering using text
- Demonstration of topic extraction using text

Why Cluster Documents?

- *Cluster analysis* is the generic name for a wide variety of procedures that can be used to create a classification of entities/objects.
  - In text analytics, the entities/objects are documents.
- The essence of all clustering approaches is the classification of data as suggested by "natural" groupings of the data themselves.
  - If we can group documents into clusters, that will help generate insights into the content of the documents

Clustering Algorithms

- Many variants of clustering algorithms for analyzing numerical data have been developed by researchers from fields such as statistics, biology, medicine, psychology, and data mining.
- Clustering algorithms can be broadly divided into four groups:
  - **Hierarchical**
  - Non-hierarchical (or partitional such as *k*-means),
  - Probabilistic (or, spectral density such as E-M, or **Expectation-Maximization**)
  - Neural network (SOM/Kohonen).

Probabilistic Clustering

- The E-M algorithm assumes that the variables are normally distributed in each cluster. Then, it applies an iterative optimization to estimate the probabilities for each observation to belong to each cluster.
- In the Expectation (E) step input partitions are selected similar to the *k*-means technique. In this step, each observation is given a weight or expectation for each partition.
- In the Maximization (M) step, the initial partition values are changed to the weighted average of the assigned observations, where the weights are those identified from the E step.
- This cycle is repeated until the partition values do not change significantly as identified by the log likelihood of the iteration.

Clustering Documents in SAS EM

- Assign each document to a cluster such that documents within the cluster are similar but documents between the cluster are dissimilar
  - Similarity may be operationalized either as distance or, cosine
- In the distance-based methods, *dissimilarity* is conceptualized as the ***distance*** between objects.
    - If two things are similar, the distance between them must be small. If two things are dissimilar, the distance between them must be large.
- In the Text Cluster node of SAS EM,
  - *Euclidean distance* is used to measure distance between clusters in the **hierarchical clustering algorithm**.
  - *Mahalanobis distance* is used in the **Expectation-Maximization (default) algorithm** to measure distance between a document and a cluster.

Text Topics

- A *topic* is a collection of terms that define a theme or an idea.
  - A topic can be derived *automatically* or, *custom-defined* by the analyst.
  - A document may contain zero, one, or many topics that are combinations of words of interest in the analysis.
- What's the difference between text clusters and text topics?
  - In text clustering, each document is assigned to only one of the mutually exclusive clusters.
  - In a text topic, each document may be assigned to multiple topics.

Plan of Analysis

- **Case:** Analyzing Android App reviews
- **Data:** Online reviews of Android Apps (anonymized).
  - If self-rating is more than or equal to 4 stars, then it is considered a positive review. Otherwise, it is considered a negative review.
  - To be analyzed separately (strategic choice) for positive and negative reviews.
- **Primary goals of this demonstration:**
  - Apply *two different* clustering techniques on the same data and explore similarities and differences in solutions
  - Apply *default* topic extraction
  - Apply *customized* topic creation

Procedure

- Follow handout titled “Demo with Android App Reviews: Text Cluster and Text Topics”