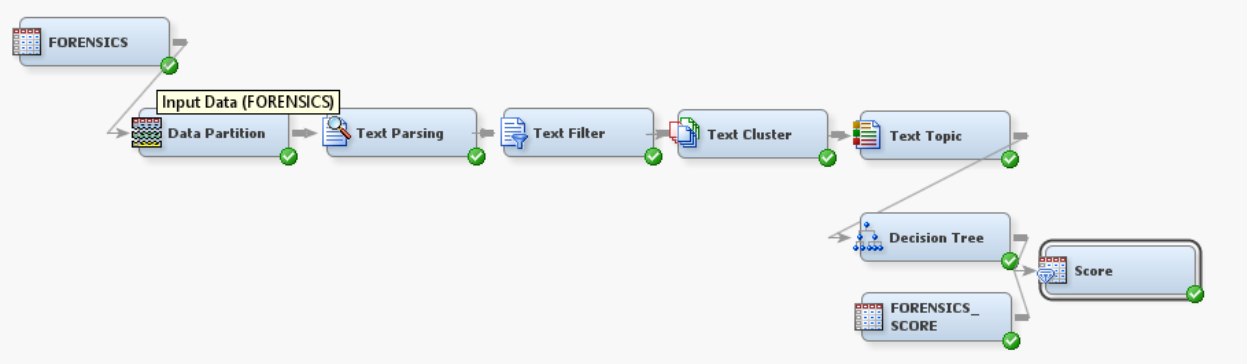
Class Assignment 8

Diane Nguyen

Part 2:

Flow Diagram



Provide a description of the problem to be solved and how text mining techniques are used in this case to solve the problem. Focus on how a textual content such as a paragraph can be used to classify a document. Detail the conversion of textual content to weights, which are in turn used as predictors in the classification. Explain why these weights can serve as predictors. You may want to do some research on your own to fully answer this question.

The Forensics data is a document containing three columns: Author, Doc ID and Text. The document text is a series of textual documents written by various authors. Through text mining, the computer program is able to identify common language, or frequency of words or phrases, used in multiple text columns in order to distinguish the author who may have potentially written the particular document. Text mining in this way uses forensic linguistics to create a predictive model in which can help in many different fields such as detective work, classification of documents in the medical or law field, ad targeting in marketing, etc. The problem to be solved is finding a way to classify the different author's works based on their different writing styles.

Within the text mining process, using a computer program, one would scan through the corpus and identify all of the potential words or phrases. Next, identify "stop words" in which you may toss out such as "a", "the", "and". You would also need to manually scan through this list and eliminate any words that would skew results such as a particular business name that includes the word "tooth" when looking through many documents dealing with dentistry. The output you are left with is a term matrix that reveals the frequency of these independent word or phrases.

In terms of textual context, a term that appears very frequently, or in all documents, would be considered a low weighted term. Similarly, a very rare term would also be considered a low weighted term. Weights can be used as predictors or indicators of different concepts because they can be grouped together. This grouping can give individual word weights a concrete meaning. Finding a terms' weight is very important in preparing for the predictive model. Local weight is determined by the frequency of the term in a document not corpus. Global frequency –

inverse document frequency (GF-IDF): builds on the previous measure, multiplying the global frequency (or how often a term appears in the corpus) by IDF. Different unique styles of writing in the document can be used to classify the document.

After completing a term matrix and determining weights, other dating mining methods such as clustering can be used to identify clusters of documents in which a particular document may belong.

Latent semantic indexing is really just a simple classification problem where each document is its own category, the training data for the category is the text extracted from the document, and where queries are just texts to classify. Many researchers have applied SVD to other classification problems. The hope is for a degree of robustness arising from the lower-order representation of terms and documents