Nate Bitting
CIS435 - Individual Assignment 5
June 1, 2014

Management Problem

The goal of this assignment is to identify themes within and across blog postings from various websites. We will apply common text analytics techniques to solve this problem through unsupervised machine learning algorithms. As with most unstructured data sets, we wanted to identify patterns or relationships from one blog to another.

Approach

Based on the data provided for this assignment, the approach we took was to tokenize the documents leveraging the Natural Language Tool Kit (NLTK) to break up the words from each of the blog postings into arrays. We then leveraged the TD-IDF vectorizer to calculate the term frequency within all blog postings. This technique will help us identify frequently used words in order identify common themes across all blogs.

The next step in the process is running this data through an unsupervised learning model, which in this case we've selected to use a normal k-means clustering model from the sci-kit learn python library. For this assignment, we chose to six clusters in order to minimize the overlapping of different clusters by ensuring the silhouette coefficient was greater than 0.2. We chose to use k-means to identify the themes or relationships between and among the various blog postings by showing the number of documents from each blog in each cluster. The last step is to look at the blog postings in each cluster to identify what themes might be present in each cluster.

Conclusion

Based on the clusters identified from the k-means clustering algorithm, we have identified the following matching words within each of the six clusters:

- Cluster 1 (Blogs 6, 7, 10) - Importance of Big Data & Web Analytics
- Cluster 2 (Blogs 2, 4, 5, 7, 10) - Marketing, Social Media & Google Analytics
- Cluster 3 (Blogs 2, 4, 5, 10) - Need for tools to drive greater optimization in the Sales process & Marketing Analytics
- Cluster 4 (Blogs 1, 2) - Businesses need to take advantage of analytics to learn more about the behavior of their users/customers (Amazon as a model) to increase growth
- Cluster 5 (All Blogs) - Data/Web Analytics, Google Analytics, Email, Social, Marketing
- Cluster 6 (Blogs 2, 5, 6, 10) - Mobile/Social Web Analytics

As you can gather above, there are some obvious overlapping between the clusters. The reason we see that is because we know all blogs are about web analytics in general and how people are using it as a competitive advantage in the market today. Text analytics is a very valuable tool from marketers to investors and can be used for a wide variety of use cases including sentiment analysis, customer preferences, product placement, etc.

Attachments

- Clusters Chart (charts/clusters.png)
- Top 25 Words for each blog (charts/blogs1-10.png)
- Console Output (console_output.rft)
- Cluster Analysis (data.xlsx)
- Python Script (NateBitting_Assignment5.py)