BUS5WB - Data Warehousing and Big Data
Assignment 03: Big Data Analytics
Marks: 30%
Assignment Type: Individual
Release Date: Thursday 3rd May 2018
Due Date: Sunday 3rd June 2018

The third assignment focuses on Big Data analytics on unstructured text data using Microsoft Azure. You are required to derive insights by applying big data distributed processing and machine learning techniques.

Dataset – TripAdvisor reviews

The dataset contains ~10000 reviews of hotels in Vietnam. The fields are;

Field name	Description
PostId	Unique ID for each review
Subject	The heading of the review
Rating	The rating given by the user, ranging from 1 to 5
Hotel	Name of the hotel
Hotel-Star	Star rating of the hotel
Review	Textual description of the review. The review field text has been pre-processed by removing non-alphanumeric characters and reduced the size to be max of 1000 characters.
Sentiment	Sentiment value extracted for the 'review' using text analytics API in Microsoft Cognitive Services. Sentiment score ranges from 0 to 1, 0 being negative and 1 being positive. (You may verify the sentiment score with the subject and review fields.)
Sentiment Category	Sentiment class generated based on the sentiment value. < 0.9: Very positive < 0.75 and <= 0.9: Positive <0.5 and <= 0.75: Neutral <= 0.5: Negative

What you are required to do

1. HDInsight to aggregate reviews

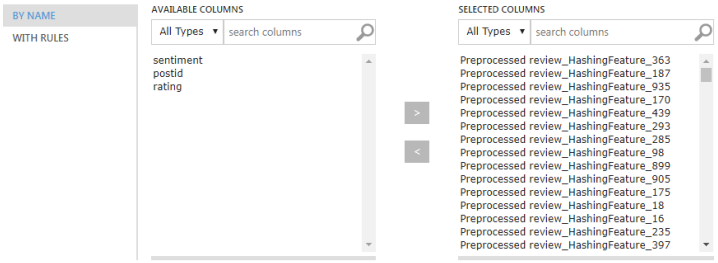
Develop an aggregate of these reviews using your knowledge of Hadoop and MapReduce in Microsoft HDInsight.

a) Follow the same approach as the Big Data analytics workshop (using wordcount method in HDInsight) to determine the contributory words for each level of rating and sentiment category.
b) Present the workflow of using HDInsight (you may use screen captures) along with a summary of findings for each level of rating and sentiment category. MapReduce documentation for HDInsight is available here.

IMPORTANT: STUDENT ACCOUNTS HAVE LIMITED AZURE CREDITS. YOU MUST CREATE AND DECOMMISSION (DELETE) THE HDINSIGHT CLUSTERS EACH TIME YOU ATTEMPT THE ASSIGNMENT. IF YOU ARE PLANNING TO WORK ON THE ASSIGNMENT ACROSS MULTIPLE DAYS, REMEMBER TO DELETE AND RECREATE EACH TIME.

2. Azure Machine Learning for sentiment analysis

Using Azure ML Studio to cluster user reviews based on sentiment score. For text clustering, you should use the 'review' field. In Filter based feature selection module, use 'sentiment' field in order to cluster reviews based on sentiment score. Download the cluster outputs into a csv file to interpret the results and derive insights. You will need to calibrate algorithmic parameters by using different Number of Centroids and Distance Metric to derive meaningful clusters. Exclude sentiment, rating or postid as selected columns to train the clustering model. Use only the preprocessed hashing features.



Provide the following,

a) A screen capture of the completed model diagram.

b) Details of parameters used for 1) feature hashing module, 2) filter based feature selection module and 3) K-Means clustering module

c) Details of the approach you chose for clustering and interpretation of clusters.

3. Findings

Summarise your findings from 1) and 2), on user rating, hotel rating and sentiment towards accommodation options in Vietnam. Consider the challenges you faced in conducting Big Data analytics on a real-life text dataset.

<u>Deliverables</u>

1. A report on the three activities.
   - The report should be compiled in Microsoft Word only, font size 11.
   - Report should not exceed 10 pages. Diagrams, tables and any other visualisations/ screen captures should be in the main body of the report.
   - Make realistic assumptions on missing information and state these in the report.
2. A compressed folder of data files that would be useful to assess your work.

Criteria	Pass	Credit	Distinction	High Distinction
HDInsight to aggregate reviews 5 marks	A minimal attempt at using HDInsight	A basic attempt at using HDInsight	A good attempt at using HDInsight	A complete attempt at deriving insights using HDInsight.
Azure Machine Learning for sentiment analysis 10 marks	A minimal attempt at clustering and analysis.	A basic attempt at clustering and cluster analysis.	A good attempt at text clustering and cluster analysis.	A complete attempt at clustering and cluster analysis.
Findings 10 marks	Basic summary of findings.	A fair effort that captures some of the key findings.	A good effort, accounting for most potential findings.	A comprehensive effort, accounting for all findings and further analysis.