Peoples' Pulse

Big Data Project Code 36 – SRS Document

IBM Career Education

Disclaimer

This Software Requirements Specification document is a guideline. The document details all the high level requirements. The document also describes the broad scope of the project. While developing the solution if the developer has a valid point to add more details being within the scope specified then it can be accommodated after consultation with IBM designated Mentor.

Table of Contents

INTRODUCTION 1
Development Environment 1
System Users 1
Assumptions 1

REQUIREMENTS 2
Sentiment Analysis 2
Sentiment Polarity Model 2

DEPLOYMENT MODEL 4

PROJECT TIPS 5

DATA SOURCING GUIDELINES 6
Twitter dataset Structure 6

TESTING GUIDELINES 7

SUGGESTED READING 8
Tools 8
Probability & Statistics 8

INTRODUCTION

A Research and Innovation wing of leading university wanted to uncover insights buried in the millions of daily online conversations. To understand what insights advertisers could gain from online conversations, the lab launched a research project that analyzed social media posts related to various subjects. The first assignment they took up was to determine how movies would open based on the social buzz? The ability to understand public sentiment in real time is very predictive of how a movie would open and what advertising worked. One of the biggest challenges in conducting sentiment analytics on social conversations is the amount of data. “How do you analyze it when you’re at the end of a fire hose?” Through a sentiment-analytics project the scholars intend to analyze millions of tweets, Facebook posts and other social media conversations to help uncover trends in near-real time. For this project, only twitter feed is being considered for sentiment analysis.

The solution will be developed using MapReduce paradigm and subsequently deployed on IBM Bluemix, a PaaS platform on Cloud providing IBM Analytics for Hadoop service. This document is the primary input to the development team to architect the proposed sentiment polarity model for this project.

Development Environment

The development will be carried out using Eclipse Version 4.2 or above. The IBM InfoSphere BigInsights Eclipse tools will have to be added to your Eclipse development environment. These tools will simplify development and deployment of applications to the BigInsights server using Java MapReduce, JAQL, Pig and Hive. They also support developing text analytics programs, such as extractors, that run on IBM InfoSphere BigInsights.

System Users

The research team shall use the outputs of the sentiment model to further execute advanced analytical models for classifying the type of sentiments.

Assumptions

1. The output generated from this project would be in JSON format.
2. A reading on sentiment analysis and various ways it can be accomplished is important.
3. The source data links being shared is a pre-processed dataset used in academic projects for training the advanced sentiment model. The developer of this project may validate the outcome the model being developed against the outcomes already included in the dataset.

REQUIREMENTS

It is required to create a sentiment polarity model for determining the sentiment score of a tweet. These tweets could be associated with a #tags which specifies the context in which the tweet is being posted. The model is expected to classify a sentence in a tweet as 'Positive' or 'Negative' or 'Neutral'. The outcome of this model will become the primary input for analyzing the classified tweets to draw inferences on the subject matter. The sentiment model helps process, filter and analyse the millions of Twitter messages as the data streams in, and uses natural language processing (NLP) capabilities to determine whether each tweet is positive or negative. The essential concepts of such a model are explained below:

Sentiment Analysis

Sentiment Analysis is a research topic today with the rise of content flowing from social media and various web channels. Humans have been doing this since ages, now it's the turn of machines to analyze what's the mass opinion on a given topic. For example, TV today is measured the same way as it was measured 50 years back. The current ratings system just tells you what channel is turned on across the subscriber base.

It tells nothing about whether the viewers are engaged and what they feel about the programs. What happens when one begins to analyze a million tweets around a piece of programming? Consider, the Oscars and if the social media buzz is analyzed, one can see what people thought about the programming and the advertising, every segment of it, in real time. What this means is that producers can actually see where the sentiment turned south in their programs, and then look at the tweets to understand what viewers didn't like. In some cases, the outcome of a reality show could be changed by audience response.

While there are many approaches being pursued to arrive at sentiment analysis, it has been decided to use a simple "Sentiment Polarity Model" to get started quickly. Intuitively, this makes sense, as the current problem on hand requires analyzing tweets of various people on a subject to arrive at the type of tone it carries i.e. positive, negative or neutral views.

Sentiment Polarity Model

The objective is to find the affinity with a particular type of tone in the tweet being analyzed. The following guidelines will aid in determining polarity of the sentences in a comment:

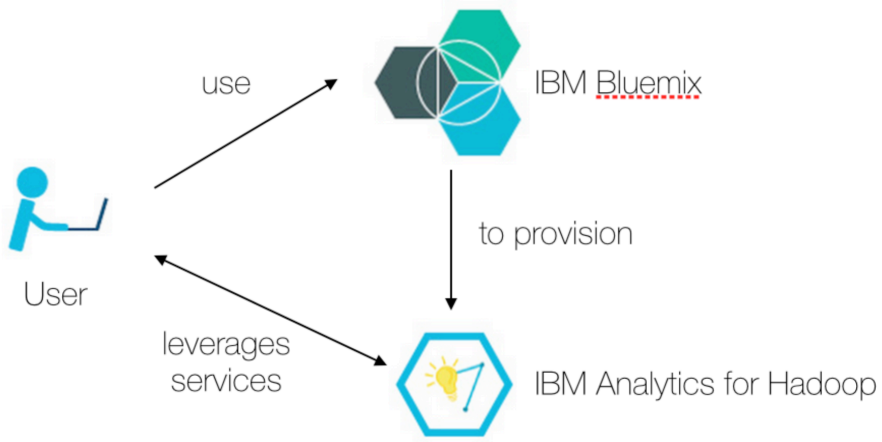
1. From the list of tweets; each line of a tweet text is taken up independently for review.
2. After removing the stop words, run the balance words through the positive words dictionary. Number of successes at this stage will add up the positive score by 1.
3. The same process is followed with negative words dictionary. Number of success at this state will add up the negative score by 1.
4. For each tweet now there are positive and negative points. If the positive and negative points add up to 0, the comment can be tagged as 'Neutral' . If the

positive points are > negative points, the comment can be tagged as positive, else it will be tagged as negative.

To summarize, the target sentiment polarity model will *generate an assessment of positive, neutral or negative emotions* for the twitter feed being analyzed using the dataset. This in turn will become the primary input for advanced sentiment analysis.

DEPLOYMENT MODEL

The deployment model is outlined below.



Once the IBM Analytics for Hadoop instance is provisioned, the available service can be easily used starting from simply uploading a file, running a MapReduce code, Big Sheets, and many more.

This project will primarily require uploading meme data and running a MapReduce program (to be developed) that will generate the “sentiment polarity model”.

PROJECT TIPS

Big Data Problems may sometimes “appear” to be very simple; and one may be tempted to solve them with traditional methods. For example, counting frequency of occurrence of every word in documents. This is indeed a simple problem as long as documents are not “too many” and are not arriving “too frequently”. Now imagine there is a stream of millions of documents coming in! Clearly with traditional methods, it will be difficult to match the processing speed with data arrival speed (velocity), volume and on occasions its variety. Therefore, focus on scalable algorithms, smart visualizations, and requisite knowledge of math - especially statistics will be critical to success.

DATA SOURCING GUIDELINES

Big data solutions solve problems by ingesting extremely large of volumes of data for various operations to be carried out on them before the results are shared with the end user or the stream of output is generated for another application's input.

The following guidelines would come in handy to source the data for your projects.

Twitter dataset used for academic purposes has been shortlisted for this project. You may download from the following links:

1. Tweets training dataset - http://stanford.io/1mxxOnl
2. Dictionary of Positive and Negative words can be downloaded from http://bit.ly/1Rv5SwF URL.

Any commercial use of this data is strictly prohibited.

Twitter dataset Structure

The data has been processed so that the emoticons are stripped off. Also, it's in a regular CSV format.

0 - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)

1 - the id of the tweet (2087)

2 - the date of the tweet (Sat May 16 23:58:44 UTC 2009)

3 - the query (lyx). If there is no query, then this value is NO_QUERY.

4 - the user that tweeted (robotickilldozr)

5 - the text of the tweet (Lyx is cool)

Only **column number 5** is to be used in this project.

The required content can be saved in the local disk or drop box and subsequently copy them into Hadoop File System.

TESTING GUIDELINES

It's easy to think that, if we know how to test a standard application, we know how to test the Big Data storage and application. Surprisingly so, it's not the case! Volume, Variety and Velocity of data make things really complex to test. While testing, mostly you are not dealing with structured data with a fixed schema; mostly the data is unstructured and a loosely defined or dynamic schema. The rate at which data is generated clearly exerts a pressure on speed of processing. Following must be kept in mind while planning the testing:

1. Plan on unit testing early and frequently during development. This is simply because big data testing is challenging, you may not be able to view source data using spreadsheets owing to sheer magnitude of the data.
2. Do not rely on eyeballing data or outputs as mechanism for verification. Create Test plan for each data set and the transformations stages it will go through in the entire process.
3. Big Data developers and testing team have to work with' Unstructured or Semi Structured' data (Data with dynamic schema) most of the time. Thus the testing activity requires additional inputs on 'how to derive the structure dynamically from the given data sources' from the business/development teams.
4. When it comes to the actual validation of the data, considering the huge data sets for validation, 'Sampling' strategy comes to rescue. But even that is a challenge in the context of Big Data Validation. This provides a tremendous opportunity for the testers who are innovative and who would go the extra mile to build the utilities that can increase the test coverage of BIG Data while increasing the test productivity as well.
5. The testing process should be strengthened on reuse and optimization of the test case sets, otherwise due to sheer size of the requirements to be tested will become unmanageable.

SUGGESTED READING

The project is aimed at making the student understand concepts of (a) Design and Development using IBM Analytics for Hadoop, IBM Infosphere Biginsights, Bluemix platform; and (b) Concepts and use of algorithms, models or visualisations for Big Data problems.

Tools

The following reading reference is easy to understand and should be read to get a clear understanding of capabilities of the tools and how you would leverage them to execute a project.

Resource	URL
IBM BigInsights Knowledge Center	http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.welcome.doc/doc/welcome.html
IBM InfoSphere BigInsights for Hadoop Community	https://developer.ibm.com/hadoop/
InfoSphere BigInsights Quick Start Edition	http://www-01.ibm.com/software/data/infosphere/biginsights/quick-start/tutorials.html
IBM Bluemix Dev – Hands on with Hadoop in Minutes	https://developer.ibm.com/bluemix/2014/08/26/hands-on-with-hadoop-in-minutes/

Probability & Statistics

Big data problems require an understanding of Probability and Statistics, which is pre-requisite for most modeling exercises. You may use your own reference content for solving the problems or may refer to the fundamentals from the following links.

Resource	URL
Introductory Statistics: Concepts, Models and Applications	http://www.psychstat.missouristate.edu/sbk00.htm
Statistical Thinking for Managerial Decisions	home.ubalt.edu/ntsbarsh/business-stat/opre504.htm