BIGDATA ENGINEERING

Critique Mining

BalaManikandan Gopalakrishnan - Rachan R Hegde - Malika Thapa

Contents

Introduction

In this report, we will look at the big data system set up and the steps involved in processing of E-commerce reviews to identify the emotions involved with products and their reviews

Data Set

The data set consists of reviews collected from amazon electronics. The data set was grouped into 2 categories. One set consists of the metadata information about the product (like product name, brand, category …) and the second one consists of the actual reviews and basic reviewer information. The volume of data under consideration is ~10GB (> 10 Million Records)

There are two major applications for such system

- ***Exploratory Data Analysis*** – To understand the trend and distribution of products
- ***Prediction*** – To design a mathematical model based on reviews and concepts of LIWC to predict the emotions associated with the product

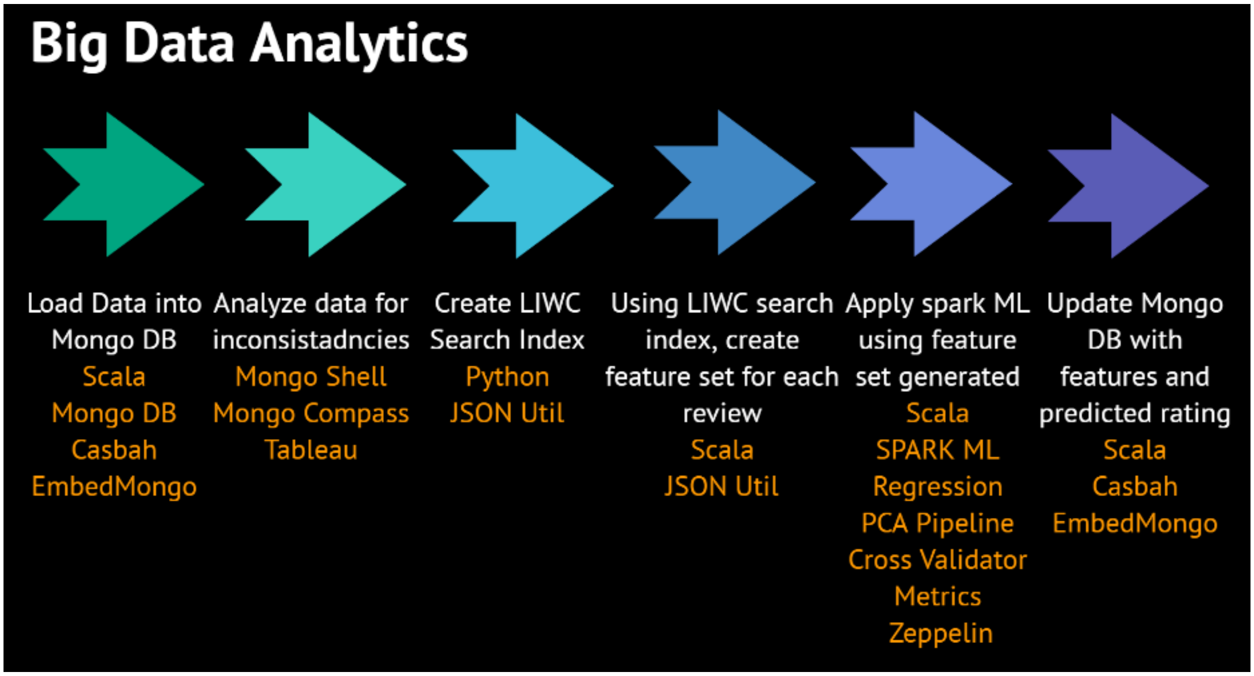
LIWC

LIWC is the gold standard in computerized text analysis. Learn how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations. Based on years of scientific research, LIWC is more accurate, easier to use, and provides a broader range of social and psychological insights.

Use cases

- Actor: Data Engineer/ Scientist
    - Load data into Mongo DB (NoSQL Data Loader)
    - Convert LIWC dictionary into TRIE (Search Index Generator)
    - Create Feature Set for ML (LIWC Feature Generator)
    - Apply ML on feature set to predict actual rating (ML Pipeline/ Zeppelin)
- Actor: Business Executive
    - Analyze brand performance based on average review rating and categories (Tableau)
- Actor: Inventory Manager
    - Analyze product inventory and brand distribution and pricing (Tableau)

Approach



Big Data Analytics

Load Data into Mongo DB
Scala
Mongo DB
Casbah
EmbedMongo

Analyze data for inconsistadncies
Mongo Shell
Mongo Compass
Tableau

Create LIWC Search Index
Python
JSON Util

Using LIWC search index, create feature set for each review
Scala
JSON Util

Apply spark ML using feature set generated
Scala
SPARK ML
Regression
PCA Pipeline
Cross Validator
Metrics
Zeppelin

Update Mongo DB with features and predicted rating
Scala
Casbah
EmbedMongo

- The data as discussed above was in JSON format. The data had to be converted into strict JSON format before it could be inserted into Mongo DB for exploratory data analysis
- After importing the data into Mongo DB, analysis was performed to handle data inconsistencies
- LIWC 2007 dictionary was converted into TRIE format which is used as a search index during feature set generation which is discussed below
- Then each of the review was broken down into words and was compared against the TRIE index generated. LIWC comes along with ~68 different moods and each word is compared with the dictionary and mood counter is incremented. These moods act as a feature set to predict the LIWC score
- After generating the feature set and the score, it was updated back to Mongo DB for visual analysis

Highlights

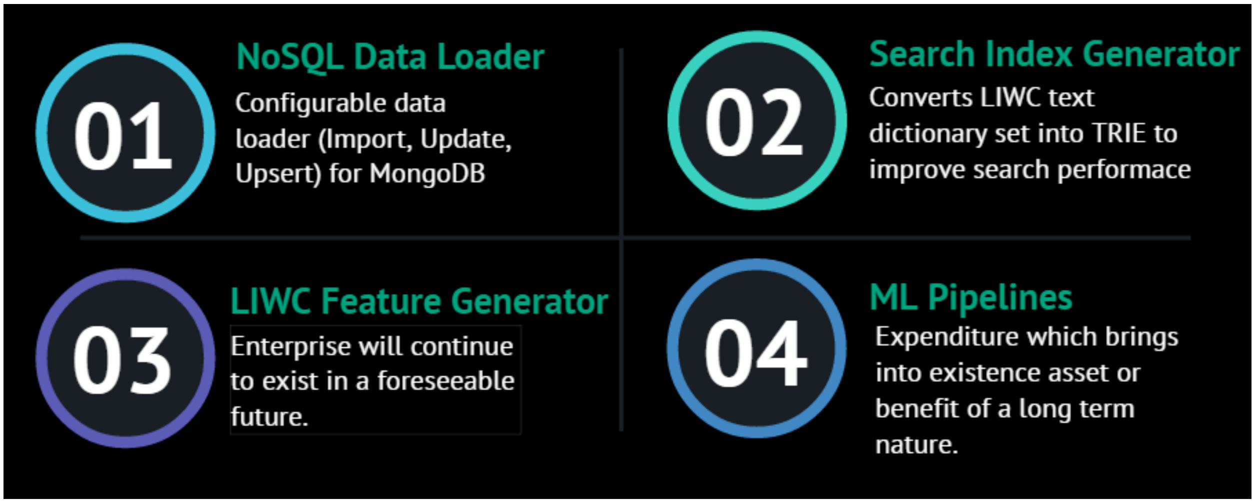
- The modules performing these tasks are completely independent (They can be reused)
- Functional programming aspects are highly followed
- The entire system was set up in EMR (with 1 Master, 9 Slaves, Mongo DB, SPARK, Zeppelin)
- Fully configurable ML pipeline
- Parallelized using SPARK
- Dashboards using Tableau
- The system is highly distributable

Acceptance Criteria

- Data loader module to parse JSON data and to import, update, upsert data into Mongo DB
- Text analyzer module and feature set creation for ML (Map review text into one of the LIWC dimensions)
- Enable business users with dashboard to search for better performing brands in select categories
- Spark to parallelize the process and Spark ML to predict the star rating of reviews
- Setting a distributed environment using Amazon EMR to execute these process

Reusable Components

Following are some of the reusable components from the setup



01 NoSQL Data Loader

Configurable data loader (Import, Update, Upsert) for MongoDB

02 Search Index Generator

Converts LIWC text dictionary set into TRIE to improve search performace

03 LIWC Feature Generator

Enterprise will continue to exist in a foreseeable future.

04 ML Pipelines

Expenditure which brings into existence asset or benefit of a long term nature.