**PROJECT TITLE: AMAZON PRODUCT REVIEW SENTIMENT ANALYSIS**

**Project Objective:**To Analyze the sentiments of the users using Text Data.

**Abstract:**Sentiment analysis is one of the major tasks of NLP (Natural Language Processing). In this paper, our goal is analyse the sentiments of the user based on a machine learning algorithm and visualize the result based on the analytics. Data used in this study are online product reviews collected from Amazon.com.

**Pre-requisite:**

**1.Spark:**

 Download https://downloads.lightbend.com/scala/2.11.12/scala-2.11.12.tgz from this

website.

 Unpack the tar file in ~/spark-2.11.12-bin-hadoop2.7 path.

 export SPARK\_HOME=~/spark-2.11.12-bin-hadoop2.7

 Run the spark-submit with the mongodb configuration and python code.

$SPARK\_HOME/bin/spark-submit --master "local[4]" --conf

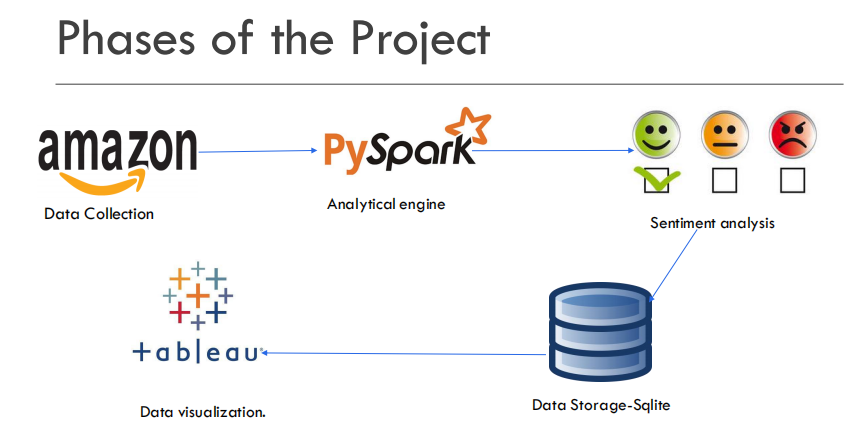
"spark.mongodb.input.uri=mongodb://127.0.0.1/Twitter.twittertext?readPreference=

primaryPreferred" --conf

"spark.mongodb.output.uri=mongodb://127.0.0.1/Twitter.twittertext" --packages

org.mongodb.spark:mongo-spark-connector\_2.12:2.4.0 spark.py

1. Anaconda Jupyter Notebook
2. Python 3.8
3. SQLite-3
4. Tableau.



The first phase is Data Collection.

The primary source of data is the review of the users from Amazon.com that is scraped using a Python library called Beautiful soup that is used to data out of HTML and XML files.

Graphical user interface, text, application, email

Description automatically generated

**Use Pandas to create a Data Frame:**

Graphical user interface, text, application, email

Description automatically generated

**Data Processing:**

After the datas are scraped from the amazon,we need to process the scraped data.The structured and semi structured data is processed using pyspark.

Pyspark is used since it is a collaboration of Apache spark and Python.It provides a wide range of libraries and is majorly used for Machine Learning and Real-Time Streaming Analytics.Here in Pyspark,the required packages are imported and we obtained a spark data frame as a result.

Text

Description automatically generated

**Spliting the data set into train and test set:**

Text

Description automatically generated

Now the obtained text should be pre-processed using the following methods

**Tokenization:**

Tokenization is the process of breaking a stream of text up into words, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Although tokenization is a slow process, with the help of Spark we can make it fast by running it in chunks/parallel.

Some important packages are imported and reviews are tokenized with

A screenshot of a computer

Description automatically generated with low confidence

Eliminating Stop Words:

Stop words are those articles in a sentence which are a bit much in any area in content mining. So we by and large overlook these words to upgrade the exactness of the analysis. In various configuration there are distinctive stop words contingent upon the nation, language and so forth In English arrangement there are a several stop words.

A picture containing table

Description automatically generated

**Hashing TF:**

Text

Description automatically generated with medium confidence

### Applying Naïve Bayesian classifier Model:

The Naïve Bayesian classifier works as follows: Suppose that there exist a set of training data, D, in which each tuple is represented by an n-dimensional feature vector, X=x 1,x 2,..,x n, indicating n measurements made on the tuple from n attributes or features. Assume that there are m classes, C 1,C 2,...,C m. Given a tuple X, the classifier will predict that X belongs to C i if and only if: P(C i|X)>P(C j|X), where i,j∈[1,m]andi≠j. P(C i|X) is computed as:

P(Ci|X)=∏k=1nP(xk|Ci)

**Result:**

An accuracy of 76% is achieved in this model.

Text, letter

Description automatically generated

**Data Storage:**

An SQLite connection has been established and the scraped data is stored into the database.

Text

Description automatically generated

**Data Visualization:**

Tableau is installed and the amazon web scraped datas are imported into the tableau for analytic visualization.

As we can see,the bar chart compares the 4 mobile phones based on the popularity .The bubble chart compares the product based on the price.

Chart

Description automatically generated

**RISKS FACED THROUGH THE TABLEAU DASHBOARD:**

The data that is stored that is the consumer analysis of the amazon products is connected to Tableau through the ODBC drivers for further analysis and visualization.

In the event that we plan the dashboard and referring view in an appropriate manner, we don't have to store colossal measures of information in Tableau. Toward the day's end clients see the visual dashboard which contains couple of GBs information (Majority of the time under 5 GB). Since, dashboard presents solidified data, for example, avg, total, estimation results, drill downs, roll ups and so forth In the event that dashboard isn't introducing RAW definite information, there is no requirement for gigantic measures of information in Tableau.

**Conclusion:**

Thus a model is developed to classify the sentiments of the user and also analysis is

done for certain popular products and data is visualized using Tableau

In future accuracy can be increased by better data preprocessing and also by using

other ML/deep learning techniques.

**References:**

Samha,Xu,Xia, Wong & Li “Opinion Annotation in Online Chinese Product Reviews.” In Proceedings of LREC conference, 2008.

Nina Isabel Holleschovsky, “The social influence factor: Impact of online product review characteristics on consumer purchasing decisions”, 5 th IBA Bachelor Thesis Conference, Enschede, The Netherlands 2015

Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." 2016

Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp„s Ratings Based on Text Reviews." (2015).

[5] Rain, Callen. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning."Swarthmore College (2013).