SENTIMENT ANALYSIS

1. Principal Investigator

Member 1 - Ashok Kumar Reddy Yannam- **yannama1@newpaltz.edu**
Member 2 - Naveen Kumar Malempati **- malempan1@newpaltz.edu**
Member 3 - Avinash Reddy Mora – **moraa4@newpaltz.edu**

Mentoring
Professor Min Chen
Department of Computer Science, SUNY-New Paltz
chenm@newpaltz.edu

1.1Individual Contribution Breakdown (list the percentage)

Task	Member 1	Member 2	Member 3	Total
Introduction	30%	35%	35%	100%
Background	35%	35%	30%	100%
Implementation	35%	30%	35%	100%
Experiment Results and Discussion	30%	35%	35%	100%
Conclusion	35%	30%	35%	100%
Other contribution and explain	35%	30%	35%	100%

2. Title of Project

Sentiment Analysis on Online product's reviews.

3. Mentoring

Professor Min Chen, Department of Computer Science, SUNY-New Paltz
chenm@newpaltz.edu

4. Introduction

4.1 Project Motivation

According to Ramteke et al. (2012) motivation for Sentiment Analysis is two-fold. Both consumers and producers highly value “customer’s opinion” about products and services. Thus, Sentiment Analysis has seen a considerable effort from industry as well as academia.

4.2 Aims and Objectives:

Primary aim of this project is to use MapReduce in creating the required data for sentimental analysis from the product reviews text information. MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chucks and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

The goal of this project is to predict the sentiment of a given text using python where we use NLTK aka Natural Language Processing Toolkit, a package in python made especially for text-based analysis. So with a few lines of code, we can easily predict whether a sentence or a review (used in the blog) is a positive or a negative review.

5. Background/History of the Study

1. **Hadoop:**

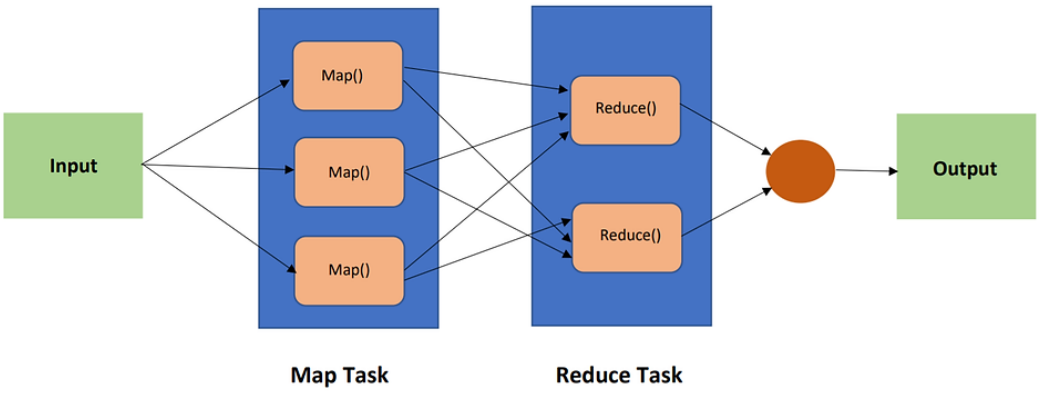
   Hadoop is an Apache open-source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

2. **Map Reduce:**

   MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers.

   Map step: The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. The worker node processes the smaller problem and passes the answer back to its master node.

Reduce step: The master node then collects the answers to all the sub-problems and combines them in some way to form the output the answer to the problem it was originally trying to solve.



3. MRJob Library:

mrjob is the easiest route to writing Python programs that run on Hadoop. If you use mrjob, you'll be able to test your code locally without installing Hadoop or run it on a cluster of our choice.

Additionally, mrjob has extensive integration with Amazon Elastic MapReduce. Once you're set up, it's as easy to run your job in the cloud as it is to run it on our laptop.

Here are several features of mrjob that make writing MapReduce jobs easier:

- Keep all MapReduce code for one job in a single class
- Easily upload and install code and data dependencies at runtime
- Switch input and output formats with a single line of code
- Automatically download and parse error logs for Python tracebacks
- Put command line filters before or after your Python code

If you don't want to be a Hadoop expert but need the computing power of MapReduce, mrjob might be just the thing for us.

4. Sentimental Analysis:

Sentiment analysis is a vital topic in the field of NLP. It has easily become one of the hottest topics in the field because of its relevance and the number of business problems it is solving and has been able to answer.

Essentially, sentiment analysis or sentiment classification fall into the broad category of text classification tasks where you are supplied with a phrase, or a list of phrases and your classifier is supposed to tell if the sentiment behind that is positive, negative or neutral. Sometimes, the third attribute is not taken to keep it a binary classification problem. In recent tasks, sentiments like "somewhat positive" and "somewhat negative" are also being considered.

5. NLTK library:

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum**.**

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3.

6. Matplotlib:

Matplotlib is a low-level graph plotting library in python that serves as a visualization utility.

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots**.**
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout**.**
- Export to many file formats **.**
- Embed in JupyterLab and Graphical User Interfaces**.**
- Use a rich array of third-party packages built on Matplotlib.

6.Approach and Implementation

This project collected reviews of a particular product from Kaggle website and data set is Amazon Alexa Reviews in a text file. The pipeline of this project is as follows:

1) Collecting data from online reviews of a product form amazon website.
2) Used MapReduce approach in collecting the required data for further analysis.
   - Here, the mapper function tokenizes the entire text file into various tokens.
   - Combiner function is responsible for counting the frequency of each distinct token.
   - Reducer takes the part of appending the word with respect to its count value.

We have used MRJob package from python to perform the above functions.

Any sentiment analysis workflow begins with loading data. But what do you do once the data's been loaded? You need to process it through a natural language processing pipeline before you can do anything interesting with it.

The necessary steps include (but aren't limited to) the following:

1. **Tokenizing sentences** to break text down into sentences, words, or other units
2. **Removing stop words** like "if," "but," "or," and so on
3. **Normalizing words** by condensing all forms of a word into a single form

4. **Vectorizing text** by turning the text into a numerical representation for consumption by your classifier

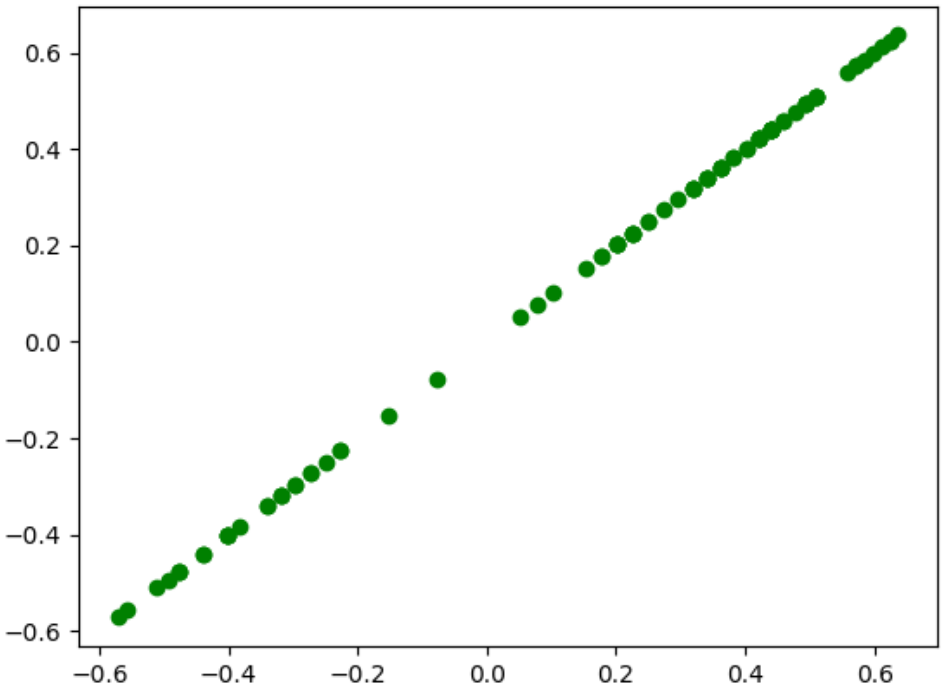
All these steps serve to reduce the **noise** inherent in any human-readable text and improve the accuracy of your classifier's results.

3) Considering the word and its frequency, in this step we are analysing the context of the reviews by classifying if the overall feedback on the product is positive or negative. This is done using *SentimentIntensityAnalyzer* function from NLTK library. We collected the polarity score of each and every token and analysed the nature of the context.

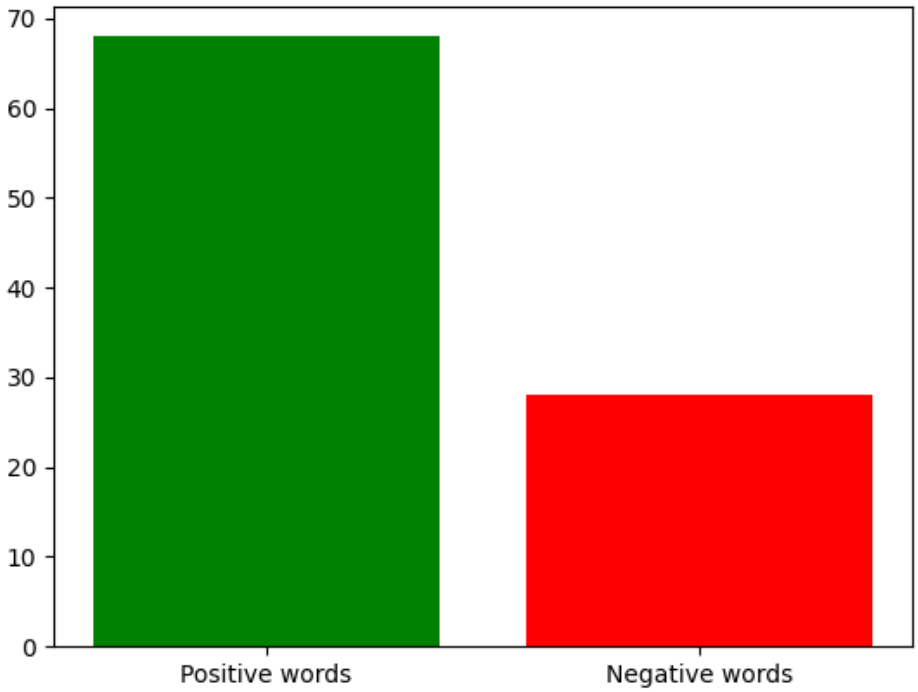
7. Experiment Results and Discussion

In this project we considered the input data from amazon product reviews and performed MapReduce and Sentimental analysis to find out whether the product is having more positive reviews or negative reviews on its own by training the data.

Results are shown in a scattered graph and a Bar graph which are showing the obtained polarity scores in the input text. -1 indicates complete negative word and +1 indicates complete positive word.



Based on the polarity scores we portrayed the differences the presence of positive and negative words through a bar graph as shown below.



8. Conclusion

In today's environment where we're suffering from data overload (although this does not mean better or deeper insights), companies might have mountains of customer feedback collected. Yet for mere humans, it's still impossible to analyze it manually without any sort of error or bias.

Oftentimes, companies with the best intentions find themselves in an *insights vacuum*. You know you need insights to inform your decision making. And you know that you're lacking them. But you don't know *how* best to get them.

Sentiment analysis provides answers into what the most important issues are. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition that isn't always right.

9. References

[1] **https://www.kaggle.com/sid321axn/amazon-alexa-reviews**

[2]https://www.geeksforgeeks.org/hadoop-mrjob-python-library-for-mapreduce-with-example/

[3]https://www.cfilt.iitb.ac.in/resources/surveys/SentimentAnalysis-Vinita.pdf

[4]https://realpython.com/python-nltk-sentiment-analysis/

[5]https://iopscience.iop.org/article/10.1088/1742-6596/1187/5/052020/meta