SENTIMENT ANALYSIS ON TWITTER DATA USING HADOOP

Sentiment analysis (opinion mining) means extracting the sentiments of people using natural language text. Sentiment classification looks for instance at emotional states like angry, sad and happy. To extract the sentiments from micro blogging sites we are using a framework called **Hadoop** which is used here for testing. Hadoop is having a large storage capacity when compared to other frame works. It can read TB's of data within short span of time. In micro blogging sites we have millions of tweets. We can find the developed articles/projects for time accuracy, extracting emoticons and performing sentiments. In this article, we are going to identify the opinions of people using twitter datasets based on their tweets, Bayesian classifier helps to classify the positive, negative or neutral tweets. Sentiment analysis is used to gather the information available from social networks. Time is more precious than money, so without wasting time on reading all positive, negative, neutral feedbacks we just go for sentiment analysis to conserve our time.

Micro Blogging has become the most popular communication tool among the internet users. Millions of messages, tweets, and shares takes place every day in several sites like Facebook, twitter, tumblr and so on. Users daily used to share their personal messages, daily activities their personal opinions on various topics and also they may have some discussions on current issues. Users feel free to post their own feelings about a particular product on their own page.

But, day by day the data users are increasing and companies are generating large amounts of data, in 2011 Facebook generated 6 billion messages per day and E-bay, 2 billion page views per day. It is a complex task for RDBMS and it is not designed to handle large amounts of data. Later APACHE HADOOP was introduced to process this big data and it uses a commodity hardware. In Hadoop we can store large amount of data in HDFS (Hadoop Distributed File System). Replication factor of Hadoop is 3 i.e., it can create 3 replicas at different data node locations in orders to reduce the loss of data. It allows developers to focus on application development and business logic. After its arrival it is helpful to store huge amounts of data and processes in a very less time.

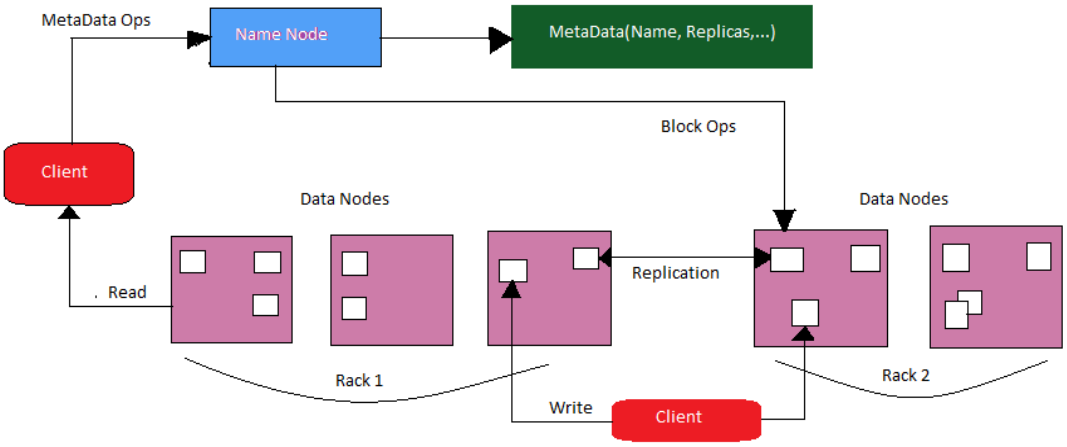

Fig 1: Architecture of Hadoop Distributed File System (HDFS)

Suppose if a company releases a new product, they would like to know the feedback of that particular product. No one is interested to give their individual reviews on that product but if the company asks for customers' opinion on any social networking site, then many users freely post their reviews on their personal pages, so organization can easily gather the feedback of the product from the users directly from here, that's why many organizations are making use of these sites.

From the feedback they can analyse what changes are to be done on their product and what are the positives and negative feedbacks on the product. They also get to know what people are expecting more on that product. These all can be easily done with sentiment analysis.

The work in this article leads to identify the number of positive, negative and neutral tweets made by users. By considering the number of the respective feedbacks we can easily identify the sentiments of them.

From the last 10 years several works has been done on sentiment analysis. Much research has been done on Twitter sentiments. Supervised classifiers like Naive bayes (NB), Support Vector Machine (SVM), Maximum Entropy (MaxEnt) are used to extract the sentimental tweets on twitter.

Naive Bayes Classifier

Naïve Bayes is a simple model for text categorization.

Class c* is assigned to tweet d, where

$c^{*}=\text{argmac}_{c}P_{NB}\ (c|d)$

$$P_{NB}(c|d):=\frac{P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d))}}{P(d)}$$

Our Approach

In this article we are going to introduce the text data of twitter by considering a dataset having more number of user tweets. We will process all the tweets and find out the total number of positive, negative and neutral tweets.

- Tweet like “Today I’m so **happy**” is a positive tweet.
- Likewise, “The hotel food is **not** at all good” is a negative tweet. Here, **not** is considered as negative tweet.
- “Yesterday I visited Switzerland. It is so **good** but because of climate my health **spoiled**” is considered to be a neutral tweet.

Tweets	**Polarity**
Fabulous	Positive
Highly valuable	Positive
Sinful	Negative
Immoral	Negative
Satisfactory	Neutral

Table: Dictionary of Emotions

By considering the polarity of words we can calculate which is having the highest priority. Consider a scenario where we are having millions of tweets on our product but we are unable to know who are liking and how many are liking the product. So we filter all the positive, negative and neutral tweets. For instance, if there are 3,000 positive tweets, 1,500 negative tweets and 1,000 neutral tweets, then we are having more number of positive tweets. Thus, we can say that the product is having good rating.

Likewise, it helps in knowing the review results of a specific product.

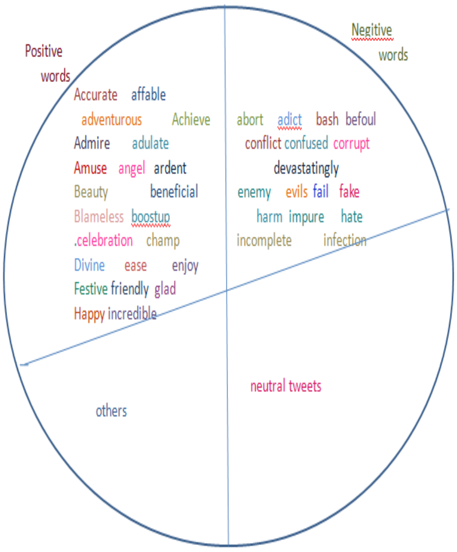

Fig 2: Sample SentiWords

Early applications of sentiment analysis were mainly focused on classifying movie reviews or product reviews as positive or negative or either positive or negative sentences.

But Opinion mining mainly focuses on detailed analysis of the sentiments expressed in texts. In our proposed scheme we are considering a twitter dataset by collecting thousands of user tweets as a sample dataset where sentiment analysis can easily calculate large amount of data in a few seconds.

Likewise we can also calculate the feedback of a particular product. In the recent days, gathering product reviews has become complex, people are not interested to give their feedback on a particular product and it is very hard for the organizations to gather the feedback directly from the people. So they are choosing the sentiment analysis approach.

Conclusion:

The code can handle and filter the positive, negative and neutral tweets, but we are not mainly focussing on time efficiency. In hadoop it is an easy task to reduce the time count of tweets. It handles thousands of tweets in less time. Generally sentiment analysis is a wide area for research. As of now we can handle up to feedbacks but we are concentrating only on textual data entered by user, but not concentrated on emotical tweets. There are no hash tags in this article. Those are left for future development.

Contact us for further details

Sowmya Sri Godavari

Hadoop Developer

sowmyag.in@mouritech.com

MOURI Tech
www.mouritech.com