# Project Proposal
# Big Data (DS-GA 1004)

Diogo Mesquita (dam740) Kenil Tanna (kyt237) Mihir Ujjwal Rana (mur214)

## 1 Project Title

Sentiment analysis of Amazon.com product reviews, and (optionally) building a recommender system using sentiment scores.

## 2 Problem Description and Goal

Sentiment analysis aims to determine the attitude of a user with respect to some topic (in this case, an Amazon product). The sentiment score represents how much a user likes/dislikes a certain product, which can in turn be a major factor in, say, deciding if (s)he won't buy a product (even if the products are similar). The goal of this project will be to perform sentiment analysis on Amazon review data and predicting the score class. Time permitting, these scores will be used as input to the recommender system to make recommendations to a user. We believe that this satisfies the project requirements and is an interesting industry-oriented problem to work on, which will be useful to us after our graduate studies.

## 3 Previous Works and References

There has been some work done on sentiment analysis for Twitter systems using a map-reduce framework, [1] describe how data is scaled at a larger scale at Twitter and use an application of sentiment analysis to show how it works. Their implementation is based on Pig which is built on Hadoop. [2] have also done big data analysis on Twitter sentiments by using Naive Bayes, and first predicting whether each work will be positive or negative and then using that to get a sentiment for the entire tweet. For Amazon review data, we couldn't find substantial work being done using a system like Hadoop, although there has been some work on sentiment analysis on subset of the data. [3] do sentiment analysis on the Amazon product review data using Naive Bayes, but they have done it using scikit-learn.
The various Collaborative Filtering methods are discussed in Chapter 9 [4]. The item-item similarity and design of collaborative filtering algorithms for Amazon product recommendations are discussed in [5].

## 4 Data Set

The (product review) data can be found here. The total size is nearly 18 GB and comprises 83.68 million rows, which is apt to be run in a map-reduce framework. For sentiment analysis, algorithms such as Naive Bayes, Logistic Regression, are parallelizable, so are user-user (or item-item) collaborative filtering for

recommender systems, which can be implemented on Hadoop/Spark since at their core, they consist of vector inner products, matrix multiplication, factorization, etc.. Every row in the review data is of the form *(reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime)*, and in ratings data is of the form *(user_id, product_id, product_rating)*.

## 5 Method and Algorithms

The input data comprising reviews (and other metadata) will be divided into train and test sets, and the following algorithms will then be used to calculate the Sentiment Scores ($\in$ [0,1]):

1. Naive Bayes Classifier
2. Logistic Regression

Time permitting, we aim to use the scores as an additional input in the ratings data for building the recommender system. To achieve this, we fully populate the Utility (user-product) Matrix using *User-User* and *Item-Item* Collaborative Filtering.
Concretely, each product rating will be predicted on a scale of 0-5 or 0/1 (i.e., like/dislike). For measuring similarity, multiple metrics such as Jaccard Distance and Cosine Distance will be explored.

## 6 Evaluation Criteria

For sentiment analysis, Accuracy will be used (by converting scores to classes (in 0-5). For collaborative filtering, traditionally, both Mean Absolute Error (MAE) and Root-Mean-Squared Error (RMSE) are used for evaluating the predictions of recommendations [6], and we will explore both metrics for our project.

## References


[1] Jimmy J. Lin and Alek Kolcz. Large-scale machine learning at twitter. In *SIGMOD Conference*, 2012.

[2] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen. Scalable sentiment classification for big data analysis using na #x00ef;ve bayes classifier. In *2013 IEEE International Conference on Big Data*, pages 99–104, Oct 2013.

[3] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, Jun 2015.

[4] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2nd edition, 2014.

[5] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003.

[6] Claude Sammut and Geoffrey I. Webb. *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st edition, 2011.