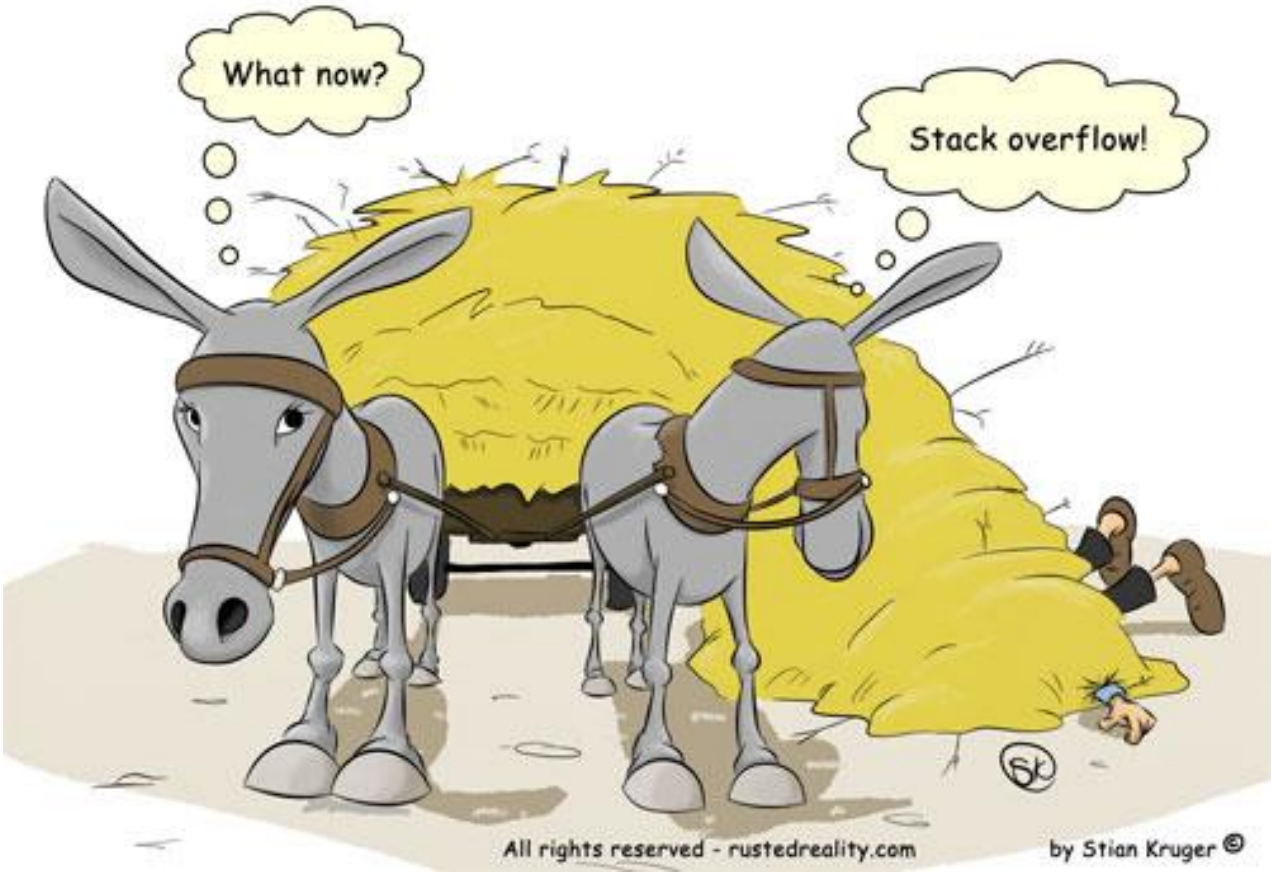
A Predictive Analytics
Project Proposal
To
Analyze Stack Overflow Dataset



Presented By:

Alok Satpathy

Aman Agarwal

Pratik Mrinal

Sankit Gupta

1. Problem Statement and Approach:

StackOverflow is a collaborative questioning and answering site designed for developers to find answers. As this online community is growing with new users joining every day, it's important to manage the content to avoid irrelevant posts like spams to ensure quality of content. Also, since one question can have multiple answers and only one answer can be marked as the best answer, it would be beneficial to automatically predict the best answer.

The dataset from StackOverflow can be organized into 5 categories - Posts, Comments, Users, Votes and Badges. The project aims at classifying each post (question/answer) into spam / non-spam and further evaluate the quality of the answer posts to predict likelihood of it being marked as the best answer. Each post has an option to upvote/downvote. While upvote increases the reputation of the owner by +5, downvote has a negative effect on reputation of both owner (-2) and the user downvoting (-1). Hence the users tend to provide a negative feedback as a response to the reply to the question, instead of downvoting it. We would like to apply Natural Language Processing (NLP) on the comments of each of these posts to understand the correlation between downvoting vs. negative comments and predict the likelihood of posts resulting into negative comments.

2. Dataset Details:

The dataset has been curated from Stack Exchange site available at the link http://data.stackexchange.com/stackoverflow/queries. It is available for sharing and reproduction licensed under Creative Commons by ShareAlike 3.0.

Stack Exchange provides a data dump of all user-contributed content on the StackOverflow, an online community for programmers to learn and share knowledge. The dataset available and used in this project was last updated in March 2016. The data was fetched via queries run on Stack Exchange Data Explorer tool. This tool facilitates querying from the huge repository and download the data as .csv file.

The original data is more than 1TB and has 188 columns distributed among 26 tables. These tables have data about posts, comments, tags, votes, users and reviews. The whole StackOverflow model is based on reliability of the posts and comments. This makes it essential for the users to know which comment should be relied upon and which should not. This is determined by the number of votes on a post. Also, StackOverflow thrives on the community collaboration and gamification lies at the core of community collaboration. The badges, points on upvotes and privileges thus are essential. This data is also available in this dataset.

For this analysis, however, we have used only part of the whole data. We used joins over some of these tables and fetched the relevant columns while carefully managing the constraint relationships among the necessary columns.

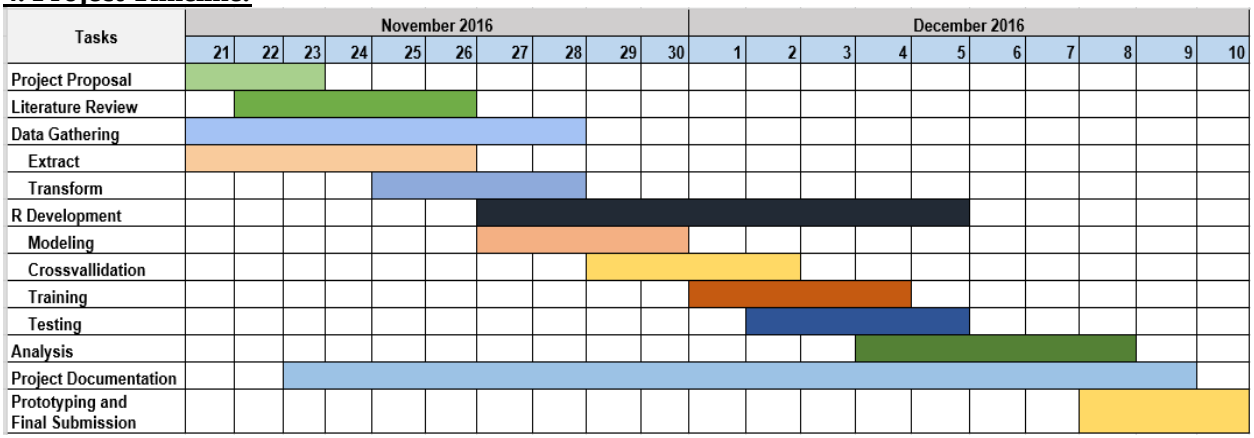
3. Possible Challenges with the R Coding:

Enlisted below are few of the challenges that are expected as part of analysis process for the given problem statements:

1. Selection of the most appropriate model, exactly detailing the dependencies between the sample features. A deviation from the most appropriate model may lead to fault in the spam filter and erroneously mark the posts as spams thereby decreasing the credibility of StackOverflow postings.
2. NLP implementation for sentimental analysis of the comments, requiring the appropriate depiction of the negative to positive emotion detection.

3. Extensive coding to extract and prepare huge data sets with high number of sample features and instances, in addition to the transformations to neglect the NA and duplicate data.
4. Shrinkage of the feature space for dimensionality reduction and related cross validation of the models to achieve an optimum balance of bias-variance tradeoff.
5. Application of the results to the real world "testing data" to achieve the minimum FDR (Type I error rate) in addition to the highest sensitivity ratio.

4. Project Timeline:



5. Roles and Responsibilities:

As a team, every member will be equally responsible to carry out the tasks mentioned in the timeline above. We will initially work independently to design models on the problem statements. Once this has been done, the team will move forward with the best models in both the segments from the 4 models designed. The team will then conduct cross validation and various testing to get the most accurate model. Additionally, every member will individually read and report the literature in this field, which will then be consolidated for a literature review report.

6. Literature Review:

1. Evaluation and Prediction of Content Quality in Stack Overflow with Logistic Regression, Daoying Qiu, December 16, 2015, Finland
2. Stack Authority: Predicting Stack Overflow Post Helpfulness Using User Social Authoritativeness, Ahmed, Kunistkiy and Maricq, November 2015, CA, USA
3. Fit or Unfit : Analysis and Prediction of 'Closed Questions' on Stack Overflow, Correa and Sureka, July 2013, Delhi, India
4. Design Lessons from the Fastest Q&A Site in the West, Mamykina, Manoim, Mittal, Hripcsak and Hartmann, ACM, 2011
5. What developers are talking about? An Analysis of Stack Overflow Data,
6. Towards discovering the role of emotions in stack overflow, Nicole Novielli, Fabio Calefato, Filippo Lanubile, Bari, Italy, May 2014
7. Analysis of Titles from the Questions of the Stack Overflow Community Using Natural Language Processing (NLP) Techniques, Tapan Kumar Hazra, Aryak Sengupta, Anirban Ghosh, August 2015
8. The Challenges of Sentiment Detection in the Social Programmer Ecosystem,Nicole Novielli, Fabio Calefato, Filippo Lanubile, Bari, Italy, May 2014