StackOverflow (Assignment 1) - Data Analytics

By
Vedant Bassi (CED15I013)
Dinesh (EVD15I002)
Sai Uday Kiran (MFD15I008)

Information gathered from : https://insights.stackoverflow.com/

More than 50 million developers visit Stack Overflow every month to learn, share their knowledge, and build their careers.

Use Cases of Data Analytics in StackOverflow

1. Classification
    a. Multi label Question classification
        i. In Stackoverflow there are 53027 tags in total.
        ii. In total there are 41622211 questions in StackOverflow as of today.
        iii. Hence on an average, it amounts to 2.54 tags per question.
        iv. To automatically attach tags to these questions using Multi Label classification, they can use several methods, including:
            1. Chain of binary Classifiers
            2. Deep Neural Networks
        v. All data has been captured from : https://data.stackexchange.com/stackoverflow/query/new

    b. User Classification for job postings
        i. Stack overflow collects yearly developer surveys to capture data about their users and understand their demographic.
        ii. They have been collecting this data for the past 7 years.
        iii. This Data helps them better understand the user so that they can serve better job suggestions for these users, and track their question and answer patterns and draw predictions between the user experience, so that they can propose better jobs to users with higher experience.
        iv. Annual Developer Survey Data is available at: https://insights.stackoverflow.com/survey
        v. They capture various dimensions while capturing this data:
            1. Geography
            2. Developer Type
            3. Contribution to Open source

4. Coding as a Hobby
5. Years Since Learning to Code
6. Years Coding Professionally
7. How Many Developers are Students?
8. Educational Attainment
9. Age, Gender
10. Platforms they code on
11. Most Loved, Dreaded, and Wanted Languages
12. Most Popular Development Environments

2. Clustering:
    a. Removing duplicate questions based on intra Cluster distance
        i. There are 41,622,211 posts on stackoverflow, after considering that they actively remove duplicate content form the website either by use bots to scan their website our by the help of moderators.
        ii. These bots read questions and encode them in a form of a feature vector based on various NLP metrics, and then these feature vectors are plotted on a higher dimensional graph and the intra cluster distance is then evaluated, if the intra cluster distance is less than a particular threshold, then the post is marked as a duplicate by the bot.
        iii. This threshold is calculated from previously marked posts by moderators and using that as training data the bots are trained.