**PROJECT PROPOSAL**

## **Problem Statement Formation**

The current problem for StackOverflow is that just like any open web forum or developer community platform, there are millions of questions posted without any moderation OR manual intervention of categorizing questions posted by the developer community, which makes it very hard for the community in general to explore which words contribute the most to which topics, and which topics contribute the most to which documents (questions on Stack Overflow, in this case).

## **Context**

Stack Overflow is a question-and-answer site for professional and enthusiast programmers.

* Every document is a mixture of topics
* Every topic is a mixture of words

Based on the problem statement, it is a complicated task to tag questions to a certain topic name since more often, the questions posted would be very open ended and every user has a different definition, semantic and syntax of what and how to post a question.

An effective data-oriented solution needs to be devised by developing models that identify and tag questions to a specific topic which could be used as a basis for future tag recommendation when another developer posts a related question OR a question on a related topic.

This would help keep the platform organized and make it a more collaborative platform for users to question and answer on topics that interest them.

## **Criteria for Success**

* Identifying tags for question text
* Predicting popularity of an answer based on number of upvotes

## Scope of solution space

Scope of the solution would be to find out tags or keywords associated to the questions, so StackOverflow can recommend and/or assign tags whenever a user is posting a question.

Build a model like this is not just for analysis; it can be used to make predictions or implement new ideas. For example, one idea for Stack Overflow would be to automatically suggest a list of possible tags for new questions based on the text of a question.

## Constraints

* The dataset is a 10% representative sample of the questions posted in StackOverflow and is not indicative of the entire population i.e., all Questions posted on Quora **(Sample bias)**
* Need to be vary of **Confirmation bias** when building the model since there could be predetermined assumption involved when building the model, so it is important to welcome inputs from multiple Machine learning and Data Science experts when trying to solve this problem.

## **Stakeholders**

* Stackflow management team
* StackOverflow Data science and Machine learning experts

## **Data Sources**

<https://www.kaggle.com/stackoverflow/stacksample>

## What is the problem you want to solve?

This is indicated in the Problem statement

## **Who is your client and why do they care about this problem?**

**Client** – StackOverflow

StackOverflow wants to provide an easy to comprehend, streamlined and user friendly interface which makes learning approachable, collaborative, effective and fun.

## **What data are you using? How will you acquire the data?**

**File descriptions**

* Questions.csv
  + Questions contains the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.
  + **Format** - CSV file that has 7 columns
* Answers.csv
  + Answers contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
  + **Format** - CSV file that has 6 columns
* Tags
  + CSV file that has 2 columns - TagId and feature tags that need to be mapped to the question text

**Data availability** – The data is available in the Kaggle community as indicated in the Data Sources section above

## **Method and Solution**

* Data Wrangling – Extraction and Cleaning, Text processing, Natural Language Processing using libraries like NLTK, SpaCy etc as needed.
* Perform Exploratory Data analysis for understanding the data to find out correlations between variables.
* Perform feature engineering by breaking the ‘questions’ dataset into training and test dataset. Use methods like tfidf on the textual variables
* Apply Multi label classification models to identify tags associated to question text.
  + Predicting tags on questions reserved for test dataset.
* Apply linear regression models to score popularity for answers by predicting number of upvotes.

## **Deliverables**

* A GitHub repo containing the work completed for each step of the project, including:
  + A project report
  + A slide deck