MLDM Case Study

Group No - 3

CB.EN.U4CSE17102	Abhishek Marpu
CB.EN.U4CSE17105	Akash G
CB.EN.U4CSE17119	Hare Sankaran RV
CB.EN.U4CSE17120	Himanshu Kumar
CB.EN.U4CSE17138	PNSS Akshay

Abstract

There are plethora of data, both structured and unstructured data getting generated each day. Recently graduated/ing engineers will find it challenging to explore new technologies as it will be time consuming to aggregate data from multiple sites. Also in a field like CS, there are plenty of outdated information which may mislead the budding developer. Our use case mainly pertains to helping and guiding those individuals using stack overflow dataset whose website is the first place every developer with problem reach out to.

Our primary goal is to perform exploratory data analysis and data mining and explore associations based on question tags and have a aggregated view of a wide range of technologies, the complex relationships among the technologies and its user base and the recent trends, which reflect the best practices followed by a large community of developers and hence giving a clear insight about the existing trends.

Problem to be solved:

The popularity of a question depends on its tags. But a new programmer in the community may lack knowledge of giving proper tags, this may lead to the question being downvoted or the question, failing to be relevant to the community to which it is forwarded. This may lead to demotivation for the new users.

Through our project, we intend to assign the tags so that the questions would be sent to appropriate experts from the domain to get the most accurate answers. Current system uses the method of self assigning tags to the questions. The users who are unaware of the domain cannot assign tags by themselves. This is where our project helps to assign most relevant tags to the questions to get answers from relevant persons who are experts in that domain.

Datasets Used:

StackOverflow Survey 2020, 2018

General Survey conducted each year to identify target audience who uses the website

https://www.kaggle.com/aitzaz/stack-overflow-developer-survey-2020

https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

Stack Overflow Data (BigQuery Dataset)

Contains Question & Answer discussion with tags, comments, votes , etc..

https://www.kaggle.com/stackoverflow/stackoverflow

Planned Approaches

1. Exploratory Data Analysis (EDA) on the survey dataset
2. Association Rule Mining (ARM) on Q&A Tags
3. Various Clustering Methods to group similar topics and similar questions

References:

1. Chen, C., & Xing, Z. (2016). Mining technology landscape from stack overflow. In A. Jedlitschka, & M. Jørgensen (Eds.), *10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement: ESEM 2016* [14] IEEE, Institute of Electrical and Electronics Engineers. https://doi.org/10.1145/2961111.2962588