Team Project Proposal

Part 1. General Information

Name: Utkarsha Shetye, Aditya Rajmane

Project Title: Technical Forum Analysis

Project Description: Technical discussion forums often help us to build our knowledge, taking help from people who are more experienced and who possess greater knowledge than ours. What if we use this information provided by the experienced people to give suggestions to the students who ask questions in the forum, by screening the website for similar posts or providing links related to the subject under consideration. What if we get links to websites which teach such courses or books available online which can help? Suggestion like this are valuable and can help us deeply understand the concepts.

Data Sources - StackOverFlow dataset(received), StackExchange dataset(received), Books Universal Links/ Youtube dataset(yet to receive the dataset)

Part 2. General Data Source Information

Data Sources	**Data Source Description**	**Data Size**
StackOverFlow	Source Website: https://archive.org/details/stackexchange This dataset consist of basic fields to identify user, their posts, comments, links, etc.	15GB
StackExchange	Source Website: https://archive.org/details/stackexchange This dataset contains various user posts in detail and can be useful to extract the content to search for the suggetions.	
Youtube/Books Universal Links dataset	Source Website: s3://datasets.elasticmapreduce/ngrams/books (https://aws.amazon.com/datasets/8172056142375670)	4.8GB

Part 3. Detailed Data Source Information

Data Sources	**Data Characteristics**	**Data Frequency**
- StackOverFlow - StackExchange - Youtube/Books Universal Links	- The data stored on the Books/Youtube server will be updated in soft realtime. - Data on the stack overflow will be updated in Real time - Data is non-static (Addition is done every minute)	- Dynamic/ real-time. - New data added every minute.

Part 4. Technologies

1. Basic: Map-reduce, MongoDB, R
2. Experiment/Risky: Kafka, Spark
3. Visualization: D3, Tableau

Part 5. References

amazon.com, stackexchange.com