Topic of study: Who is doing what in open source collaborative platform namely Wikipedia.

Motivation: Wikipedia is the largest multilingual, web-based, free-content encyclopedia that is available in the internet. This project is supported by Wikimedia foundation and it is completely open source that means it is editable by anyone. I want to analyze the edit history of different Wikipedia articles and find out interesting facts that can be extracted from it i.e., types of editors/contributors, recognizing editor characteristics from edits they make, how do articles evolve, community reactions to edits etc.. Wikipedia is built on online collaboration and information sharing among users. Since it is free and open to all, there are myriads of talking, editing and corresponding changes which are going on every day. The main motivation behind this project idea is in recent years it is observed that there are plenty of foul editing, vandalism that is taking place. To prevent this we need to identify what kind of editors are doing these and which types of articles are susceptible to these malicious activities. So in this research I have three research questions:

* What anonymous and registered people are doing (type of edits)?
* How people are reacting to those (edit over edits)?
* Potential gender bias in edit and revert war.

Method: The first task require topic modeling and editor role identification while the second task might require some kind of sentiment analysis and text mining. The third task will involve inferring user gender and study of gender effect on wiki text and corresponding edits. Some possible extension of this work might be: detecting correlation between quality of edits vs editor lifetime and bag-of-words approach to classify an edit to foretell if it is vulnerable to getting reverted eventually.

***Research challenges.*** First, the dataset of Wikipedia dump is of titanic scale. The English wiki dump produces terabytes of edit history which will be cumbersome to handle. Second, the results of analysis maybe misleading and it is difficult to determine a comprehensive metric for the study. Third, due to the nature and size of data the presence of bias in study is difficult to catch.

Dataset: The data is collected from the huge data dumps those are already available in wikimedia sites (https://dumps.wikimedia.org/enwiki/). The dumps for different language versions of Wikipedia are kept separate. It is important to note that the data are available in XML format (some previous version have SQL and HTML dumps too but are out of date) and needed to be transformed into more readable format. So I had to execute a Java tool that is already available called *mwdumper* that converts the XML to SQL. Some other tools such as mwdump.py, ImportDump.php, xml2sql (<https://meta.wikimedia.org/wiki/Data_dumps/Tools_for_importing>) are also available. For small scale analysis we can use the Wikipedia API (<https://www.mediawiki.org/wiki/API:Main_page>). Querying the revision history through revision API is also possible.