Stack Overflow Question Analysis using Topic Modeling

Nitish Ganesan (nkg8933@rit.edu)
Advisor – Dr. Carol J. Romanowski
Rochester Institute of Technology

Introduction
- Posting questions in the Internet to gather information is becoming popular and very common.
- Stack Overflow is a question-answer website specially for programmers.
- Stack Overflow users should manually select the tags to the question they wish to post.
- Topic modeling is a machine learning algorithm specially used for predicting topics.
- Using Topic modeling the tags can be automatically predicted which reduces human efforts and errors.

Goal
- To learn about the different topic modeling algorithms and choose three best suited algorithms.
- Predictions of tags by analyzing the Stack Overflow questions using the selected machine learning algorithms.
- Trend analysis of the predicted topics.

Sample Questions

Figure 1. Sample Questions

Figure 2. Topic Clusters

Figure 3 represents the accuracy and time taken for each topic model.
- The time taken by LDA was more but the model was consistent over the number of topics and passes.
- HDP chose 150 as the optimal number of topics but the accuracy was less compared to LDA.
- NMF took the least time but the accuracy is less compared to the other two algorithms.

Results

Figure 4. Topic Weightage

Figure 6 represents the popularity gained by the topics over the years.
- Php language is gaining the most popularity. C# was widely used but recently lost its attention.

Conclusion and Future work
- Question-answering websites are gaining more and more popularity. According to the trend analysis prediction, 10% questions will be posted in the stack overflow website for the upcoming year.
- In order to get accurate results, Machine learning algorithms should be chosen according to the nature of the dataset.
- Further analysis of the questions can be done to find the relationship between questions. Existing questions can be suggested to users if the relationship is found.

References