DS250

# Data Analytics and Visualization

Assignment 2

Animesh Raj : 11940120

Puja Bansal : 11940910

## Introduction

We will learn to analyze Wikipedia Graphs. We will focus on the articles in the mathematics category. Assume that there are students who are preparing for the JEE-Advanced exam.

https://jeeadv.ac.in/syllabus/combined-syllabus.pdf

We need to generate a sub-graph(s) of Wikipedia articles that are relevant to their study.

Furthermore, we will also try to give them an order in which these should be read (traversal algorithm). These traversals can be organized by subjects etc.

### Part A

Scrape and Label articles according to complexity in “Wiki Math Articles” sheet - It is done by our side(20 +20).

### Part B

Wikipedia Graph along with attributes including keywords, tags, NLP features etc.

- Had found hyperlinks from a wikipedia page to other wikipedia pages.
- Has implemented BFS to build the graph.

- To calculate node attributes, we have used NLTK keywords extraction - tokenizer, stopwords, stemming, removing keywords of length 1, Bag of words model etc.
- Word 2 vec embedding average has been used for all keywords for calculating NLP embedding. For that, TF weighted has also been implemented.
- We have also implemented a co-occurrence matrix which can be used to find word2vec embedding manually.

## Part C

Additional features based on Graph using concepts like centrality metrics, clustering coefficient.

Degree Centrality, Betweenness Centrality, Closeness Centrality, Clustering Coefficient, Page Rank has been implemented successfully.

## Part D

Development of Node classification Models.

We have developed 2 node classification models :

- Label propagation : For Label propagation, we have used sklearn library as shown in tutorial.
- ANN : We had feature vectors corresponding to every node, and for our root node we were having the labels. Based on that we have trained our model and found out the labels of other links.

## Part E

Graph Traversal :

Has implemented BFS to traverse the graph.

Article Ordering Algorithm:

Has printed it in an order such that one can read those links in that order. For this, I have difficulty labeling all nodes using ANN.

We for a node took nodes within a fixed distance from that node. Those nodes, we have sorted in their increasing difficulty level.

THANK YOU!!