

Dataset

<https://snap.stanford.edu/data/wiki-meta.html>

Plan for analysis

Part one

We would like to identify the similarities between pages, and create a way to score similarity between different documents.

We aim to do this by doing some sort of graph analysis and cluster documents into different domains.

Part two

We would like to find the breadth of domains that wikipedia editors have. For example, is an editor sticking to mathematical topics, or is that editor editing pages from mathematics to tv shows.

Teammates

Sachin Iyer and Triet Vo will be working together