# **Recommendation Engine for Wikipedia**

**A clear description of the topic:**

As a part of the final project, I would like to build a recommendation system for Wikipedia. Here, I will be collecting the data dump from [Wikipedia](https://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia). The main goals of the project are as follows:

* Collect the data and perform pre-processing on the Data set.
* Create a graph database to find the missing gaps in the data. Some of the articles are missing in few languages. Hence, we fill these gaps (Maybe by using translators)
* Build a recommendation engine on top of the graph database by prioritizing the user’s likes and interests

**Background research of related work:**

We formulate our problem as follows: given a (non-empty) set Li of Wikipedia articles that a user i has enjoyed or liked, we wish to generate a non-intersecting set Ri (Li ∩ Ri = ∅) of Wikipedia articles the user will also find interesting or enjoyable to read. Our challenge is in generating measures of similarity between articles that generalizes to allowing our recommendation engine to accurately suggest articles related to the user’s preferences, as well as out-of-the-box articles which (possibly unrelated to the user’s likes) he or she may still enjoy. An additional challenge is working with the massive dataset that comprises the 17+ million articles of Wikipedia.

The background for this project is described in this article:

<https://blog.wikimedia.org/2016/04/27/article-recommendation-system/>

**Data sources:**

<https://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia>

**What algorithms are being used and code sources:**

The algorithms used for this project are:

* Collaborative filtering
* Hierarchical clustering

**References:**

* [**https://en.wikipedia.org/wiki/Wikipedia:Database\_download#English-language\_Wikipedia**](https://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia)
* [**http://cs229.stanford.edu/proj2011/TopalovicRothfelsSaeta-ARecommendationEngineForWikipediaArticlesBasedOnConstrainedTrainingData.pdf**](http://cs229.stanford.edu/proj2011/TopalovicRothfelsSaeta-ARecommendationEngineForWikipediaArticlesBasedOnConstrainedTrainingData.pdf)
* [**https://blog.wikimedia.org/2016/04/27/article-recommendation-system/**](https://blog.wikimedia.org/2016/04/27/article-recommendation-system/)

**Who is in the group?**

* **Raksha Kaverappa-001224171**