# News Recommendation System January 2020

This project was aimed at developing a Content-based News Recommendation System for the client. The generated recommendations were to be used to ensure that only relevant content is used for sending out notifications to users. The raw textual content of news articles was to be represented numerically and compared for similarity. A variety of techniques commonly used in Natural Language Processing were used to perform this task. I was the sole resource deployed by Valiance on this and I was provided with a large corpus of news articles scraped from the internet for training purposes. I handled all project related activities independently without the need for close supervision.
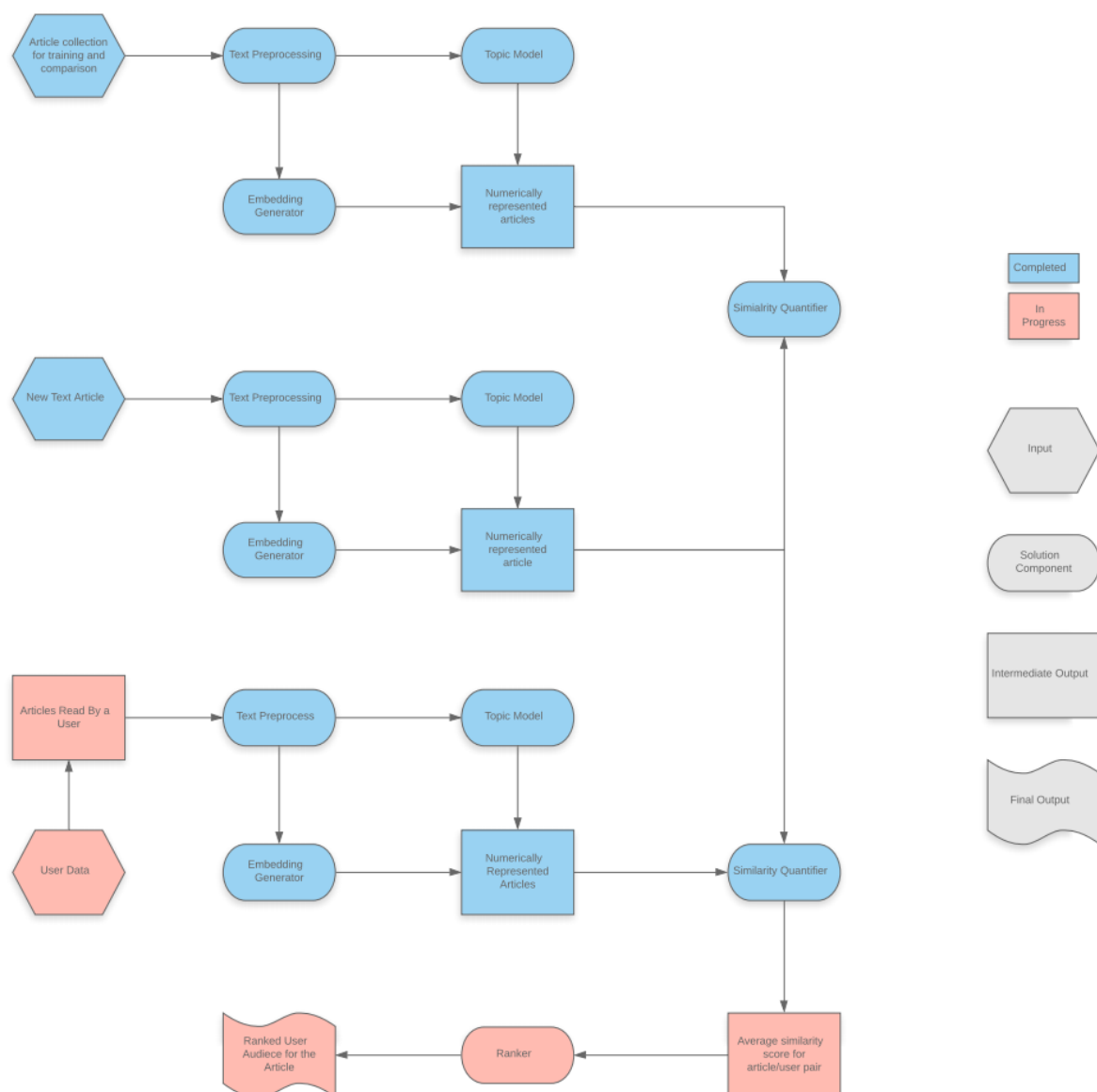
The various components of this project were as follows:

- A text pre-processing module that would perform the sanitization of the raw textual data from news articles.
    - Remove punctuations
    - Remove stop-words
    - Lemmatize the text
    - Remove other noisy elements
    - Standardize the text
- A LDA based topic model that would calculate the topic distribution in each article.
- A Doc2Vec based embedding generator that generate document level embeddings for each news article.
- A similarity module that would compute the similarity between two articles on the basis of topic distribution and document embeddings.
- A user module that would use all of the above to model a user's reading behaviour and recommend the most relevant content to him (this module is currently work in progress).

The entire pipeline for this project was created in python.