# Group 1 Status Report

Group 1

May 22, 2024

# Project Overview

- The goal of this project is to train a model that can predict the tag of a given wikipedia document.
- Bag of Words method is used to convert the text data into numerical data.
- Variant of fuzzy c-means clustering algorithm is used with the provided 'Categories' columns as class lables.
- After embedding space is generated, we use a pmf function using centroid of cluster and distance metric to generate a probability distribution.
- Lastly, the probability distribution is used to predict the tag of a given wikipedia document.

# Dataset and Pre-liminary Analysis

- Wikipedia Plaintext(23-07-01) Dataset
- Consist of 6,286,775 articles, titles, text and categories from July 1st 2023 dump
- Dataset consist of 4 columns: 'id', 'title', 'text', 'categories'
- All dataset is saved as .parquet file per starting alphabet/character

# Data Preprocessing

- Data is loaded from .parquet files
- Data is cleaned by removing special characters, numbers, and stopwords
- After cleaning, data is converted into lower-case where it's used to generate a unique words list.
- Documents are converted into numerical data using Bag of Words method

# Bag of Words

- The Bag of Words method converts text data into numerical data by representing each document as a vector of word counts.

$$\mathbf{d} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

- Here, $\mathbf{d}$ is a vector representing a document, and $f_i$ is the frequency of the $i$-th word from the vocabulary in the document.

# Document Vectorization and Clustering

- After all documents are vectorized, we apply dimension reduction(CUR decomposition) to reduce the dimension of the data.
- Afterwards, per unique category we may construct a probability distribution using the centroid of the cluster and distance metric.

# Document Cluster and probability distribution

- For each unique category, we construct a probability distribution using the centroid of the cluster and distance metric.
- The exact detail is not yet decided, but using normal distribution could yield pmf in form of:

$$p(d) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d)^2}{2\sigma^2}}$$

Where d is the distance from the centroid, and $\sigma$ is the variance of the cluster.

# Model Validation and Document Classification

- After the probability distribution is generated, we can use it to predict the tag of a given wikipedia document.
- The model will be validated using a test set of ratio 70:30, and the accuracy will be calculated depending on probability assigned to the correct tag.

# Current Progress

- Currently within process of extracting valid categories and documents from the dataset.