**Methodology**

**Methods Applied**

The principle of the recommended system is classification which means training the machine using labelled data to predict further articles. Therefore, we decided to choose Naïve Bayes and Stochastic Gradient Descent methods in the experiment which belongs to supervised learning to build the model and both methods are good at natural language processing.

The Naïve Bayes classifier is suitable for discrete data and determines the probability that an article belongs to the topic. In this case, the Bernoulli model can generate an indicator for each term of the vocabulary and use binary occurrence information, ignoring the number of occurrences, whereas the multinomial model keeps track of multiple occurrences. We think this might be an ideal model for the natural language processing because of its special handling methods.

On the other hand, Stochastic Gradient Descent is a simple and efficient approach to discriminative learning of linear classifier under convex loss function. This classifier is suitable for large-scale and spare machine learning problem and supports different loss functions and penalties for classification (L.Bottou, 2018). In this project, the SGD classifier combined multiple binary classifiers in a “one versus all” (OVA) scheme and computed the confidence score for each classifier and choose the class with the highest confidence. Since our dataset has a large scale and sparse features, the SGD classifier might be a good choice.

**Pre-processing and Feature Extraction**

The first step for data pre-processing is import the libraries and the data set. We imported the numpy and pandas packages for processing the data and also imported sklearn packages for feature extraction, machine learning and model metrics evaluation.

In order to clean and prepare the data sets for further analysis, we checked the training data. There are no missing values in article number, article words, and topic parts, and no noise data because all labels in topic part belongs to the 10 topics or “IRRELEVANT”.

Then we used 10-fold cross validation and early stopping method to utilize our data better and prevent the probability of underfitting or overfitting. The use of SGD classifier with grid search cross validation results in a stratified 10-fold cv and each fold is representative of the dataset. Comparing using cross validation and without it, the accuracy has been increased because there are more training and validation data for the model. The most important reason for using the CV is that we can find the best parameters for our model in this way, since we are able to try all of them by using a single set.

The final part of the pre-processing is the feature scaling/extraction. We tried to convert the text articles to a matrix in two different ways which are count vectorizer and tfid vectorizer. The count vectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also encode new documents using that vocabulary. However, the tfid vectorizer is to calculate word frequencies and using the frequency as the components of the resulting scores assigned to each word (Brownlee, 2017). It will tokenize documents, learn the vocabulary and inverse document frequency weightings and allow us to encode new documents.

Moreover, we think stop words might be an important thing for the result and we tried to set stop words as none which means use all words as the features for the model, and also tried to set stop words as English to remove the stopping words from the features.

**Model Selection**

After cleaning up the data and completing the feature extraction, we need to select the appropriate model to predict the topic of an article. We compared the Bernoulli Naïve Bayes and Stochastic Gradient Descent methods from three aspects, which are principle of classification process, efficiency and accuracy scores.

The Bernoulli model estimates as the fraction of documents of class that contain the term, and when classifying a test document, the Bernoulli model uses binary occurrence information, ignore the number of occurrences and treat all features as independent. As a result, the Bernoulli model typically makes many mistakes when classifying long documents (D.Manning, Raghavan, & Schutze, 2008). Also, this model will have a better performance in a large amount of training data than small amount.

However, the SGD classifier looks at the interaction between each word to a certain degree in the article and good at capturing those relations. And SGD can converge faster than Bernoulli Bayes because it performs updates more frequently. Therefore, we can get away with this because the data often contains redundant information and the gradient can be reasonably approximated without using the full dataset. There are some interactions between each word in this project, so the stochastic gradient descent model is better at capturing those.

By comparing the classification process and the efficiency of these two models, the stochastic gradient descent method has more advantages than Bernoulli Bayes method in this project, since each article contains the interrelated text and also requires a faster speed used to process the massive data. The accuracy scores verified our assumptions that the SGD classifier has a higher accuracy score 80% than the Bernoulli Bayes classifier 67%. Therefore, we used the stochastic gradient descent method as our model for this project.

# Reference

Brownlee, J. (2017, September). Retrieved from https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

D.Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

L.Bottou. (2018, 09). Retrieved from https://leon.bottou.org/projects/sgd