**UN Rapid Assessment Project:**

**Text Classification with Unlabeled Data**

**Fordham University Design Lab**

**Team Mozart**

Daizhen Zheng

Long Wan

Travis Park

Qi Zhang

**Problem Description**

The United Nations released 17 Sustainable Development Goals (SDGs). These serve to provide general direction and guidance to nations as they develop. They contain over 170 total targets which provide specific metrics for evaluating success. Countries submit National Planning Documents which outline their plans for achieving the goals and targets. Some documents can exceed 300 pages and it is very time consuming to evaluate them for compliance with all targets. The UN Development Group is seeking a tool that can aide analysts in creating the sector/target matrix which is used as a quick reference sheet when working with the National Planning Documents. Our task is to use text analysis and machine learning to read the documents and assign credit where targets are mentioned and highlight any missing information. It is important to note that we are not evaluating whether the text addresses the target satisfactorily, just if it is relevant at all.

**Methodology**

Challenges

This project posed two unique challenges. At its core, it is a classification problem. Our tool needed to be able to understand the drive behind each target, the meaning of each paragraph, and which target each paragraph is referencing. However, typical classification problems use supervised learning which requires a training data set that is already labelled with the answers for a model to mine for insights and evaluate success. This is not the case for this project. We were presented with raw text with no answers so it would be difficult to mine the data and difficult to determine if our model was accurate.

The next challenge was the number of targets. Even if there was a labelled dataset, there were significantly too many possible labels for a model to predict. With over 170 different possible predictions it would be very difficult for a model to uncover the nuanced differences between two targets that had very similar vocabulary but different goals.

Solutions

To overcome the problem of unlabeled data we transformed each paragraph and target into a term document matrix with Term Frequency – Inverse Document Frequency (TF-IDF) weights. This means a table with a row for each paragraph and a column for each word in the entire document. At the intersection if the paragraph contained that word more frequently than usual it would have a high score. TF-IDF has the benefit of elevating the importance of rare words and lowering the importance of very common words. With each paragraph and target represented as a vector we could determine similarity between each paragraph and target. The top 5% of scores were consistently good matches and the bottom 5% were not matches. The middle 90% however was ambiguous and could not be determined by similarity alone. By labelling the top 5% most similar as matches and the bottom 5% as non-matches we had created a sparsely labelled dataset.

There are several machine learning techniques available for classifying with sparsely labelled datasets such as semi-supervised learning or Positive Naïve Bayes Classifiers that we applied successfully.

Now that we had a labelled dataset, the next step is to train a model and start making predictions. Our models had a very low score initially due to the large number of potential classes. To mitigate this, we grouped each target by its goal which left only 17 goals to predict. This number is still high, but with some editing of the language of the goals they were made distinct from each other. A two-step classification process was implemented. First train a classifier to predict each paragraph’s goal, then limit the available targets to predict to only those under the goal for which it was labelled.

**Other Applications**

The lessons learned from this project can be applied to other text analytic projects that are unlabeled which could greatly increase the number of text datasets available on which to build models. By labelling a few records for which we are most confident using a rule based or term frequency approach, we can then use those labels to train models where previously it would have been impossible. Additionally, in some cases where the number of classes to predict is high, if those classes can be grouped, models can utilize a two-step approach to accomplish a high degree of accuracy.

**Future Improvements**

To increase the size of the training set for the models, a larger corpus of text could be gathered to increase the number of paragraphs that are strong matches using similarity and increase the number of records available to train the models.

A method of determining what sector the paragraph discussing (education, government, business, energy, etc) is required to complete the sector/target matrix.