Topic Extraction from Unstructured NGO Documents

Muhammed Daniyal Hamid, MS Data Science
Sidra Effendi, MS Data Science

BACKGROUND & PROBLEM STATEMENT

- NGO organisations produce different types of reports each year.
- The different reports give information about an organization's functioning and is largely left unused.
- We aim to see the kind of reports the different NGOs are producing. For example, types of reports in war zones vs reports for NGOs working with women and children.

OBJECTIVES

- Try machine learning and NLP techniques to successfully group unstructured documents collected from NGO websites.
- Extract insights from these document groupings to see which topics various NGOs publish documents regarding.

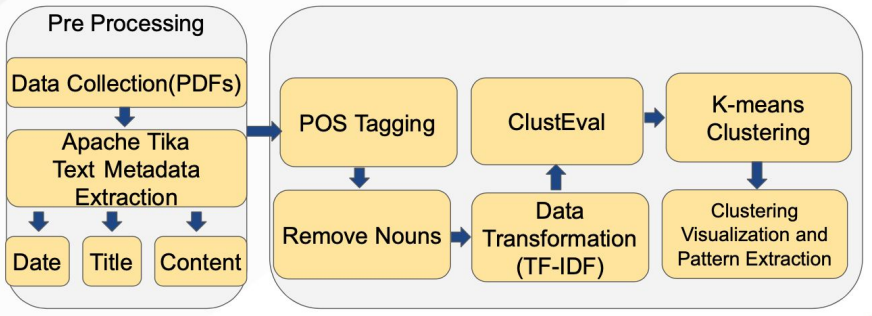
DATA

We have collected data from the Data4Good center. Dataset contained a total of 5000 pdfs from 34 different NGOs. The pdfs are different types of reports like, news report, situation report, program report, annual report, etc. Many of the pdfs fall into unidentified categories.

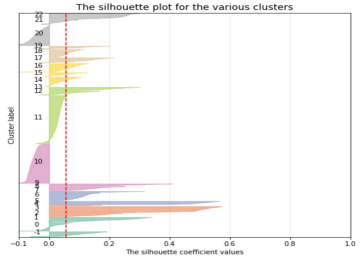
PRE-PROCESSING STEPS

- The first step was to extract metadata from the PDFs and use it to get rid of non-english documents by using language detector on the content.
- There were a lot of duplicate documents and they were removed using fuzzywuzzy on the document title. In the end we were left with around 3000 pdfs.
- Applied NLP on the content to remove Nouns from the text after POS tagging. The motivation for removing the nouns came from the fact that for a particular organisation, the reports might have common organisation name and common places mentioned and this could affect clustering of the report types.

METHOD



- Natural Language Processing
- Document vectorization:
  - TF-IDF
  - Doc2Vec
- Machine learning methods:
  - K-means clustering
  - Agglomerative clustering
- Using PCA to generate clustering visualization.



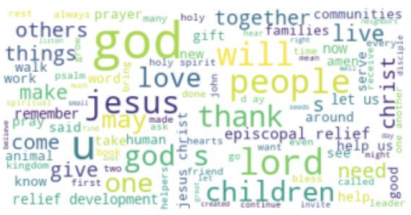
RESULT

While evaluating the word clouds the most coherent ones were with the k-means clustering algorithm using TF-IDF vectorization on our documents.
Both for k-means and agglomerative clustering algorithm the optimum number of clusters returned by clusteval is 24. Moreover, the silhouette[1] score for k-means is 0.056 and for agglomerative 0.051 using tf-idf. Doc2Vec performed worse in our evaluation.

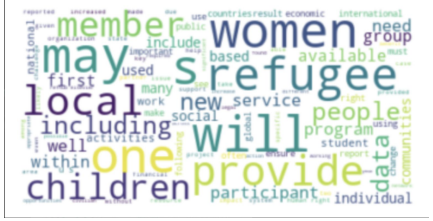
INSIGHT

We observed NGOs working for refugees were part of cluster with word cloud containing words- 'refugees', 'provide', etc. NGOs which were on a Christianity relief mission fell into cluster with words like 'Jesus', 'Lord', etc. in the cluster. Similar pattern was observed in other NGOs and clusters.

Cluster: 22



Cluster: 0



NEXT STEP

For the clusters which don't have a extractable label from document Title yet, along with the generated word cloud, topic modelling can be used to find the correct labels/topics for the report type.

REFERENCE

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.