COGNITIVE CATEGORISATION OF START-UPS TO IDENTIFY EMERGENT INDUSTRIES

Muhammad Faaez Riaz

Executive Summary

Objective

The objective was to build this NLP pipeline so the analysis can be reproduced every quarter which would then be used to facilitate financial analysis and investment decisions at the client. The result produced is an Industry Categorization for the companies in our dataset based on their description. The produced categorization would be used by the client to identify emergent industries within the startup world in Eastern and Central Europe.

Approach

The approach was to apply NLP techniques on company descriptions gathered from their website or crunch base in order to identify emergent industries within the start-up world and categorize all companies into these industries.

To build our NLP Pipeline we'll be using the Doc2Vec model to convert the company descriptions into vector embeddings and will then be running a clustering on the produced vectors to do the categorizations which will be based on the cluster value. The first stage will be to do the entire process on a subset of 5000 companies from our data and afterwards based on the results we'll do the analysis on the entire dataset.

Business Problem

Flashpoint is an international technology investment firm that manages equity and debt funds. Due to stiff competition in the venture capital world, there is a constant need to innovate in order to make better investment decisions. The company invests in software companies, their sweet spot is software-as-a-service solutions in the initial revenues stage (early stage start-ups). Therefore, industry categorization of start-ups and identification of emergent industries is an area of interest for the company because it can significantly help with financial analysis conducted by the company and improve investment decisions. However, it is not easy to categorize start-ups by their industry by using generic tagging.

Data

The client had collected data for over 600, 000 companies collected from crunch base and other sources with information including organization ID, name, website URL, description, finances, fund raise and industry category assigned by CrunchBase. Originally, the data was stored in the client's Postgre SQL Database, but to build our pipeline we exported the tables we needed to csv.

Infrastructure

We used Azure ML Studio with built-in support for Jupyter Labs for our project and used a STANDARD_D11_V2 compute instance to power our notebooks. Jupyter Notebooks and Python was our choice to build the NLP Pipeline in and the end result would be a CSV.

Data Preparation and Cleaning

After discussions with flashpoint, we narrowed the scope of the analysis to limit only to companies that had obtained funding; this would exclude non-funded startups, NGOs and Public Sector organizations. As part of Data cleaning we performed the following steps: Spell Check, Removal of URLs, Special Characters, Other Company names, frequently occurring words, Stop Words

Summary of Methodology

Used Python Pandas for all data related work and Matplotlib for all visualizations. The steps in our NLP Pipeline were:

1. Spell Check
2. Data Preparation and Cleaning.
3. Tokenization of Data (Includes NLP steps like stemming, lemmatization etc.): We used NLTK and Spacy to tokenize our textual documents and tagged them using the unique company ids.
4. Creating a Word Vectorization by running Doc2Vec model on our data: We used the Doc2Vec implementation of the Gensim library to get create our model and tuned it by inferring the records in our training set again.
5. Creating an Industry Categorization using clustering on the output/vectors produced by the Doc2Vec Model: We used k-means clustering from sklearn. The optimal number of clusters as validated for the 5000 records subset was 60 clusters.
6. Labelling Clusters: We used tf-idf and frequency analysis for analyzing the most frequently occurring terms in each cluster and used them to assign labels to each cluster.

Results and Validation

Our final output was a csv that contained the original company information alongside a cluster/category for each company. Overall, for the 104,000 companies we produced 800 unique categories and the client adopted a manual validation approach.

As part of the validation, the client picked random clusters within the results and manually verified that all companies within those cluster were actually similar/belonged to the same industry.

Conclusion

The results from the overall dataset showed a lot of promise and the client generally agreed with the broad industry categories that were produced. A further expansion of the project using more textual data and other approaches for categorization like using LDA (Topic Modelling) can also be done to improve results.

Overall value of the analysis can only be done when the client performs financial analysis (for eg. Capital inflows into various industries) using our results.