Online Based Enterprise Characteristics Experimental Statistics: United Kingdom

Aims

The aim of the project was to develop a proof of concept pipeline for the unsupervised clustering of companies based on web scraped text data.

This involved:

- handling data quality issues with web scraped data;
- writing software to run the pipeline;
- applying various cluster evaluation metrics.

The eventual goal of such a clustering is to go beyond the standard industrial classification (SIC) of companies and create a dataset which could be used to:

- identify emerging trends not captured by SIC;
- identify issues with SIC (such as misclassifications);
- impute SIC where such data is missing;
- provide a stratification variable.

Data Sources

Four datasets were used in the project: acquired web scraped data, company information from Companies House, the Standard Industrial Classification of Economic Activities (SIC) for the UK, and pretrained word vectors.

The web scraped data provided by a third-party consists of text extracted from the most likely candidate website from search engine queries for companies registered with Companies House. Each record has a score indicating a degree of confidence in the match between the business and the website. The text data is extracted from the HTML of the home page and an about page if it exists. This data was linked to company information from Companies House.

Companies House provide publicly available data on companies registered in the UK. This formed the target population for the web scraped data. Companies House also provide SIC data which was used as a comparison for clustering produced by this project. SIC is identical to the Statistical Classification of Economic Activities in the European Community (NACE) to the four-digit class level. The three components of the source data: web scraped text, company information and SIC are treated as a single dataset in the study. Pre-trained word vectors were used to improve cluster quality.

Pre-trained word vectors from the word2vec model trained on part of the Google News dataset [1] are used to embed the web scraped text data into a word vector space of 300 dimensions.

Methodology

The pipeline developed (see Figure 1) has four main stages: data cleaning, vectorisation, dimensionality reduction and clustering. This is followed by an evaluation of the clustering through cluster metrics and an SIC prediction task.

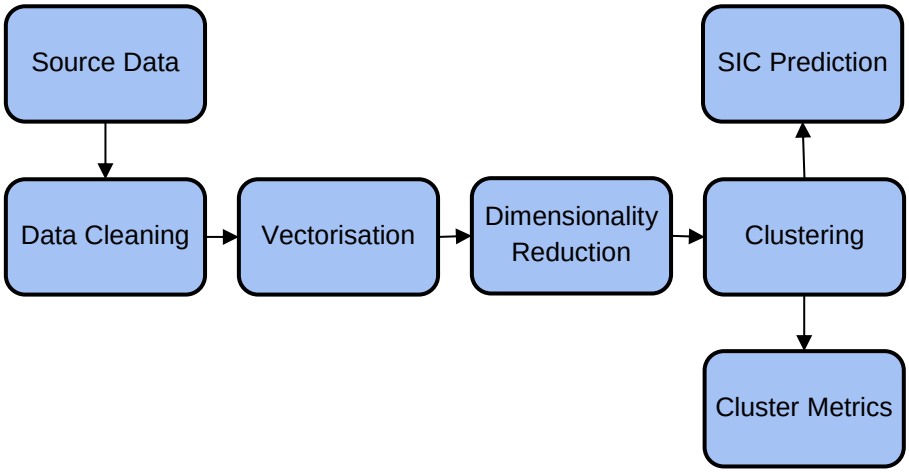

Figure 1: Simplified diagram of the processing pipeline.

Data collection

A third-party data provider used a list of companies registered with Companies House to find candidate websites from search engine queries. The candidate websites were then assigned a score based on heuristics (e.g. number of times the company name appears in the text of the homepage). The score (see Figure 2) was used to take a stratified sample of the data, this was manually checked and used to create a bootstrapped estimate of the ROC curve. This was then used to set a threshold for data quality.

The web scraping methodology itself was not the focus of this study, though it is important to note that the data acquired only represents a small (and biased) fraction of the target population.

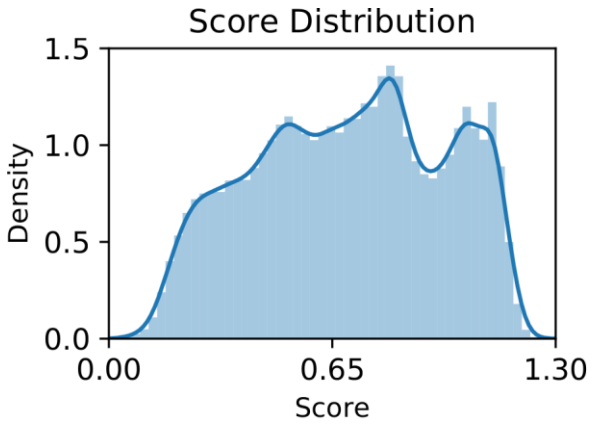

Figure 2: Score distribution of data provided from third-party. Score is based on heuristic features of websites likely to be good matches for the target. The scoring process was largely treated as a black box during the study.

Data cleaning

The main stages in data cleaning are:

- Deduplication (based on scraped domain and text extracted)
- Removal of bad data (e.g. error pages)
- Removal of punctuation/stopwords
- Lemmatization
- Filter for English language words only
- Removal of websites with only a few words extracted

During the development of the cleaning pipeline topic analysis with latent Dirichlet allocation [2] was used for data exploration. This was an effective means for discovering issues, for example clear topics around vocabulary specific to HTTP, JavaScript and cookies led to efficient cleaning rules. From a practical point of view the visualisation tool: pyLDAviz (https://github.com/bmabey/pyLDAvis) was useful for exploring topics.

Vectorisation

To vectorise the text data a pretrained word2vec model was used to first embed each word into a vector space. The vector for a given website is then formed by taking the mean of the vectors for the words from that website. This is not a state-of-the-art approach, but it is simple to perform and is used as a benchmark for document classification (for example see [3]).

The word2vec model is trained on the "Google News" dataset and contains 300-dimensional vectors for around 3 million words and phrases [1].

Dimensionality Reduction

Clustering data in a 300-dimensional space is challenging (see [4] & [5] for more on "the curse of dimensionality"). To enable clustering of the data we first decrease the number of dimensions. This is done with Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). In essence, UMAP finds a low dimensional projection of the data which preserves the topological structure (see [6] for details). In our case we use UMAP to step down the dimensionality from 300 to 10. Enough of the local structure is preserved to support density-based clustering algorithms.

Unsupervised clustering

The clustering itself is done with a density-based approach: HDBSCAN [7]. This has few parameters to tune and as a density-based method is flexible regarding cluster shape. In our case we use a hard clustering (i.e. no fuzzy membership), a data point is either in a cluster or labelled as an outlier. This simplifies the analysis of the clustering at the cost of leaving many values labelled as outliers.

Cluster validation

Cluster evaluation can be thought of as internal: how well are the clusters formed within the space; and external: how do they compare to some "ground truth" clustering. As the aim of the study is to improve on SIC it makes sense to treat SIC as such a ground truth. Evaluation methods can be further broken down into global and local metrics (i.e. do they refer to the clustering as a whole or give a metrics for individual clusters). The following presents two closely related methods of evaluation external to the clustering based on SIC as a "ground truth". The first is a method of local cluster

evaluation using an impurity metric and the second is a global metric based on using the clustering to predict SIC.

A perfectly valid clustering of companies could still be completely different to SIC. In our case it would be reassuring to see some level of correspondence with SIC, as evidence that similar concepts are being captured. We want a clustering which is similar enough to SIC to allow for comparison and yet different enough to yield new information about the economic activity of the companies in question.

Impurity, an external metric, is one example of a cluster evaluation metric used in this study. The Gini impurity of a cluster, C, containing k classes of a given level of the SIC hierarchy is defined as:

$$I_C = 1 - \sum_{i=1}^{k} {p_i}^2$$

Where p_i is the empirical probability of SIC class i in cluster C (i.e. the proportion of the cluster drawn from that class).

Impurity gives a cluster level metric. If all clusters are highly impure it is reasonable to assume that companies are being linked in a way quite different to SIC. Clusters with low impurity are more likely to be being linked in ways which would be recognisably similar to SIC.

An alternative, although closely related, approach is to train a tree-based model to predict SIC from the cluster assignment. A performance measure such as the macro averaged F-Score can then be used to compare different clusterings and to assess how closely a clustering follows the structure of SIC.

SIC prediction can also be carried out before clustering, based purely on the vectors in the embedding (either in the 300-dimensional word2vec space or in the reduced 10-dimensional space). This can give a sense of how much of the SIC structure that is possible to extract from the embeddings with a complex model is captured by the clustering. In this study neural networks were used for prediction of SIC from the embeddings.

Results

The project aimed to develop a method to perform unsupervised clustering on websites associated with companies. The preliminary results presented here cover the embedding of websites in a vector space and the clustering itself with two of the methods explored for cluster evaluation in terms of SIC: cluster impurity and SIC prediction.

Embedding

Each website in the dataset is mapped to a vector in a 10-dimensional space. The dimensionality reduction step helps avoid the break down of distance metrics in high dimensional spaces and maintains enough local structure to support meaningful density-based clustering. See Figure 3 for a visualisation showing clear spatial structure of an embedding of 130,000 websites.

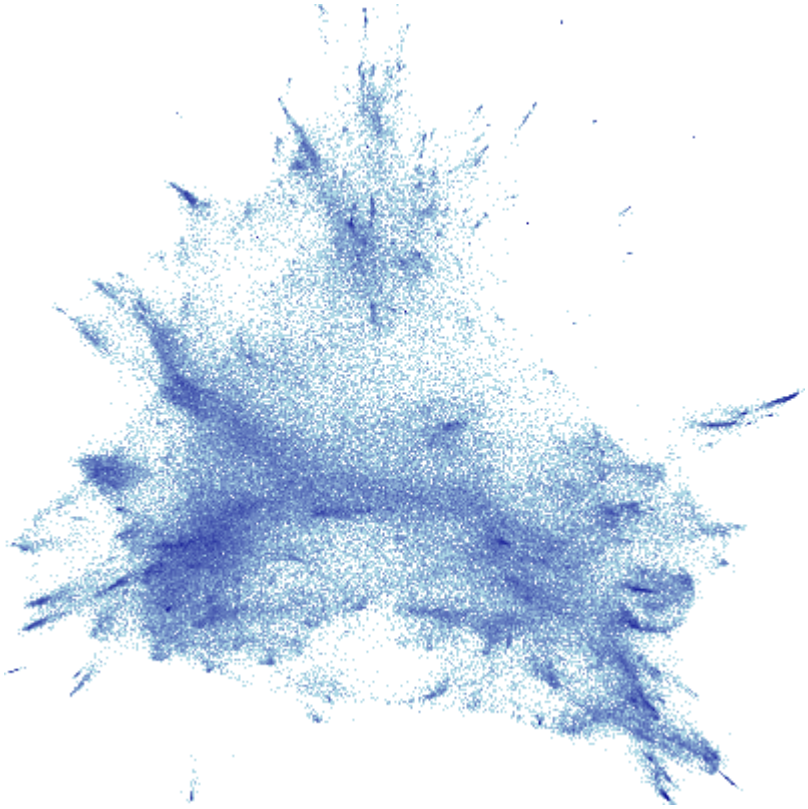

Figure 3: Vector embedding of 130,000 websites based on text data. This visualisation is of 2 dimensions of the 10-dimensional embedding (i.e. after UMAP is applied to the output of the word2vec vectorisation). Clear structure is visible, although as a note of caution it is possible for some of the apparent structure to be created as an artefact of the dimensionality reduction procedure itself.

Clustering

The number of companies that don't get assigned to a cluster under this method seems to be high. For example, in one clustering of 130,000 websites only 44% were assigned to a cluster. A clustering with this level of coverage would be limited in its usefulness. There are methods to deal with this, for example the soft clustering extension to HDBSCAN. However, we were reluctant to take an approach that would likely lead to more tenuous cluster associations before first establishing how well formed the hard clusters were. As such improving the clustering coverage is left for future work.

Cluster Impurity

See Figure 4 for a plot of cluster impurity. Here the upper and lower bounds are formed by drawing clusters from a multinomial distribution with the empirical SIC class distribution from the same dataset. It is clear that some of the clusters have structural similarities with SIC.

A cluster with a single SIC category would be perfectly pure (and hence would have impurity 0). Clusters within the bounds are consistent, in terms of impurity, with those that would be formed by chance. A systematic study of clusters with low impurity relative to clusters of a similar size has not yet been carried out.

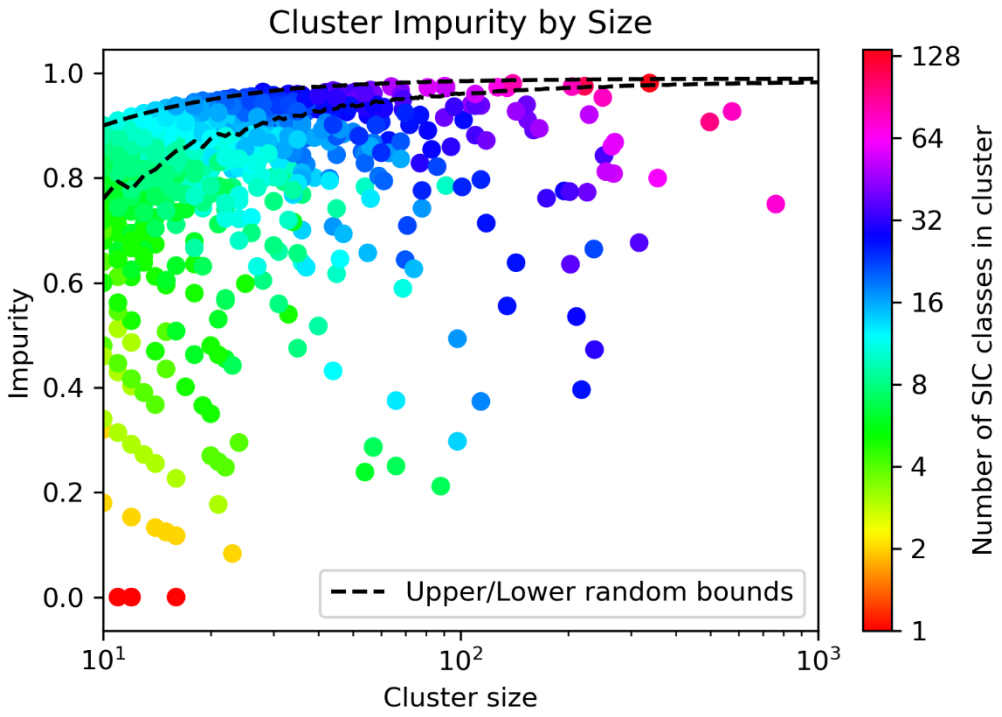

Figure 4: Plot of Gini impurity of around 2,000 clusters by cluster size where points are coloured by the number of the 4th level of the SIC hierarchy classes in a cluster. Bounds formed by bootstrap sampling random allocations of SIC classes from a multinomial distribution based on the empirical SIC distribution (bounds covering 99.9% of samples drawn).

SIC Prediction

Using cluster assignment to train a random forest and then reporting on model performance gives an alternative perspective on cluster evaluation. For example, using the same clustering as in Figure 4 to predict the top-level of the SIC hierarchy (21 SIC categories rather than the around 530 at the fourth level) gives a macro averaged F-Score of 0.31 ± 0.006 (1 standard deviation, 10 fold cross validated with 10 repeats). Similar predictions can be carried out at different levels of the hierarchy, although the metrics drop off as the number of categories to predict increases. On its own this number is difficult to interpret, especially considering the class imbalance in the data, however it does provide a way of comparing multiple attempts at clustering. It can also serve as a useful test of whether the clustering is losing information on SIC by comparing results to models trained directly on the embedding rather than clustering.

Conclusion

The methodology developed should be considered a work in progress. On manual inspection a significant proportion of the clusters do appear meaningful. However, the data used for clustering in the study was only a small fraction of the target population, with differences in the SIC class distribution. Regardless of clustering performance this limits what we would be able to say about the general applicability of this approach.

Regarding clustering performance, the clustering algorithm marks many companies as outliers rather than associating them with a cluster. It is also clearly very different to SIC. Even if the clustering produced is valid in terms of the text found on websites these two issues may need to be addressed before this could be used to augment SIC.

A full analysis of the stability of the clustering and derived metrics has not been completed. There has also not been enough work done to demonstrate the value the clustering adds to SIC to make a recommendation regarding adoption of the methodology.

References

[1] "word2vec," [Online]. Available: https://code.google.com/archive/p/word2vec/. [Accessed 31 10 2019].

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning research,* pp. 993-1022, 2003.

[3] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *International conference on machine learning*, 2014.

[4] R. E. Bellman, Adaptive control processes: a guided tour, Princeton University Press, 1961.

[5] M. Steinbach, L. Ertöz and V. Kumar, "The challenges of clustering high dimensional data," in *New directions in statistical physics*, Springer, 2004, pp. 273-309.

[6] L. McInnes, J. Healy and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *preprint (arXiv).*

[7] R. J. G. B. Campello, D. Moulavi and J. Sander, "Density-based clustering based on hierarchical density estimates.," in *Pacific-Asia conference on knowledge discovery and data mining*, 2013.