Consumer Data Research Centre

An ESRC Data Investment

CACI

Topic extraction and document classification on textual survey data with Unsupervised modelling techniques

Eirini Milaiou[1], Guy Lansley[1] and Chase Farmer[2]
[1]University College London, [2]CACI UK

Project Background

In most customer surveys there is plenty of information in the form of comments in raw unstructured texts. Thus, the necessity of taking this information into account for understanding customer behaviours leads to the need for analysing this data within frameworks from text mining to extracting underlying patterns. Topic modelling is a highly popular area of text mining for documents to automatically understand their content and extract high quality information, without human annotation. In this project the main aim is to identify and to capture the underlying topics and to cluster the user's comments into those topics, using unsupervised topic modelling techniques. The research was based on data from a large survey of shopping centres across the UK.

Data and Methods

The main challenge was to handle the structure of the documents, which are short and without proper syntax text messages. To achieve the set goals, topic extraction models were implemented to reveal the latent themes in the collection and to cluster the documents accordingly. Biterm, LDA with variational inference and Gibbs sampling and Topic modelling with distributed representation of words were the algorithms which were implemented and tested for this problem. Biterm topic modelling and Gaussian Mixture Models with distributed representation of words were examined due to their good performance on short documents. LDA, even though it is not addressed by the literature as the most appropriate algorithm for modelling short documents, it was favoured because the documents from the data typically only represent singular topics. Most documents express their topics clearly and in few words. In that way, it is assumed that LDA can capture the underlying topics by operating on short documents. The evaluation of topic models is not a straightforward task due to the lack of labelled/test data. The evaluation was approached as a three level procedure including qualitative and quantitative methods, based on the themes interpretation, topic coherence and the successful clustering of the documents. For the best results, cosine similarity of the topics was also conducted, it identified that no pair of topics exceeded similarity of 0.4 illustrating that the themes are satisfyingly discrete.

Key Findings

LDA with Gibbs sampling on single documents outperforms the rest of the models. The most satisfying model produced distinctive and informative topics, close to the industry's expectations and it also performed well on classifying comments into the appropriate topics. Following an exploration of perplexity scores, it was concluded that this particular dataset can be described by 13 distinct topics. Additionally, a combination of the extracted features of the documents with the numerical variables in the dataset highlights some patterns regarding the attributes of each shopping centre, patterns which would be inefficient to be extracted by human inspection. For instance, analysis of certain shopping centres using the composition of survey topics and specific ratings was carried out, which indicated potential issues the centres exhibit. Moreover, clustering of these centres was also conducted using the distribution of comments per topics in order to identify broad trends across the data.

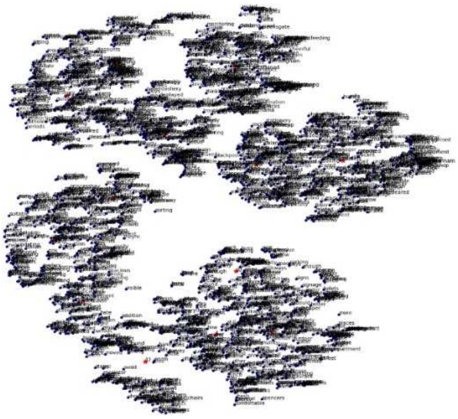

Figure 1. Grouped words using the means from the components of a Gaussian Mixture Model

Value of the Research

The novelty of this work is that it assesses various topic extraction techniques on short comments, a useful tool for survey analysis. Additionally, for the evaluation of the results, a three level assessment is suggested in order to encounter the problem of lacking labelled data. However, the experiments on a real-world dataset from industry were successful and are useful for achieving quick insight on large volumes of textual data.