A very-short-text clustering method based on distributed representation to identifying research capabilities of a Higher Education Institution

Keywords- Short Text, Clustering, Expansion representation, latent factors, word embeddings, Semantic similarity.

Purpose. Text documents are an important source of data for tech mining techniques. Usually text databases include document sufficiently long to apply conventional text mining techniques. However in some tech mining tasks, such as capabilities identification process, we have database with very short texts, which represent a challenge for conventional text mining techniques. The problem has to do with the small number of terms that fail to provide enough statistical information to find any kind of relationships among the documents in the collection. The main purpose of this work is to show how to generate thematic clusters using only the titles of the research projects in one Higher Education Institution.

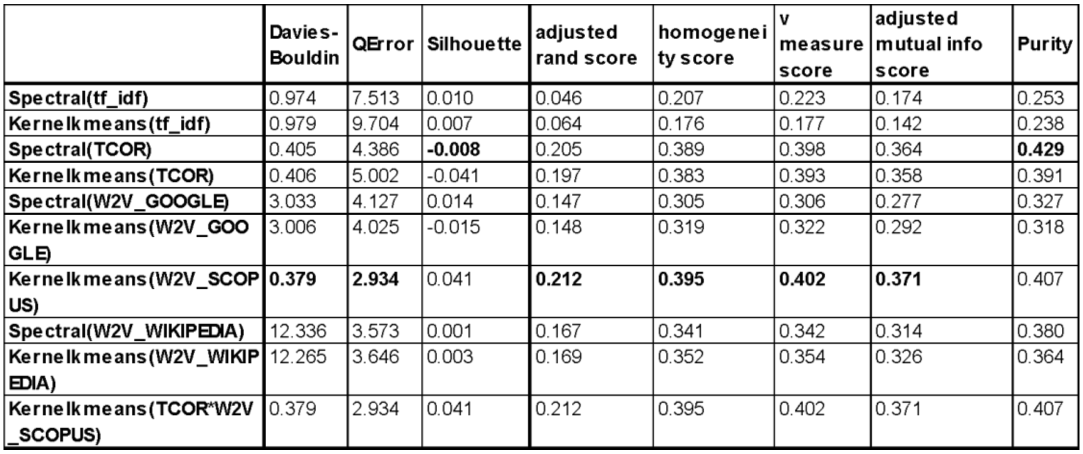
Working with short-text collections has become an important area of research in information retrieval and data mining, due to the proliferation of data sources with reduced textual information, e.g. blogs, reviews, tweets and other social-network and message-sharing platforms. Many researchers have focused efforts on techniques and applications of short-text clustering. In most of these works short texts correspond to documents with a handful of sentences. However, there are not works that concentrate on classification of very-short texts, i.e., that do not span more than one sentence.

Methodology. This is an exploratory study with an experimental design. The first step of the study is to find the best semantic representation for a specific short-text collection. To this end we tried two approaches, one of them uses term co-occurrence representation (TCOR). It is based on co-occurrence statistics that represent the semantic of a particular word by the terms that co-occur with across all text (Cabrera, Escalante, & Montes-Y-Gómez, 2013). Unfortunately the tiny length of the documents make less probable that different texts have terms in common making the co-occurrence matrix sparse. To alleviate this, we collected a set of related documents (scientific papers) using the short texts as queries, and then extracting word co-ocurrence statistics from the extended set of documents. In the second approach, we used a pre-trained word embedding to extend the semantic representation of the words using external knowledge in an efficient implementation of the continuous Bag-of-Words (Wang et al., 2016). We used a word2vec model learned from a Google News dataset, which contains 300-dimensional vectors for 3 million words and phrases. We also tested other word embeddings trained with other document databases: entries from Wikipedia and there were 1.151.090 extracted words, and 1.045.144 different abstracts from the Scopus database.

Using as input these representations we evaluated three different clustering algorithms (Spectral Clustering, Kernel k-means and Non-Negative Matrix Factorization). The results were evaluated using three intrinsic measures (Davies-Bouldin, QError, Slihouette) and five extrinsic measures (Homogeneity Score, V-measure, Adjusted MI, Purity).

The different algorithms and configurations were evaluated in a dataset of article titles obtained from the Scopus database using 20 different queries. For each query, 100 most relevant documents that contain all of the terms of the query in the keywords field were kept. The 20 different queries were used as ground truth for the extrinsic measures.

Findings. Table 1 presents the results obtained. They show that the performance of word2vec trained with Scopus data Base (W2C_SCOPUS) method, which uses term selection, is the best compared to the other two words embeddings (word2vec with Wikipedia and Google News). The TCOR representation has a competitive performance compared to W2C_SCOPUS, in fact it got the best performance for two of the measures. We can conclude that the performance of methods using external knowledge related to the document collection was better than the performance of traditional methods. Similar conclusions were found generating thematic clusters using only the titles of the research projects of National University of Colombia.



	Davies-Bouldin	**QError**	**Silhouette**	**adjusted rand score**	**homogeneity score**	**v measure score**	**adjusted mutual info score**	**Purity**
Spectral(tf_idf)	0.974	7.513	0.010	0.046	0.207	0.223	0.174	0.253
Kernelkmeans(tf_idf)	0.979	9.704	0.007	0.064	0.176	0.177	0.142	0.238
Spectral(TCOR)	0.405	4.386	**-0.008**	0.205	0.389	0.398	0.364	**0.429**
Kernelkmeans(TCOR)	0.406	5.002	-0.041	0.197	0.383	0.393	0.358	0.391
Spectral(W2V_GOOGLE)	3.033	4.127	0.014	0.147	0.305	0.306	0.277	0.327
Kernelkmeans(W2V_GOOGLE)	3.006	4.025	-0.015	0.148	0.319	0.322	0.292	0.318
Kernelkmeans(W2V_SCOPUS)	**0.379**	**2.934**	0.041	**0.212**	**0.395**	**0.402**	**0.371**	0.407
Spectral(W2V_WIKIPEDIA)	12.336	3.573	0.001	0.167	0.341	0.342	0.314	0.380
Kernelkmeans(W2V_WIKIPEDIA)	12.265	3.646	0.003	0.169	0.352	0.354	0.326	0.364
Kernelkmeans(TCOR*W2V_SCOPUS)	0.379	2.934	0.041	0.212	0.395	0.402	0.371	0.407

Table 1. Clustering performances obtained by using TF-IDF original based on the Scopus titles dataset, best TCOR, and best result using word2vec based on for represent titles in a vector space model with the different external sources (Wikipedia, Google News and Scopus data base).

References

Cabrera, J. M., Escalante, H. J., & Montes-Y-Gómez, M. (2013). Distributional term representations for short-text categorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7817 LNCS, pp. 335–346). doi:10.1007/978-3-642-37256-8_28

Cagnina, L., Errecalde, M., Ingaramo, D., & Rosso, P. (2014). An efficient Particle Swarm Optimization approach to cluster short texts. *Information Sciences*, *265*, 36–49. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025513008542

Pinto, D., Benedí, J., Rosso, P., & Benedi, J.-M. (2007). Clustering narrow-domain short texts by using the Kullback-Leibler distance. *… Linguistics and Intelligent Text Processing*, *4394*, 611–622. doi:10.1007/978-3-540-70939-8

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806–814. doi:10.1016/j.neucom.2015.09.096