

# Applying text mining to understand the academic internships market

Rui Almeida and Nuno Escudeiro

Instituto Superior de Engenharia do Porto  
{1150727,nfe}@isep.ipp.pt

**Abstract.** The academic internships market evolves fast. Understanding its dynamics is crucial to keep the market balanced, as internship offers submitted by companies should be aligned as much as possible with the searches made by students. Applying text mining approaches to help analyse internship descriptions might give an updated view of the academic internships market during a certain period of time. Machine learning techniques, such as document clustering, are helpful to extract patterns from textual data. A set of clustering algorithms and dimensionality reduction techniques were compared on a collection of 2163 internship descriptions. Results indicate that reducing the number of dimensions can improve cluster assignments, despite the loss of information.

**Keywords:** Text mining, Document clustering, Dimensionality reduction, Data visualization

## 1 Introduction

Education plays an important role in society, especially as demand for specialized people increases in the job market. In 2018 there were 17.5 million higher education students in Europe [1]. Internships are often the first professional experience that students have.

To promote innovation in the academic internship market, the Praxis <sup>1</sup> web platform helps to connect students with companies, higher education institutions and research labs. Organizations are able to share national or international internships while students can search them, using keywords and filters, and apply to them.

Unsupervised clustering algorithms are useful for exploratory data analysis. It is hypothesized that text mining techniques can be used to help understand the current state and evolution of the academic internships market. Combinations of different clustering algorithms and preprocessing techniques were evaluated by visually inspecting the clusters generated and through internal cluster validation measures.

---

<sup>1</sup> <https://praxisnetwork.eu>

## 2 State-of-the-art

Automatic processing of text by means of computer applications commonly requires a collection of text documents (corpus) to be transformed into a numeric representation, as most data mining algorithms require structured numeric data [2]. A simple approach is to create document-term matrix, where each row corresponds to a document, each column corresponds to a term, and each value contains the number of appearances of that term in a document. As certain words appear in many documents, alternative weight schemes can be used, such as the term frequency-inverse document frequency (TF-IDF), which reduces the importance of words that appear in many documents [3].

The resulting matrices are sparse and have a high number of dimensions. Dimensionality reduction techniques, such as stop word removal, stemming and latent semantic analysis (LSA), are frequently used preprocessing steps to improve the performance of other algorithms [4].

Alternative methods to obtain vector representations for words, such as word2vec [5] and GloVe [6], have been extensively researched in recent years, showing good results in document clustering and classification tasks [7]. These algorithms are trained on huge corpus and are able to capture semantic similarity by representing similar words close to each other in a dense vector space. However, when dealing with smaller corpus, these new methods can perform worse than LSA at capturing relevant word associations [8].

To cluster documents in distinct groups, based on their similarity, hierarchical clustering [9] and k-means variants [10] are widely used hard clustering algorithms in various domains. The cosine similarity is generally chosen for text analysis, as metrics like the Euclidean distance are not appropriate for high dimensional and sparse data [11]. For soft clustering, topic modeling algorithms like Latent Dirichlet Allocation and non-negative matrix factorization are able to discover latent topics that occur in a corpus [12].

Hierarchical clustering has the advantages of producing a dendrogram, which is a tree-based representation of the clusters, and the number of clusters does not have to be pre-specified. A limitation of hierarchical algorithms is the time complexity, at least quadratic, since the similarity between all documents has to be calculated.

Spherical k-means, which uses the cosine similarity instead of the Euclidean distance, is often used for document clustering [13]. It has the advantages of scaling linearly with the number of documents, however it is not deterministic, as it is sensitive to the initial random seed selection, and the number of clusters must be pre-specified.

Determining the ideal number of clusters is a difficult task, as their quality must be evaluated and can be subjective. Possible strategies include the visual inspection of the keywords and documents in the clusters or calculating metrics like the mean silhouette. The silhouette of a document measures how similar it is to other documents in its own cluster, when compared to other clusters, ranging from -1 to +1 [4]. This is an area of active research, as statistical measures do not always correlate well with the interpretability of topics [14] [15].

An effective way to compare cluster assignments between different algorithms, is to do a two-dimensional projection of the data and apply colors corresponding to the clusters for all data points. A state of the art nonlinear dimensionality reduction algorithm for embedding high-dimensional data in two or three dimensions, for data visualization purposes, is t-distributed Stochastic Neighbor Embedding (t-SNE) [16].

### 3 Methodology

To group internships based on their descriptions, different clustering algorithms and preprocessing techniques were tested to achieve better results. The first step was to construct the corpus by extracting and cleaning the data from a database. Afterwards, a list of stop words was created to exclude words that appear too frequently in English and in this domain, such as "work" or "skills".

The corpus contains 2163 internships submitted in Praxis between October 2013 and November 2019. Most of them are written in English. The following preprocessing steps were done to transform the description of internships into a numeric representation:

1. Lowercase conversion, replace diacritics, remove numbers and HTML tags.
2. Perform tokenization to split each word.
3. Remove stop words.
4. Compute TF-IDF feature vectors.
5. Dimensionality reduction with LSA (25, 50, 100, 150 and 300 dimensions)
6. Apply clustering algorithms (hierarchical clustering, spherical k-means).

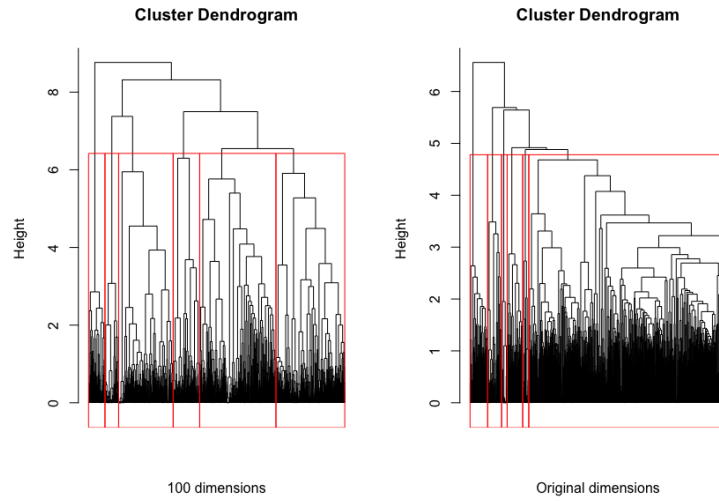
To compare the resulting clusters, the most important words for each cluster were inspected. They were obtained by calculating the mean TF-IDF score of each word in each cluster and sorting them. The mean silhouette was selected as the internal validation measure to compare the cohesion of the different clustering algorithms.

### 4 Results and discussion

The removal of stop words significantly improved the clusters generated. Stemming was excluded from the preprocessing steps as it had minimal impact. These statements are based on the visual inspection of the clusters.

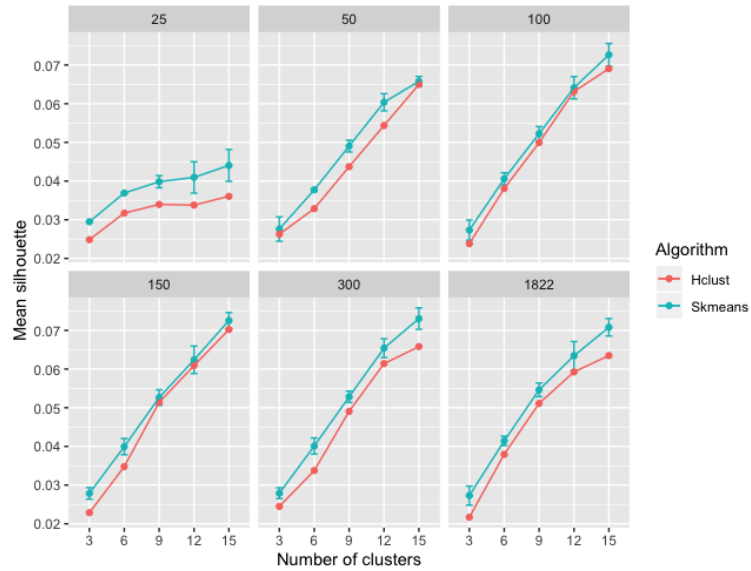
Hierarchical clustering had significantly better results after reducing the number of dimensions with LSA. The dendrogram in Figure 1, shows that using the original TF-IDF matrix causes one cluster to be dominant. With 100 dimensions, the clusters are much more balanced. Spherical k-means did not have the same issue with high dimensionality.

To evaluate how many dimensions are necessary, the mean silhouette was calculated with 30 random initializations for the different combinations of algorithms, number of clusters and LSA dimensions. The results are presented in



**Fig. 1.** Dendrograms with a cut of 6 clusters for different dimensions

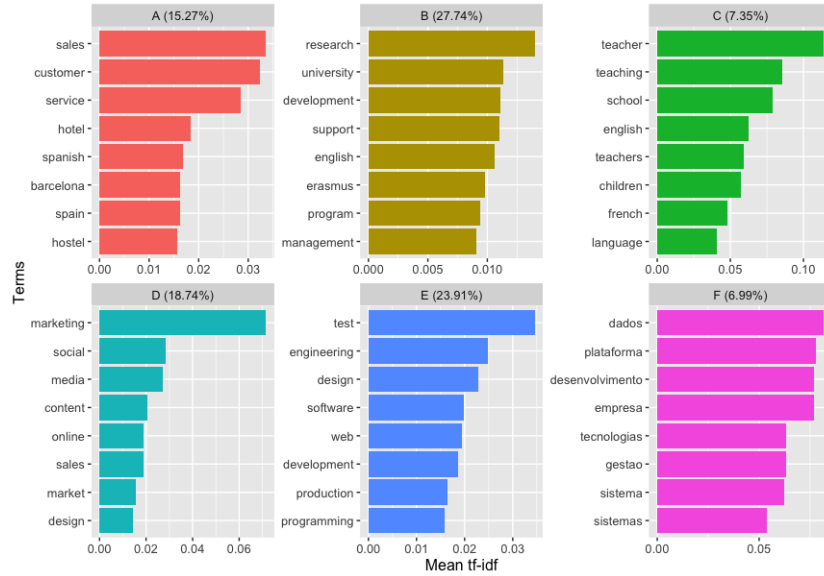
Figure 2. Spherical k-means has slightly higher silhouette and with 100 or more dimensions, the silhouette does not significantly decrease. Visual inspection of the clusters also suggests that quality drops with less than 100 dimensions.



**Fig. 2.** Mean silhouette comparison

On the current corpus, the number of clusters should be lower than 10. Higher values frequently result in clusters with internships created by a single organization, as a few of them submit similar internships. Figure 3 presents the most relevant terms for each cluster when using spherical k-means with 6 clusters and the dimensions reduced to 100 with LSA. The following clusters can be identified:

- **Cluster A** - Sales, customer service and hospitality. Many of these are submitted in Spain.
- **Cluster B** - University research and management.
- **Cluster C** - Education, mostly related with teaching other languages.
- **Cluster D** - Marketing.
- **Cluster E** - Engineering, software development and design.
- **Cluster F** - Internships written in Portuguese.



**Fig. 3.** Most important terms per cluster

Looking at the evolution of these clusters over the years in Figure 4, we can see a significant decrease in the education internships when compared to other areas. Engineering internships also appear to be decreasing, but most internships written in Portuguese are related with software engineering. There is an increase on the internships related with sales, customer service, hospitality and marketing. To inspect each cluster in more detail, term frequencies and term correlations can be used to plot word clouds and co-occurrence networks, respectively. Figure 5 demonstrates a word co-occurrence network with the words that appear most often together in internships from the engineering cluster.

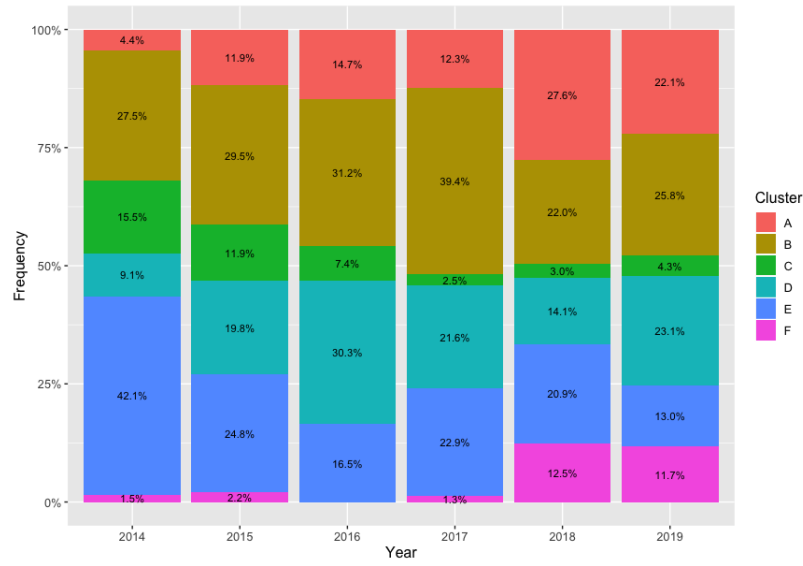


Fig. 4. Clusters evolution

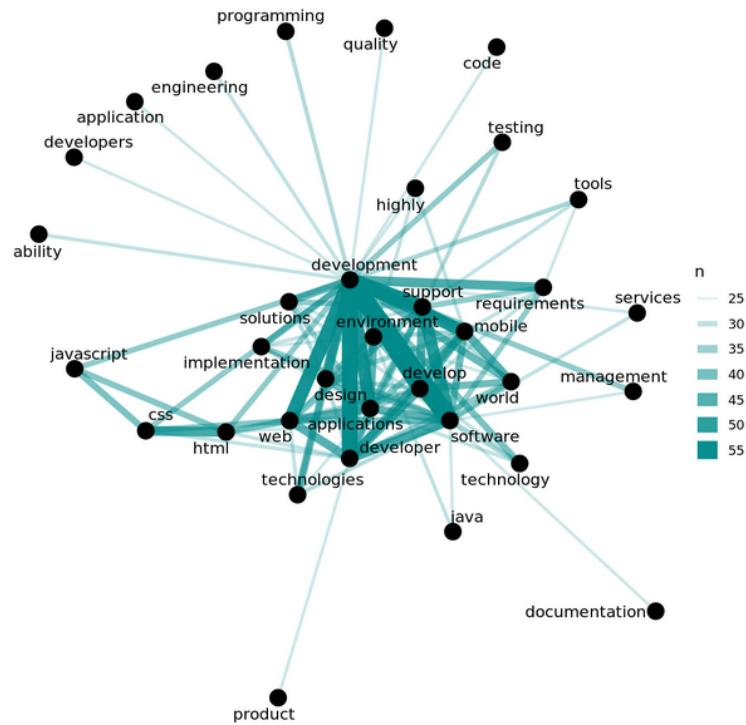
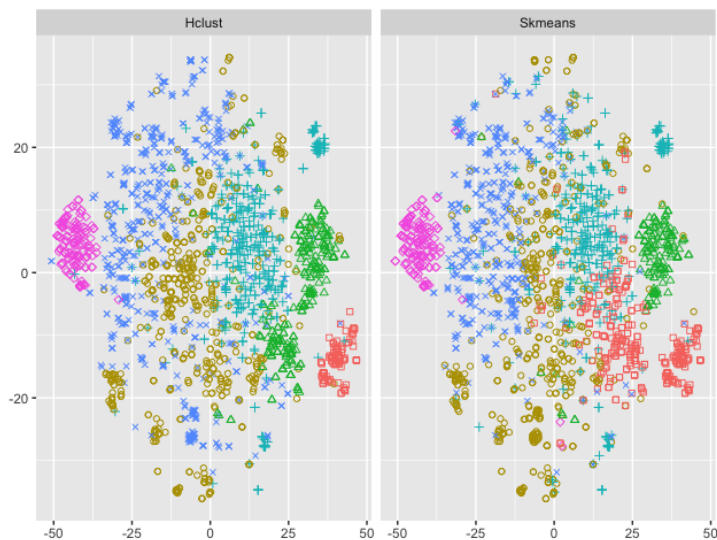


Fig. 5. Co-occurrence network of the engineering cluster

A comparison of the cluster assignments made by the algorithms, on 100 LSA dimensions, is displayed in Figure 6 using a t-SNE projection. It shows that assignments are very similar in both algorithms.



**Fig. 6.** t-SNE projection of clustering algorithms

## 5 Conclusions and future work

Through the use of dimensionality reduction techniques and document clustering algorithms, it was possible to cluster internships based on their descriptions, providing helpful information about the academic internships market.

Spherical k-means produced better clusters than hierarchical clustering when using the original TF-IDF matrix, however they produced similar clusters after performing LSA.

To assert the quality of the clustering algorithms, manual inspection of the clusters was required, as the chosen evaluation metric did not always correlate with human judgments. Additional research could be done regarding topic coherence evaluation metrics, to match more closely to what humans consider meaningful topics.

Another interesting path is to use newer techniques to generate word vectors on both smaller and bigger datasets. Finally, more clustering or topic modeling algorithms could be evaluated.

## References

1. Eurostat. Tertiary Education Statistics - Statistics Explained. url: [https://ec.europa.eu/eurostat/statistics-explained/index.php/Tertiary\\_education\\_statistics](https://ec.europa.eu/eurostat/statistics-explained/index.php/Tertiary_education_statistics). Accessed 17 Nov. 2020.
2. Charu C. Aggarwal. Data Mining: The Textbook. English. 2015 edition. New York, NY: Springer, Apr. 2015. isbn: 978-3-319-14141-1.
3. Zhang Wen, et al. A Comparative Study of TF\*IDF, LSI and Multi-Words for Text Classification. In: Expert Systems with Applications, vol. 38, no. 3, Mar. 2011, pp. 2758–65. ScienceDirect, <https://doi.org/10.1016/j.eswa.2010.08.066>.
4. Frizo Janssens, Wolfgang Glänzel, and Bart De Moor. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. en. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07. San Jose, California, USA: ACM Press, 2007, p. 360. <https://doi.org/10.1145/1281192.1281233>.
5. Mikolov, Tomas, et al. Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Curran Associates Inc., 2013, pp. 3111–3119.
6. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.
7. Meng, Yu, et al. Spherical Text Embedding. Advances in Neural Information Processing Systems, vol. 32, 2019. [experts.illinois.edu, https://experts.illinois.edu/en/publications/spherical-text-embedding](https://experts.illinois.edu/en/publications/spherical-text-embedding).
8. Marwa, Naili, et al. Comparative Study of Word Embedding Methods in Topic Segmentation. Procedia Computer Science, vol. 112, Jan. 2017, pp. 340–49. [www.sciencedirect.com](http://www.sciencedirect.com), <https://doi.org/10.1016/j.procs.2017.08.009>.
9. Zhao Ying, et al. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, vol. 10, no. 2, Mar. 2005, pp. 141–68. Springer Link, <https://doi.org/10.1007/s10618-005-0361-3>.
10. Sculley, D. Web-Scale k-Means Clustering. Proceedings of the 19th International Conference on World Wide Web, Association for Computing Machinery, 2010, pp. 1177–1178. ACM Digital Library, <https://doi.org/10.1145/1772690.1772862>.
11. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of Similarity Measures on Web-Page Clustering. en. In: Workshop on artificial intelligence for websearch (AAAI 2000) 58 (July 2000), p. 7.
12. Chen, Yong, et al. Experimental Explorations on Short Text Topic Mining between LDA and NMF Based Schemes. Knowledge-Based Systems, vol. 163, Jan. 2019, pp. 1–13. ScienceDirect, <https://doi.org/10.1016/j.knosys.2018.08.011>.
13. A. Zanasi. Text Mining and its Applications to Intelligence, CRM and Knowledge-Management. en. WIT Press, Sept. 2007. isbn: 978-1-84564-131-3.
14. Morstatter, Fred, and Huan Liu. In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. Journal of Machine Learning Research, vol. 18, no. 169, 2018, pp. 1–32.
15. Mehta, V., et al. Evaluating Topic Quality Using Model Clustering. 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2014, pp. 178–85. IEEE Xplore, <https://doi.org/10.1109/CIDM.2014.7008665>.
16. Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. In: Journal of Machine Learning Research 9 (2008), pp. 2579–2605.