Grouping Mixed Documents: Mexico Job Offers Case Study

Moreno Galván Elizabeth(✉), Miguel Félix Mata Rivera, and Carmona García Enrique Alfonso

UPIITA-IPN, Av. Instituto Politécnico Nacional 258, 07340 Gustavo A. Madero, Mexico City, Mexico
elizabeth.moreno.galvan.05@gmail.com

Abstract. A mixed dataset is composed of structured and unstructured documents whose heterogeneous data formats complicate not only their processing but also their content analysis. Finding the semantic correspondence among documents stored in a mixed dataset requires identifying, combining, and assembling diverse techniques from many knowledge fields to analyze and reveal possible patterns among them. In the case of text documents, it has been addressed by processing semantic properties and linguistic relationships between words through vector representations and word embedding models. In this paper, we present a methodology to calculate the content proximity in mixed documents, using three techniques: similarity measure, doc2vec embedding model, cosine similarity, and K-means algorithm. The study is centered on the Mexican job market documents to find relationships among mixed documents of job offers. The results show that creating groupings of mixed documents, based on their semantic and cosine similarity, allows them to reveal patterns.

Keywords: Text classification · Mixed dataset · Embedding model · Similarity · doc2vec · K-means · Cosine similarity · Job offer · Vector space model

1 Introduction

Document similarity is a Natural Language Processing NLP [1] technique that has been widely used in various applications such as text classification [2], document semantic similarity detection [3, 4], authorship identification [5], information retrieval (IR) [6], answer questions chatbots [7], text summarization [8], sentiment analysis [9, 10] and applied in other complex study areas like medical or educational to name a few. In labor market area, employers spend great time writing and publishing available jobs information into text documents called job offers in order to find the most suitable professionals whose meet all the requirements needed. In this sense, the job offer is a mixed text document which is formed by structured and unstructured information. To find the correspondence among documents stored in such mixed dataset implies discover associations and patterns, also by clustering and classification. In this sense, what motivates this work is to report experiments applying the combined used of data analysis techniques, to make useful inferences about document similarity problem.

M. F. Mata-Rivera et al. (Eds.): GIS LATAM 2020, CCIS 1276, pp. 151–159, 2020.
https://doi.org/10.1007/978-3-030-59872-3_11

2 Literature Review

Writing a text document that meets some objective, for example a letter or research article is a common task. In some cases comparing such documents with others has become a necessary activity, such as authorship checking or find similar documents to be consulted and referenced in an research. At present, these processes are hindered due to the large number of documents available thanks to digital means, so carrying out search and consultation activities involves exhaustive work too complex to be carried out by a human being. That is why grouping, classifying and find correspondences among text documents, based on content analysis and the measurement of similarities between them, is a task that have been addressed by processing semantic properties and linguistic relationships between words present into documents. In this sense, several techniques have been proposed that can be grouped into two main categories [11]: content-based and knowledge-enriched based:

In the first group, the most used method is the Vector Space Model VSM [12], where each document is represented as a m-dimensional weighted vector and the dimensions correspond to individual characteristics or terms. The result is called the bag-of-words model. The limitation of this model is that it does not take into account polysemy (the same word can have multiple meanings) and synonymy (two words can represent the same concept).

The second group, includes Latent Semantic Analysis (LSA) [13], used to extract a latent semantic structure in documents by applying the reduction of dimensionality to the matrix of term – document. In the same group, Embedding Words, as word2vec [4, 14–18], is a technique that learns to read enormous amounts of texts and memorize which words seem to be similar in different contexts. An analogous method, also known as document embeddings Paragraph Vector or Doc2Vec [17], was presented by the same word2vec author, T. Mikolov [18], which represents documents as fixed-length, low-dimensionality vectors. Regarding unsupervised learning applied with Processing Natural Language NLP, there is a wide preference to apply K-Means technique [19–22] clearly stands out since several algorithms, both unsupervised and supervised, require that the text be preprocessed within the vector space, also implementing techniques such as bag of words, in addition to the use of other more such as binarization, tokenization, linguistic analysis and stemming.

As has been showed, the main difference between those techniques is that the first group uses only textual information contained within documents, while the second group enriches these documents by extracting information from other sources, usually knowledge bases.

3 Data Collection

The data was extracted from words that correspond to a job offers for some common professions such as Doctor, Lawyer, Secretary or Business Management. A total of 10,682 job offer records were collected from the Linkedin job search platform website. The dataset consists of all job offers published in spanish, with location in Mexico City. Table 1 describes the general structure of the dataset.

Table 1. Linkedin document sections.

Section	Description	Structured
Título	Job title offered	Yes
Empleador	Company name	Yes
Ubicación	Workplace	Yes
Descripción	Job offer details such as salary, schedule, requirements or functions	No
Nivel de experiencia	Experience Level: Prácticas/Sin Experiencia/Algo de responsabilidad/etc.	Yes
Tipo de empleo	Workday: Jornada completa/Medio tiempo	Yes
Función Laboral	Job function: Desarrollo empresarial/Ventas/Gestión de proyectos/Tecnologías de la Información/Consultoría/Ingeniería/, etc.	Yes
Sectores	Business sector: Alimentación y bebidas/Recursos Humanos/Ventas al por menor/Telecomunicaciones/Contabilidad/, etc.	Yes

4 Materials and Methods

4.1 Methodology

Based on the KDD methodology which has been widely used for data analysis, the applied methodology parts of the conception that there are groups of mixed documents available on the social network, thus consists of three main phases subdivided into phases ranging from 1) data collection to conform the experimental data warehouse, 2) pre-processing phase applying natural language techniques, for later apply analysis and also grouping techniques, and finally 3) present the results obtained which will show the possibility of grouping mixed text documents.

By this method, shown in Fig. 1 the resulting analysis will be obtained from the application of statistical, analytical, NLP, machine learning and data mining tools, and the generation of visualization models that help illustrate the value of information, making it a valuable resource for creation of business strategies.

The methodology description is resumed as follows:

4.2 Phase 1 Collection and Storage

To the conformation of the dataset, since obtaining data from the social media, the data source must have the property of being mixed, since as mentioned above, the treatment and combination of applied techniques is focused on this type of document. In this way, every document characteristic such as title, is stored as a study dimension.

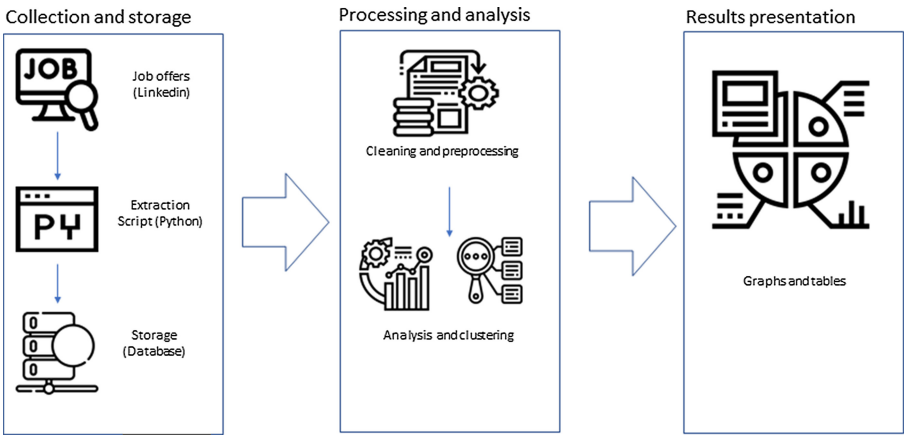

Fig. 1. Followed methodology.

4.3 Phase 2 Processing and Analysis

The computational treatment of textual information involves a mathematical modeling process and the use of paragraph embeddings technique, in which the texts are treated as objects and the words are represented in the vector space. Once the VSM has been obtained, it is possible to perform operations such as finding similarities between the vector representations by applying techniques such as calculating distances by cosine similarity [18] (see Fig. 2).

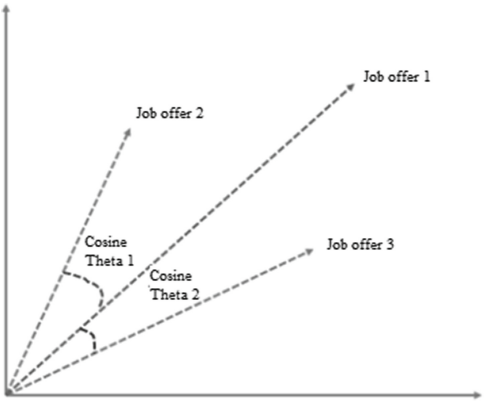

Fig. 2. Vectorial representations of job offers

The technique for grouping job offer documents proposed in this methodology is the K-means algorithm [9], one of the most widely used non-hierarchical cluster methods, which divides existing data into one or more clusters.

4.4 Phase 3 Results Presentation

To present the grouping results from the combined application of content analysis techniques, the representation was chosen by dispersion diagrams with colors that graphically support the distinction of the groups obtained and the differences between them obtained from each exploration.

5 Results

In order to carry out the documents grouping and classification, two experiments were carried out: grouping job offers by cosine similarity and grouping by K-Means algorithm.

5.1 Grouping Job Offers by Cosine Similarity

The corpus of documents corresponding to job offers from various fields was divided into training (70%) and test sets (30%). Thus, the model was trained with 7153 records in 100 epochs, with a window of 10 words. Later, to verify that it has learned all the words and if they have a contextual meaning, the search for the words "Sales", "Medical" and "Lawyer" was carried out, whose scatter plots are shown in Fig. 3 showing all the documents that are close or similar to the test documents contextually.

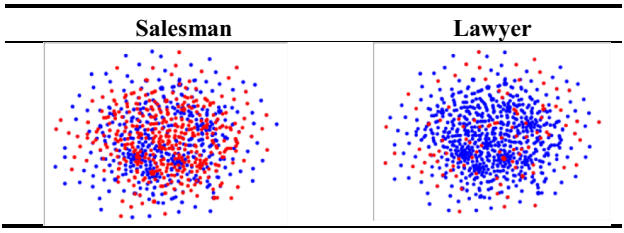

Fig. 3. Similarity between job offer documents by doc2vec testing model

From the doc2vec model, a *corpus* of job offer documents were compared one by one with the rest of the *corpus* using the cosine similarity formula 1.

$$similarity = cos\,\theta = A \cdot B \Big/ ||A||\,||B|| = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Since the similarity score is between 0 or 1, a minimum threshold can be set such as 0.6, in this way similar document sets were created with a similarity score of any pair greater than 0.6.

A pair of experiments where applied:

i) In the first experiment, it has been obtained a word2vec model, in which, the records were grouped from all the words contained in the dataset and its corresponding context.

ii) The second experiment consists in the application of doc2vec model, Fig. 4 shows a fragment of the set of similar documents obtained through the calculation of the 60% cosine similarity of a document (id: 369), using the doc2Vec trained model.

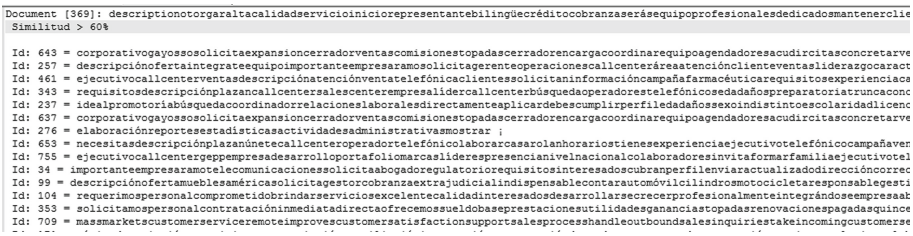

Fig. 4. Fragment of a list of documents similar to a job offer

Table 2 summarizes the groupings obtained as a result from both experiments whose cosine similarity calculation was greater than 60% (0.6).

Table 2. Groups by cosine similarity over 60%.

Embedding technique	Similarity score	Groups formed	Score of records inside each group	Dataset length
Word2Vec	60%–65%	255	20–30	10 682
Word2Vec	66%	1671	4	10 682
Word2Vec	Over 67% to 70%	7	3–4	10 682
Doc2Vec	60%	14	70–245	3311
Doc2Vec	60%	328	25–89	7153

Table 2 exemplifies that by doc2vec embedding model, the documents can be classified more effectively than by word2vec embedding model.

5.2 Grouping by K-Means Algorithm

Through this experiment, a training set from the *corpus* of job offers was modeled by the word2vec vector model. After the K-means algorithm application was produced the scatter plot in Fig. 5.

The graph shows that the words contained in job offer documents are very similar in context, so this method does not offer groupings that allow any representative distinction to be made between them.

On the other hand, an experiment where a set of the of job offers *corpus* were trained by doc2vec vector model and grouped using the K-means algorithm produces the Fig. 6 scatter plot.

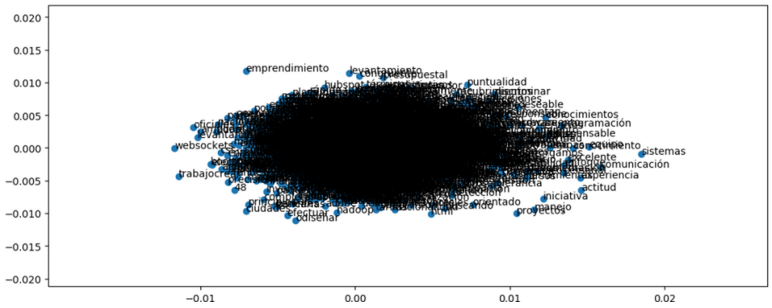

Fig. 5. Fragment of a list of documents similar to a job offer

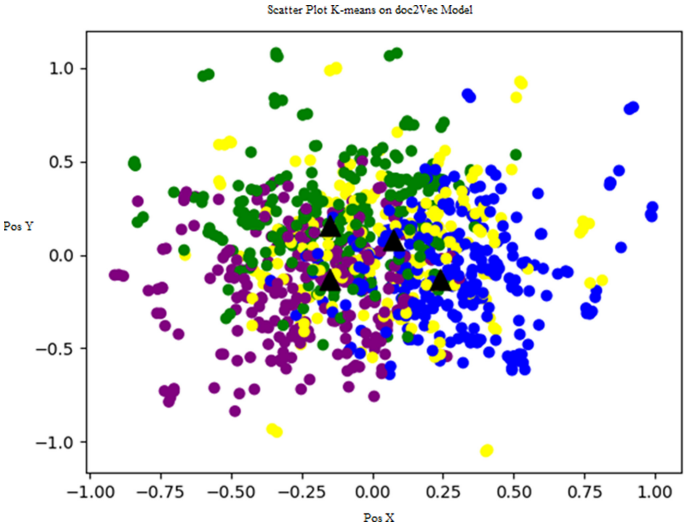

Fig. 6. Clustering by K-means algorithm application on doc2vec modeled set

As it can be observed, document sets have been obtained keeping some relation to each other, so it is possible to detect constant associations between documents.

6 Conclusions

In this paper, a methodology to explore job market in Mexico has been presented. emphasis is placed on focusing the study on detecting the causes of the phenomenon beyond the statistical approach. In this sense, the difficult in the analysis of documents from de job market such as job offers, lies in that the documents are not structured. To support the study of datasets containing this kind of documents, the document embedding model has been a recent technique, but that has shown visible advantages for content analysis and the detection of semantic relationships between documents, therefore, throughout this work some of the classic techniques of classification and text preprocessing have been proved within embedding models.

Experiments shown the great difficult to group the job offer documents from the word2vec model, since it has not generated satisfactory results, because it has been detected that within the job offer documents there is a common terminology, with few differentiators making it difficult to distinguish between them. However, to group the same documents from the doc2vec model, has given very good results allowing clustering by both proved techniques: cosine similarity and k-means, which means it is possible to create groups from similarity founded in documents no matter their mixed nature.

On the other hand, when it is required to work with other types of associated documents, such as resumes, which are used by recruiters to find the best candidate for the job offered, the problem of finding the person-job relationship is based on that there is no categorization, standardization or regulation that defines what the characteristics of a job offer are and this study opens up a study area that can help the adequate construction of these documents.

7 Future Work

The development of a methodology focused on finding correspondence, associations and patterns between job offers is proposed to determine if there is any grouping criterion by measuring similarity between them. To this end, work will be carried out on a greater number of experiments and analyzes, such as candidates profiles and relations between both, job offers and profiles with the expectation of impacting in a multidisciplinary way on the social and economic aspects of Mexico.

References

1. Oghbaie, M., Mohammadi Zanjireh, M.: Pairwise document similarity measure based on present term set. J. Big Data **5**(1), 1–23 (2018). https://doi.org/10.1186/s40537-018-0163-2
2. Aldrin, F.J., Zapata, C.M., Isaza, F.A.: Una Propuesta para la Asistencia al Proceso de Interpretación de Textos utilizando Técnicas de Procesamiento del Lenguaje Natural e Ingeniería de Software. In: Avances en Sistemas e Informática Proceedings, vol. 4, no.3, Medellín, ISSN 1657-7663 (2007)
3. Palmer, Z., Wu, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)
4. Elekes, A., Englhardt, A., Schäler, M., Böhm, K.: Toward meaningful notions of similarity in NLP embedding models. Int. J. Digit. Libr. **21**, 109–128 (2020). Springer-Verlag GmbH Germany, part of Springer Nature (2018)
5. Rozi, F., Sukmana, F.: Document grouping by using meronyms and type-2 fuzzy association rule mining. J. ICT Res. Appl. **11**(3), 268–283 (2017)
6. Alonso, I., Contreras, D.: Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach. Expert Syst. Appl. **44**, 386–399 (2016)
7. Robert, D.: The return of the chatbots. Nat. Lang. Eng. **22**(5), 811–817 (2016). Cambridge
8. Mutlu, B., Sezer, E.A., Akcayol, M.A.: Multi-document extractive text summarization: A comparative assessment on features. Knowl.-Based Syst. **183**, 104848 (2019)
9. Vilares, D., Gómez, C., Alonso, M.A.: Universal, unsupervised (rule-based), uncovered sentiment analysis. Knowl.-Based Syst. **118**, 45–55 (2017)

10. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)
11. Benedett, F., Beneventano, D., Bergamaschi, S., Simonini, G.: Computing inter-document similarity with Context Semantic Analysis. Elsevier Inf. Syst. **80**, 136–147 (2019)
12. Sidorov, G.: Syntactic n-grams in Computational Linguistics, pp. pp. 5–19. Springer, Cham (2019)
13. Botana, G.: La técnica del Análisis de la Semántica Latente (LSA/LSI) como modelo informático de la comprensión del texto y el discurso, Tesis Doctoral, U. Autónoma de Madrid (2010)
14. Rezaeinia, S.M., Rahmania, R., Ghods, A., Veisi, H.: Sentiment analysis based on improved pre-trained word embeddings. Expert Syst. Appl. **117**, 139–147 (2019)
15. Nguyen, H.T., Duong, P.H., Cambria, E.: Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowl.-Based Syst. **182**, 104842 (2019). Elsevier
16. Church, W.K.: Emerging trends: Word2Vec. Nat. Lang. Eng. **23**(1), 155–162 (2017)
17. Markov, I., Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., Gelbukh, A.: Author profiling with Doc2vec neural network-based document embeddings. In: Pichardo-Lagunas, O., Miranda-Jiménez, S. (eds.) MICAI 2016. LNCS (LNAI), vol. 10062, pp. 117–131. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62428-0_9
18. Quoc, V.L., Mikolov, T.: Distributed Representations of Sentences and Documents. Google Inc., Mountain View (2014)
19. Usino, W., Prabuwono, A.S., Hamed, K., Allehaibi, S., Bramantoro, A., Hasniaty, A., Amaldi, W.: Document similarity detection using k-means and cosine distance. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **10**(2) (2019)
20. Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, Y., Kambhatla, N.: Prospect: a system for screening candIdates for recruItment. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Managementpp, pp. 659–668. ACM, (2010)
21. Li, Y., Tripathi, A., Srinivasan, A.: Challenges in Short Text Classification: The Case of Online Auction Disclosure. Association for Information Systems, AIS Electronic Library (AISeL) (2016)
22. Kumar, A., Pandey, A., Kaushik, S.: Machine learning methods for solving complex ranking and sorting issues in human resourcing. In: IEEE 7th International Advance Computing Conference (2017)