Word Embeddings to analyze Peruvian computing curriculums

Jeffri Murrugarra-Llerena [1] **Nils Murrugarra-Llerena** [2]

Abstract

Nowadays, there is a lack of guidelines in computing curriculums in Peru. Most of them do not follow the ACM/IEEE international standards. Also, accreditation of computing programs is a time-consuming and manual task. In order to tackle these issues, we propose an automatic process to identify a curriculum according to ACM/IEEE standards. Our initial solution combines word embedding and visualization techniques. Our results confirm disorder and confusing in Peruvian curriculums. For future work, we plan to collect more data and try more elaborate techniques such as metric learning and recurrent neural networks.

1. Introduction

ACM/IEEE standard is accepted by several universities around the world and has been successfully implementing in top-tier universities such as Illinois, MIT, and Berkeley. ACM groups computing careers in five groups: Computer Science (CS), Computer Engineering (CE), Software Engineering (SE), Information Science (IS), and Information Technology (IT) (Shackelford et al., 2005). However, Latin American universities had incorporated some of these groups, and in some universities, it is difficult to categorize its computing careers into the previous five groups. This confusion happens because some computing careers are a mixture of the five core groups (Sabin et al., 2016).

Specifically, Peru has approximately one hundred computer careers nationwide. Along with these careers, there are twenty-eight different denominations. Many careers have similar names (Final, 2018), but they have different curriculums and inconsistencies in what is offered (Colegio de Ingenieros, 2006). This may result in confusing guidelines to identify computing careers in Peru, and professors are not highly specialized in core areas. Also, students acquire questionable skills that do not follow international guidelines. This dubious preparation also affects companies, because they can not find job candidates to fulfill all their necessities. (Sabin et al., 2016).

In the literature, we find some works that examine university curriculums. (de Alburquerque et al., 2010) and (Prietch, 2010) employ curriculums to analyze the market offer and propose to standarize curriculums in the Brazilian context. Also, the CS curriculum from the University of São Paulo is examined every year to determine if its quality follows the current innovations (Macêdo, 2016). Although the goal of these works are different, they have in common that their process is manual, which is time-consuming. For example, curriculums analysis from the accreditation process takes around three months (ICACIT, 2019). (Murrugarra-Llerena et al., 2011) analyzes curriculums from Peru and Brazil with a semi-automatic approach using only course titles, which may produce incomplete results due to disorder in Peruvian curriculums. Complementary, we aim to tackle this problem and contribute with an automatic tool that provides a fast, simplified and efficient process.

We evaluate two-word embeddings on computing curriculums. Then, we analyze our results with a non-linear visualization technique and hierarchical clustering. We find strong relations between Peru and USA curriculums, but also overlapping data.

2. Approach

Dataset: We collected curriculums of CS, CE, SE, IT and IS programs from the United States and Peru, which are grouped below. Each curriculum is organized in a text file, that contains the course titles and their descriptions (elective and mandatory courses)[1].

- **Curriculums according to ACM recommendations**: We select the best universities in USA, who also appears in ABET program[2]. In total, we collected 25 CS curriculums, 25 CE curriculums and an additional 25 combining SE, IT and IS. We collected them with web

[1]Department of Computer Science, Federal University of Rio Grande do Sul, Porto Alegre, Brazil [2]Snap Inc, Los Angeles, USA. Correspondence to: Jeffri Murrugarra-Llerena <ju.jeffri.v@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]https://github.com/Artcs1/DL_CURRICULUM/tree/master/Data

[2]https://github.com/Artcs1/DL_CURRICULUM/blob/master/list_universities_usa.x

scrapping and beautiful soup libraries from python.

- **Curriculums from Peru**: We collected and translated nine curriculums of the most representative peruvian universities: UNT, PUCP, UTEC, three of UPC, UNH, UNJFSC, UNP.

Data preparation: First, we convert all text to lower case and extract all words by tokenization. Then, we verified that all characters are in UTF-8 format, remove stop words with NLTK library and finally, we ignore words that appear in less than 5% and more than 95% of the documents.

Feature extraction: We used Gensim library[3] to load word2vec (Mikolov et al., 2013) and glove (Pennington et al., 2014) models, that were trained in wiki-corpus (100,000,000 bytes of plain text from Wikipedia) and Gigaword5 + Wikipedia 2014.

After model loading, we consider absent words. We created a random vector to represent unknown words. Then, each curriculum is represented as the average of their constituent words.

Evaluation: We performed experiments with T-SNE (van der Maaten et al., 2008) visualization technique for a deeply understand of our embeddings. Also, we performed a Hierarchical clustering analysis with ward linkage to understand similarities.

3. Discusion

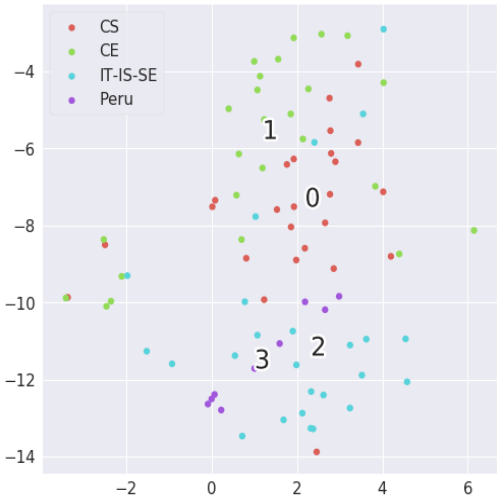
We used T-SNE to verify if our embeddings ensure cohesive and separable groups for classification. In Figure 1 (a) and (b), we find that IS, IT and SE are related for T-sne, which ensures that our union due to the lack of data is consistent. Also, CS overlaps with the other careers and it is in the middle of them. This is an interesting finding because CS is the core career that links all of the 5 computing programs.

We also note that Peruvian curriculums are related to USA groupings, and they do not present any outlier. We observe that most of them are related to IT, IS and SE. However, in our data, only three careers claim that associations. This confirms that there is disorder in Peruvian programs.

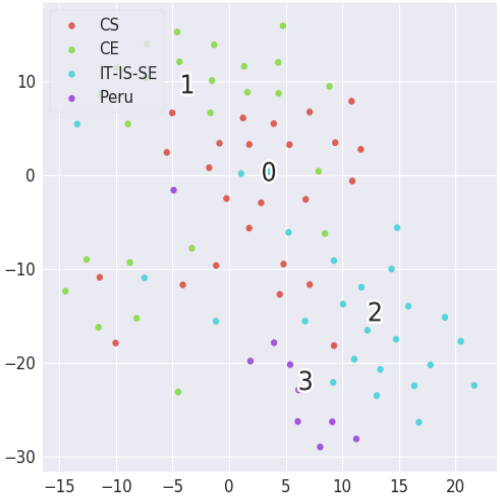
To further understand the similarity between the curriculums, in Figure 1(c), we employ hierarchical clustering and we observe that the three main groups are preserved, but the overlap is still present.

In future work, we will increase our data from Peru and USA. We also will use metric learning to obtain better embeddings. First, we will learn a metric space using the USA data, and then evaluate them on the Peruvian data. We expect to find more cohesive and separable groups among the

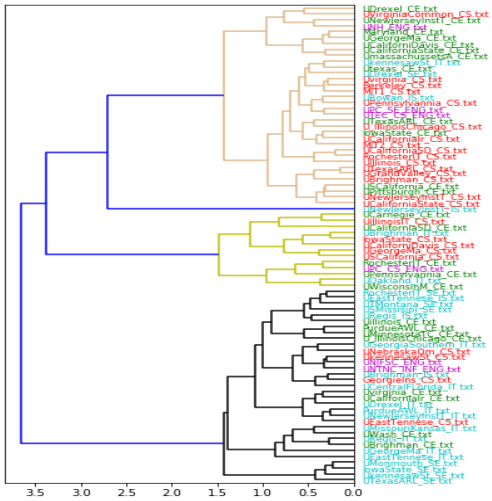
[3]https://github.com/RaRe-Technologies/gensim-data



(a) T-SNE visualization for word2vec embbedings



(b) T-SNE visualization for glove embbedings



(c) Dendogram for glove embbedings

Figure 1. Results of USA and Peru curriculums

five core standards. Also, we can model the problem as a classification task, which will follow a similar procedure as before. We aim to classify Peruvian curriculums among the USA organization. Finally, we aim to experiment with recurrent neural networks to preserve order information on the curriculums and find better embeddings.

References

Colegio de Ingenieros, P. Denominaciones y perfiles de las carreras en ingeniería de Sistenams, Computación e Informática. *Consejo Departamental de Lima*, pp. 143, 2006.

de Alburquerque, P. L. Z. et al. Uma Análise da Oferta e Abordagem Curricular dos Cursos de Bachelorado em Sistemas de Informação no Brasil. *WEI*, pp. 897–906, 2010.

Final. Plan curricular 2018 de la escuela profesional de ciencia de la computacion. *UNSM*, pp. 25, 2018.

ICACIT. Accreditation process. urlhttp://www.icacit.org.pe/web/acreditacion/sobre-acreditacion/ciclo-de-acreditacion.html, 2019.

Macêdo, B. D. Nova grade curricular do BCC-IME-USP. *WEI*, pp. 10, 2016.

Mikolov, T. et al. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*, 2013.

Murrugarra-Llerena, N. et al. Comparacao de Grades Curriculares de Cursos de ˜ Computacao Baseada em Agrupamento Hierarquico de Textos. *WEI*, 2011.

Pennington, J. et al. Glove: Global vectors for word representation. pp. 12, 2014.

Prietch, S. S. Mapeamento de Cursos de Licenciatura em Computação seguido de Proposta de Padronização de Matriz Curricular. *WEI*, pp. 921–930, 2010.

Sabin, M. et al. Latin American Perspectives to Internationalize Undergraduate Information Technology Education. *Working Group Report*, pp. 22, 2016.

Shackelford, R. et al. Computing curricula 2005. Technical report, ACM, IEEE and AIS, 2005.

van der Maaten, L. et al. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, pp. 2579–2605, 2008.