Word Embeddings to analyze Peruvian computing curriculums

Jeffri Murrugarra-Llerena and Nils Murrugarra-Llerena
Federal University of Rio Grande do Sul, Snap Inc

Introduction

- ACM/IEEE standard is accepted by several universities around the world and has been successfully implemented in top-tier universities. ACM groups computing careers in **Computer Science** (CS), **Computer Engineering** (CE), **Software Engineering** (SE), **Information Science** (IS), and **Information Technology** (IT)**.**

Motivation

- Latin American universities had incorporated some of these groups, but in some universities, it is **difficult to identify them**.
- Peru has approximately **100 computing careers** nationwide and there are **28 different denominations**. So many careers have similar names, but they have different curriculums and inconsistencies in what is offered **[8]**. This may result in **confusing guidelines** to identify computing careers in Peru.

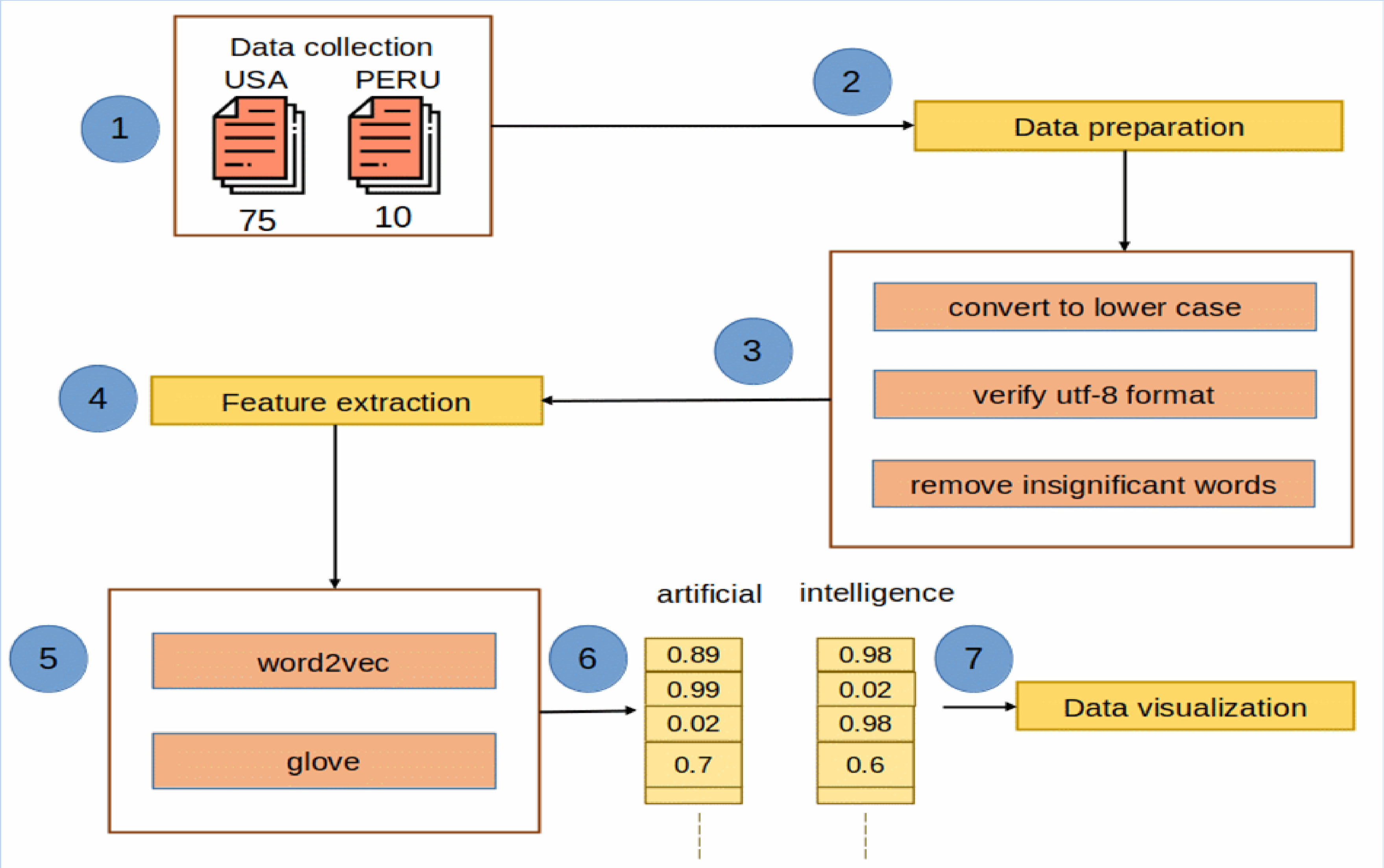
Key idea

- To tackle this problem we compile curriculums among universities from USA that follow ACM standards (accepted by **ABET accreditation program**) and compare with Peruvian ones. Then, we performed visualization techniques to compare these two countries data and determine how complex is the problem.
- We assume three main groups **CS, CE, IS-IT-SE** due to the lack of data
- The last group is also related to **organizational needs**.

Related work

- Prior work employ curriculums to analyze the market offer and propose to standardize curriculums in the Brazilian context, but have a manual process that is time-consuming **[4]**.**[5]** 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 accurate process.

Approach



1 Data collection
USA 75
PERU 10

2 Data preparation

3
convert to lower case
verify utf-8 format
remove insignificant words

4 Feature extraction

5
word2vec
glove

6

artificial	intelligence
0.89	0.98
0.99	0.02
0.02	0.98
0.7	0.6

7 Data visualization

- We collected curriculums of **CS, CE, SE, IT and IS** programs from the United States and Peru. Each curriculum is organized in a text file, that contains the course titles and their descriptions.
- 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.
- We used **Gensim library** to load **word2vec [1]** and **glove [2]** models, that were trained in **wiki-corpus** and **Gigaword5**. Then, each curriculum is represented as the average of their constituent words.
- We performed experiments with **T-SNE [3]** visualization technique for a deeply understand of our embeddings. Also, we performed a Hierarchical clustering analysis with ward linkage to understand similarities.

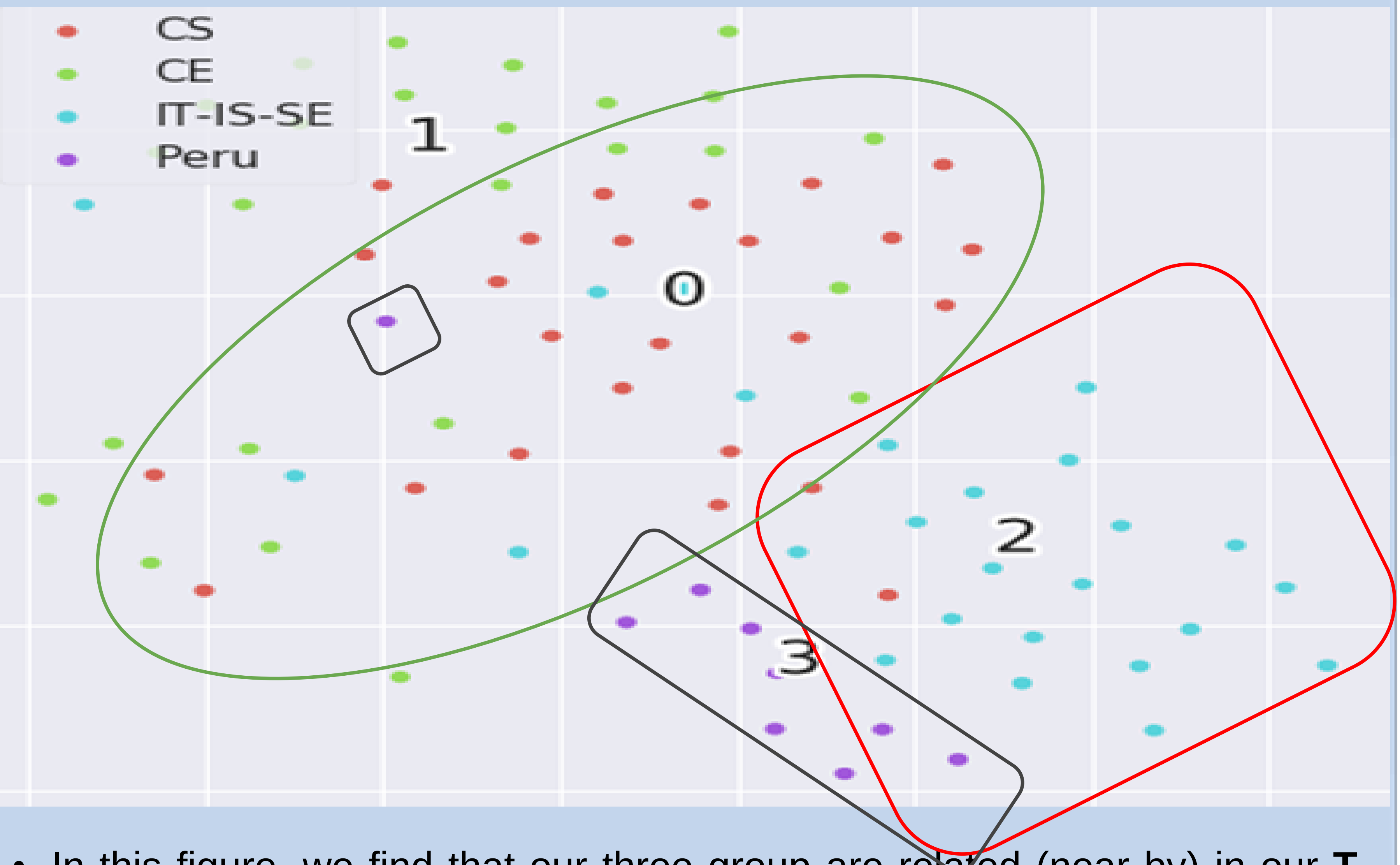
Experimental setup and Dataset

- We do a compilation of curriculums, each curriculum has the name of their courses and a description. We do that with web scrapping technique.

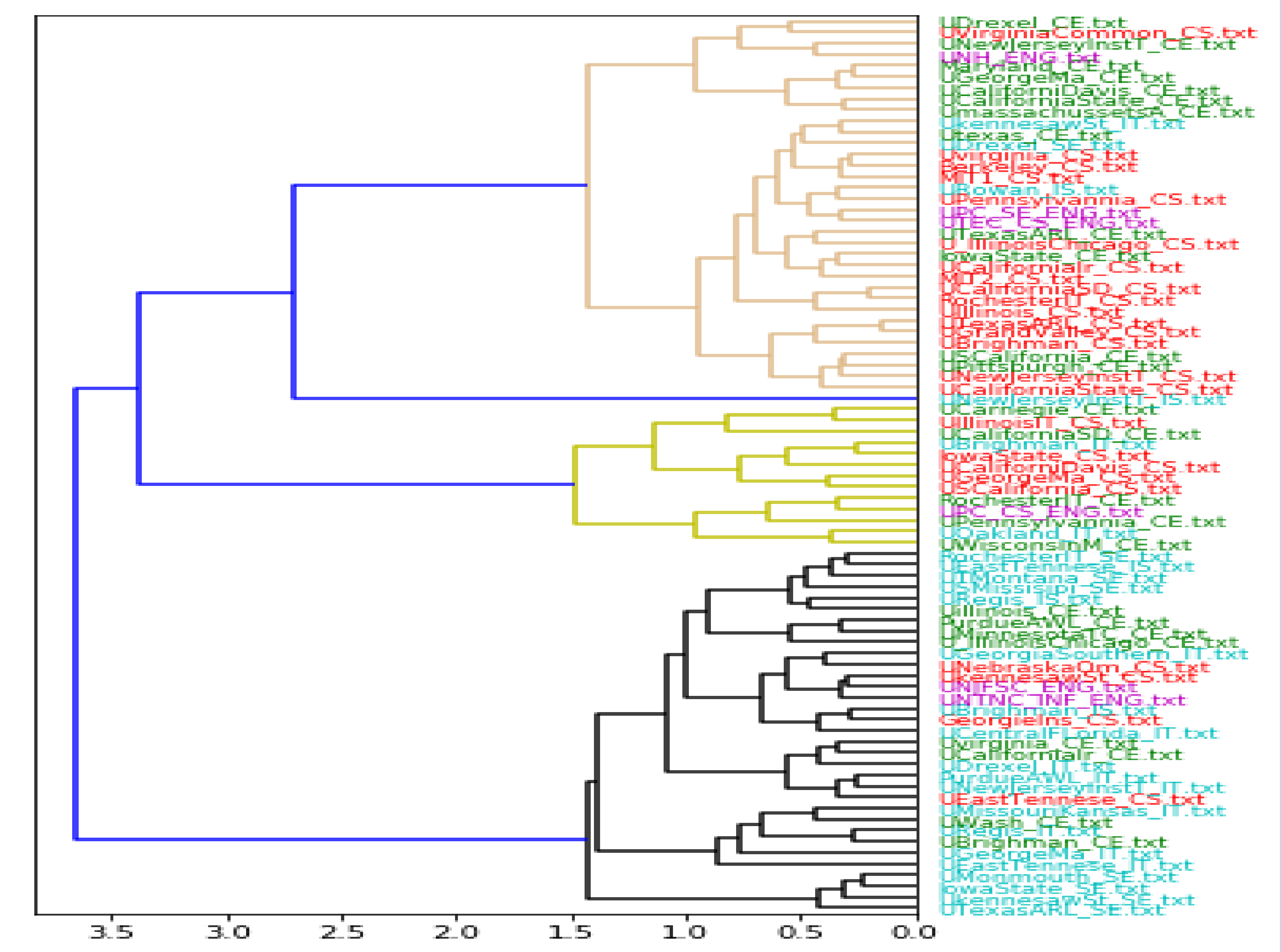
USA curriculum	CS	CE	Other
Goal	100	100	100
Current	25	25	25
In Future	73	74	74

- Data from Peru are from the most representative universities

Qualitative results



- In this figure, we find that our three group are related (near by) in our **T-sne visualization**, which ensure our union due to the lack of data is consistent.
- In the green ellipse, **CS overlaps with the other careers** and it is in the middle of them. CS is the core career of this five.
- From the black squares, we also note that Peruvian curriculums are related to USA groupings, and **they do not present any outlier**. Also, we observe that most of them are related to our three groups. However, in our data, only three careers claim that associations.
- This confirms **disorder in Peruvian programs** and **difficulty of categorize with naive techniques**.



- To further understand the similarity between the curriculums, in the previous figure, we employ hierarchical clustering and we observe that the **three main groups** are preserved, **but the overlap** is still present.

Contributions

- A new dataset that contains a description of each course. This dataset is available on **https://bit.ly/3eqmJk0**. We believe this dataset will expand research in computing analysis and will server other researchers.
- We do a preliminary analysis about the current state of computing curriculums in Peruvian universities.

Future work

- In future work, we will increase our data from Peru and USA and apply more sophisticated techniques as metric learning and **RNN(LSTM [6]**, **GRU [7]**) to preserve order information on the curriculums and find better embeddings.

References

[1] Tomas Mikolov (2013) Efficient estimation of word representations in vector space. International Conference on Learning Representations
[2] Jeffrey Pennington (2014) Glove: Global vectors for word representation
[3] Laurens van der Maaten (2008) Visualizing data using t-sne. Journal of Machine Learning Research 9.
[4] Pereira L. Z. de Alburquerque (2010) Uma Análise da Oferta e Abordagem Curricular dos Cursos de Bachelorado em Sistemas de Informação no Brasil. WEI
[5] Murrugarra-Llerena, Nils (2011) . Comparação de Grades curriculares de Cursos de ̃ Computação Baseada em agrupamento Hierárquico de Textos.WEI, 2011.
[6] Sepp Hochreiter (1997). Long Short Term Memory. Neural Computation 9.
[7] Yoshua Bengio (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS 2014
[8] Colegio de Ingenieros(2006) Denominaciones y perfiles de las carreras en ingeniería de Sistemas, Computación e Informática. Consejo Departamental de Lima