Monograph presented to the Department of Electrical Engineering at PUC / Rio as part of the requirements for obtaining the title of Specialist in Business Intelligence.

Correspondence of medical records to clinical trials for patient eligibility using NLP

Mônica Soares Brandão

Supervisors: Prof. Leonardo Alfredo Forero Mendoza, PhD

Collaboration: Prof. Cristian Muñoz, PhD

Received: December 6th, 2020

English Version

Abstract: A clinical trial is a type of scientific study used in medicine, psychology, and other sciences. Is used to test the effectiveness of a given therapeutic approach in a patient population, or to collect information about side effects of a given treatment. One of the phases of a clinical trial process is the recruitment of patients, where they will be exposed to the tests proposed by the trial.

Each clinical trial has eligibility criteria or restrictions that can include or exclude patients, making them eligible or not for the study. An efficient patient recruitment is crucial to the smooth running of the clinical trial. But it is also an arduous process for recruiters try to find patients who meet these criteria.

Natural language processing is the technology used to help computers understand the natural language of human beings. Natural language processing, usually abbreviated as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using Natural Language.

The ultimate goal of NLP is to read, decipher, understand and make sense of human languages in a way that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages.

1. Introduction

Evidence-based medicine relies heavily on clinical trials to translate research into clinical practice [1] [2]. For this reason, of the various types of experimental studies, the most frequently used and that provides stronger evidence, is the randomized clinical trial (RCTs - Randomized Clinical Trial).

Different from observational studies in which the researcher does not interfere with exposure, this study the researcher plans and actively interfere in the factors that influence the individuals in the sample.

Good clinical practice for recruiting patients is challenging for the clinical trial process. Finding patients who meet the eligibility criteria has been a manual, arduous and time-consuming process for recruiters according.

Each clinical trial has a protocol in which it clearly describes who is eligible to participate in the study [3].

In this work we focus on clinical trials that are in the recruitment stage.

Natural Language Processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process successfully large amounts of natural language data (text and speech) [4].

The technologies based on NLP, are becoming more and more widespread. For example, phones and portable computers support pre-visible text and handwriting recognition; web search engines give access to information closed in unstructured text; translators allow us to retrieve texts written in Chinese and read them in Spanish.

The use of NLP allows to provide more natural human-machine interfaces and sophisticated access to stored information, linguistic processing has started to play a central role in the multilingual information society. It also allows to relate words and documents grouping in clusters (Figure 3) which is the main approach used in this work.

2. Objectives

This work was developed with the purpose of demonstrating that it is possible to optimize and streamline the process of clinical trials recruiting, suggesting eligible patients, through Artificial Intelligence, listing some techniques that can meet this purpose.

Today many problems have been solved using machine learning (Machine Learning-ML) and Neural Networks (Neural Network) that help to mine data in a fast and efficient way using also statistical models that can have high accuracy being quite assertive .

For this work, we will work with 288 patient records from the NLP Re-search Data Set - N2C2, kindly provided by Hardware Medical School and with only one clinical trial related to Diabetes Mellitus, chosen from the ClinicalTrail.gov website.

The patient records for being unstructured texts, free text narratively written, with medical terminologies or ontologies, we realized that the use of Natural Language Processing or English, Natural Language Processing (NLP), would be effective, granting more speed and accuracy.

We will outline on the next pages, existing NLP techniques, showing the approach of each one as well as the results that each one achieves. The learning model that best fits this objective is unsupervised learning. Different of supervised learning that we simply describe as, one that, from a previously defined set of labeled data, wants to find a function that is capable of predicting unknown labels, the unsupervised learning once there is no previously defined labeled data set, aims to discover similarities between objects.

We will use the following techniques here:

- TF-IDF
- Cosine Similarity
- LDA
- Doc2Vec

3. Materials and methods

Finding patients who are eligible from thousands of medical records in different hospitals seems an impossible task. Medical portfolios are free text and are usually written in a narrative style, which can vary from one professional or health institution to another.

With this, medical terminologies and ontology were created to standardize and standardize medical terms in clinical records.

The clinical trial chosen at clinicaltrials.gov was NCT03986073, related to type 2 Diabetes Mellitus [5]. All 288 patient records provided by Harward Medical School N2C2 NLP Research Data Sets were analyzed in this work and were in XML format.

The idea in this paper is, from a clinical trial previously chosen at ClinicalTrial.org, to use some NLP techniques to suggest which patient records from Harward Medical School N2C2 NLP Research Data Sets, which are unstructured notes from the data repository of Partners Healthcare research patients [4], most correlate with the clinical study and in terms of being used as a suggestion of eligibility for recruitment for the clinical trial in question.

In the first analysis, as there is no definition of a predictor variable, we are facing an unsupervised model problem, and this will be the path adopted in this work.

In addition, as a textual problem, Natural Language Processing (NLP) applies perfectly.

The computer language chosen for this work was Phyton and the platform Google Collaboratory.

3.1 Methodologies

As an unsupervised model it is possible in this work to use four approaches:

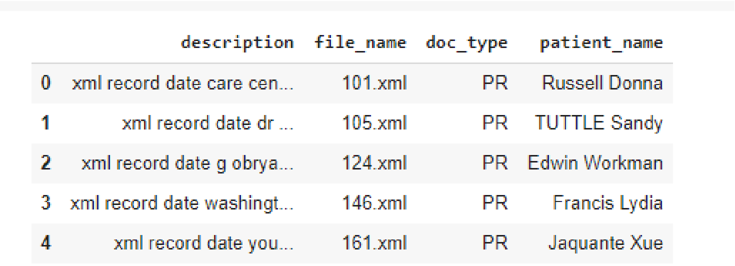
- T-SNE clustering by neighborhood
- Cosine Similarity
- Doc2Vec training
- LDA topics

The Term Frequency-Inverse Document Frequency representations (Tf-Idf) and bag of words will be also used.

The purpose of exploring the different approaches is to show their differences and the applicability of each in the process of patient eligibility for a clinical trial.

3.1.2. Pre-processing text

Medical records are narrative texts and therefore it was necessary to clean them to remove words or symbols that are not relevant to the model, so as capital letters to avoid any sensitive case, punctuation, numbers and stop words that we will see later. That is why for all medical records the text was cleaned (Figure 1)



	description	file_name	doc_type	patient_name
0	xml record date care cen...	101.xml	PR	Russell Donna
1	xml record date dr ...	105.xml	PR	TUTTLE Sandy
2	xml record date g obrya...	124.xml	PR	Edwin Workman
3	xml record date washingt...	146.xml	PR	Francis Lydia
4	xml record date you...	161.xml	PR	Jaquante Xue

Figure 1 – Data pre-processed

3.1.3. Tokenization

It is part of the pre-processing and its purpose is to section a textual document in minimum units, which express the same original semantics of the text. The term is used to designate these units, which often correspond to only one word in the text.

After cleaning the text, the tokenization was performed, separating the words and generating a list of words.

3.1.4. Stop words

Apply Stop words will remove common terms or everything that can be disregarded from the text, without changing its meaning. In the case of this work, we need to apply the stop words to remove the words without relevance from our “corpus” (bag of words).

3.1.5. Term frequency-inverse document frequency (tf-idf)

The TF-IDF weight is a weight that is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate the importance of a word for a document in a collection or corpus. The importance increases proportionally to the number of times that a word appears in the document, but it is compensated by the frequency of the word in corpus [8]. Tf-IDF is calculated by multiplying a local component such as term frequency (TF) with a global component, that is, inverse document frequency (IDF) and optionally normalizing the result to unit length. As a result, words that occur frequently in documents will be reduced [8].

We used the TF-IDF to compare the clinical trial file against the 288 patient records to try to find the records most similar to the clinical trial chosen. Since each record of a single patient, finding the records most similar to the clinical trial, we can suggest that these patients have a high chance of eligibility.

3.1.6. Cosine similarity

Cosine Similarity is a metric used to determine how similar documents are, regardless of their size.

Mathematically, it measures the cosine of the angle between two projected vectors other than zero in a multidimensional space that measures the cosine of the angle between them. The cosine of 0 ° is 1 and is less than 1 for any angle in the interval (0, π] radians.

In this context, the two vectors are matrices containing the words count of two documents. When plotted in a multidimensional space where each dimension corresponds to a word in the document, the similarity cosine captures the orientation, the angle of the documents, and not the magnitude (for the magnitude should calculate the Euclidean distance).

The Cosine Similarity is advantageous because even if the two similar documents are distant by the Euclidean distance because of the size, they could still have a smaller angle between them. The smaller the angle, the greater the similarity.

.

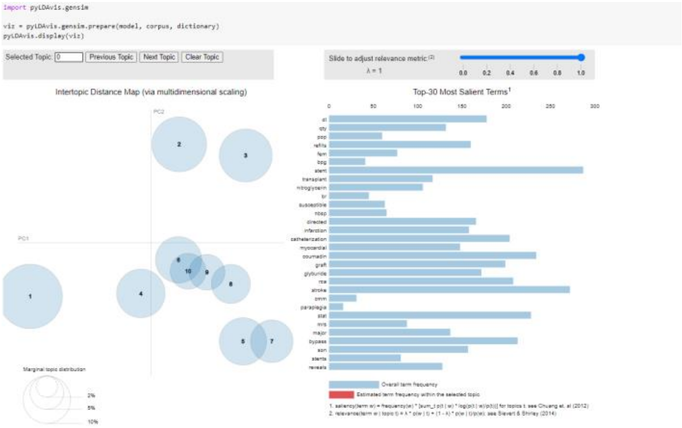
3.1.7. LDA - Latent dirichlet allocation

In natural language processing, the Latent Dirichlet Allocation of English Latent Dirichlet Allocation (LDA) is a model that assumes that documents are nothing more than a probability distribution of topics and topics nothing more are than a word probability distribution, the LDA calculates the probability that a document is mainly this topic or that topic (for example, document N is 77% topic 1, 10% topic 2, 8% topic 5 and 5% topic 7) based on the words it contains [6].

LDAvis are tools for creating an interactive visualization using a web page, from a topic model that has been adjusted to a corpus of text data using Latent Dirichlet Allocation (LDA). Given the topic model's estimated parameters, it calculates various summary statistics as input to an interactive view created with D3.js that is accessed through a browser. The goal is to help users interpret the topics in their LDA topic model.

The LDA topic model algorithm requires a document word matrix as the main input that can be created previously in CountVectorizer.

After using LDA, the result was very dispersed at the topical level and did not seem an adequate model (Figure 2).



```
import pyLDAvis.gensim

viz = pyLDAvis.gensim.prepare(model, corpus, dictionary)
pyLDAvis.display(viz)
```

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
50
100
150
200
250
300
dl
qty
pop
refills
fpm
bpg
stent
transplant
nitroglycerin
br
susceptible
nbsp
directed
infarction
catheterization
myocardial
coumadin
graft
glyburide
rca
stroke
pmm
paraplegia
stat
mrs
major
bypass
son
stents
reveals
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2 – First LDA model

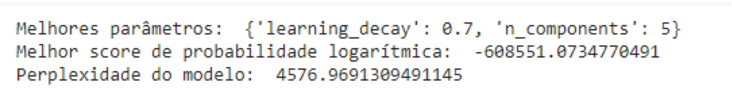
GridSearch was used to find the best model LDA. The most important adjustment parameter for the LDA is the number of topics (n_components), as we do not know how many topics we provide it with a list of different assumed values that we can define as n_components.

Another parameter that was changed was the learning rate (learning_decay), this rate also has no default value.

Before applying Grid Search, adjustments were made to the data vector (CountVectorizer) by changing the parameters token_patern to consider as tokens only in alphanumeric words greater than 5 characters.

Grid Search builds several LDA models for all possible combinations of reported parameter values. As he tests each combination, he takes a long time processing an output.

The result of the Grid Search points out the best parameters for the LDA model, which in our case was in 5 topics (Figure 3).



```
Melhores parâmetros:  {'learning_decay': 0.7, 'n_components': 5}
Melhor score de probabilidade logarítmica:  -608551.0734770491
Perplexidade do modelo:  4576.9691309491145
```

Figure 3 – Grid Search results

A comparison was also made between the performance scores of the models (Figure 4).

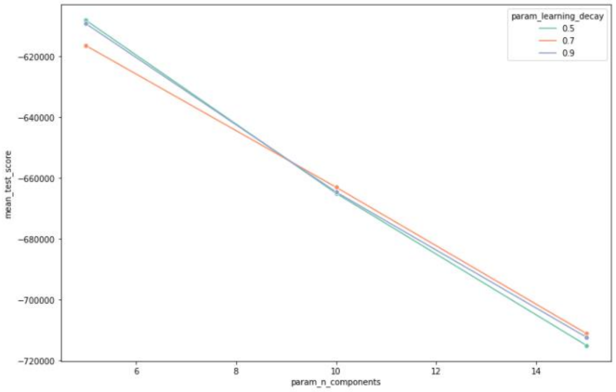
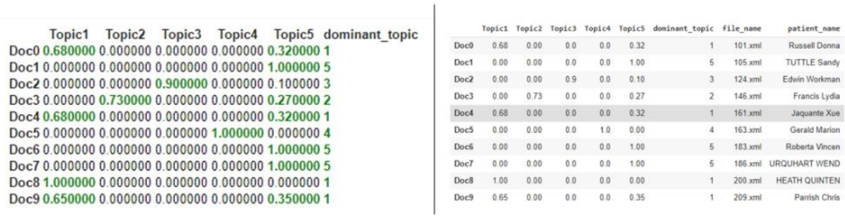


Figure 4 - Sensitivity analysis of the LDA model regarding the number of topics used.

One approach that the LDA shows us is that to classify a document as belonging to a specific topic, it is to see which topic has the greatest contribution to that document and assign it. We can see in the figure below, highlighted in green, all the main topics of a document and identify the dominant topic (dominant_topic) (Figure 5).

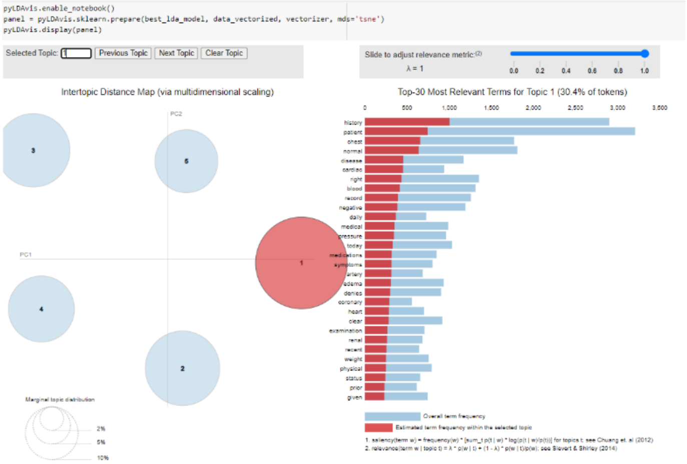


	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic
Doc0	0.680000	0.000000	0.000000	0.000000	0.320000	1
Doc1	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc2	0.000000	0.000000	0.900000	0.000000	0.100000	3
Doc3	0.000000	0.730000	0.000000	0.000000	0.270000	2
Doc4	0.680000	0.000000	0.000000	0.000000	0.320000	1
Doc5	0.000000	0.000000	0.000000	1.000000	0.000000	4
Doc6	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc7	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc8	1.000000	0.000000	0.000000	0.000000	0.000000	1
Doc9	0.650000	0.000000	0.000000	0.000000	0.350000	1

	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	file_name	patient_name
Doc0	0.68	0.00	0.0	0.0	0.32	1	101.xml	Russell Donna
Doc1	0.00	0.00	0.0	0.0	1.00	5	105.xml	TUTTLE Sandy
Doc2	0.00	0.00	0.9	0.0	0.10	3	124.xml	Edwin Workman
Doc3	0.00	0.73	0.0	0.0	0.27	2	146.xml	Francis Lydia
Doc4	0.68	0.00	0.0	0.0	0.32	1	161.xml	Jaquante Xue
Doc5	0.00	0.00	0.0	1.0	0.00	4	163.xml	Gerald Marion
Doc6	0.00	0.00	0.0	0.0	1.00	5	183.xml	Roberta Vincen
Doc7	0.00	0.00	0.0	0.0	1.00	5	186.xml	URQUHART WEND
Doc8	1.00	0.00	0.0	0.0	0.00	1	200.xml	HEATH QUINTEN
Doc9	0.65	0.00	0.0	0.0	0.35	1	209.xml	Parrish Chris

Figure 5 - Analysis of dominant topics (style x dataframe object), correlated documents to patient records

Another analysis that can be done is the number of documents per topic.
With these adjustments we see how the adjusted model using pyLDA is (Figure 6).



```
pyLDAvis.enable_notebook()
panel = pyLDAvis.sklearn.prepare(best_lda_model, data_vectorized, vectorizer, mds='tsne')
pyLDAvis.display(panel)
```

Selected Topic: 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6 - Adjusted model highlighting topic 1

3.1.8. Doc2vec

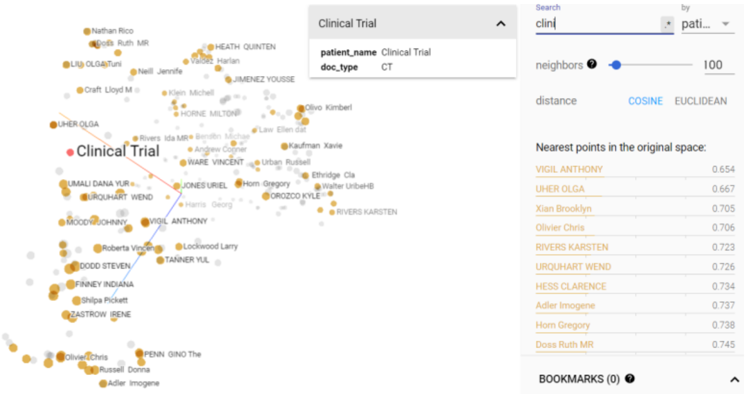
One of the most efficient techniques for representing a word is Word2Vec. Word2vec is a computationally efficient predictive model for learning word embeddings from raw text. He plots the words in a multidimensional vector space, where similar words tend to be close to each other.

The words around a word provide the context for that word. Doc2Vec is another widely used technique that creates the embedding of a document regardless of its length. While Word2Vec calculates a resource vector for each word in the corpus, Doc2Vec calculates a resource vector for each document in the corpus. The Doc2vec model is based on Word2Vec, with only the addition of another vector (paragraph ID) to the entry [9].

The purpose of Doc2vec is to create a numerical representation of a document, regardless of its length.

The technique we used to output the model was to rank the documents most similar to the clinical trial.

We concluded this technique by applying the Tensor Board Embeddings Projector, Figure 22, which is a great tool to analyze your data and see the values incorporated into each other. The panel allows searching for specific terms and highlights words that are close in the embedding space. T-SNE was the method used to be useful for exploring neighbors and finding clusters (Figure 7).



Clinical Trial

patient_name Clinical Trial
doc_type CT

Search clini by pati...

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

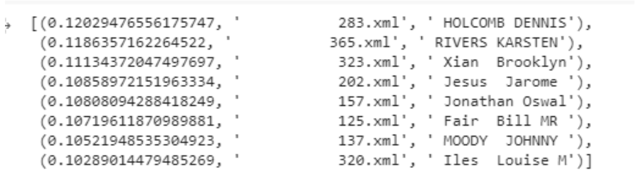
VIGIL ANTHONY	0.654
UHER OLGA	0.667
Xian Brooklyn	0.705
Olivier Chris	0.706
RIVERS KARSTEN	0.723
URQUHART WEND	0.726
HESS CLARENCE	0.734
Adler Imogene	0.737
Horn Gregory	0.738
Doss Ruth MR	0.745

Figure 7– Tensor Board based in T-SNE

4. Results

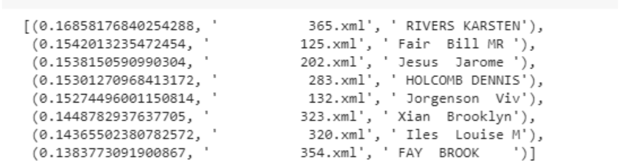
This chapter will show the results obtained by the various methodology or techniques applied on the NLP Research Data Set - N2C2 database.

We found that the application in TF-IDF, obtained a result similar to the application of Cosine Similarity, despite having different approaches, since the TF-IDF measures the term frequency instead of distance between vectors containing the word count, as is the case with Cosine Similarity (Figure 8).



```
[(0.12029476556175747, '          283.xml', ' HOLCOMB DENNIS'),
 (0.1186357162264522, '          365.xml', ' RIVERS KARSTEN'),
 (0.11134372047497697, '          323.xml', ' Xian  Brooklyn'),
 (0.10858972151963334, '          202.xml', ' Jesus  Jarome '),
 (0.10808094288418249, '          157.xml', ' Jonathan Oswal'),
 (0.10719611870989881, '          125.xml', ' Fair  Bill MR '),
 (0.10521948535304923, '          137.xml', ' MOODY  JOHNNY '),
 (0.10289014479485269, '          320.xml', ' Iles  Louise M')]
```

Figure 8 – Top 10 using TF-IDF model



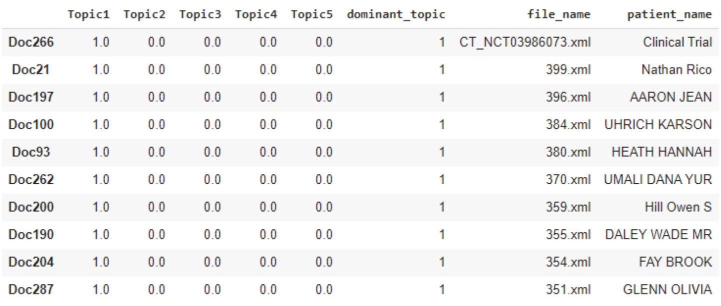
```
[(0.16858176840254288, '          365.xml', ' RIVERS KARSTEN'),
 (0.1542013235472454, '          125.xml', ' Fair  Bill MR '),
 (0.1538150590990304, '          202.xml', ' Jesus  Jarome '),
 (0.15301270968413172, '          283.xml', ' HOLCOMB DENNIS'),
 (0.15274496001150814, '          132.xml', ' Jorgenson  Viv'),
 (0.1448782937637705, '          323.xml', ' Xian  Brooklyn'),
 (0.14365502380782572, '          320.xml', ' Iles  Louise M'),
 (0.1383773091900867, '          354.xml', ' FAY  BROOK    ')]
```

Figure 9 – Top 10 using Cosine Similarity

The other technique used, the LDA shows us the topic approach.

We can see that by this model our clinical trial was classified as topic 1.

Verifying that the clinical trial was classified in topic 1, we searched the list of 10 documents in topic 1 that would be possible eligible records (Figure 10).

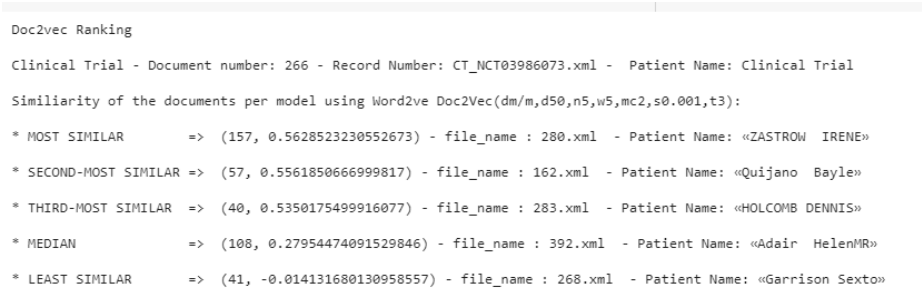


	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	file_name	patient_name
Doc266	1.0	0.0	0.0	0.0	0.0	1	CT_NCT03986073.xml	Clinical Trial
Doc21	1.0	0.0	0.0	0.0	0.0	1	399.xml	Nathan Rico
Doc197	1.0	0.0	0.0	0.0	0.0	1	396.xml	AARON JEAN
Doc100	1.0	0.0	0.0	0.0	0.0	1	384.xml	UHRICH KARSON
Doc93	1.0	0.0	0.0	0.0	0.0	1	380.xml	HEATH HANNAH
Doc262	1.0	0.0	0.0	0.0	0.0	1	370.xml	UMALI DANA YUR
Doc200	1.0	0.0	0.0	0.0	0.0	1	359.xml	Hill Owen S
Doc190	1.0	0.0	0.0	0.0	0.0	1	355.xml	DALEY WADE MR
Doc204	1.0	0.0	0.0	0.0	0.0	1	354.xml	FAY BROOK
Doc287	1.0	0.0	0.0	0.0	0.0	1	351.xml	GLENN OLIVIA

Figure 10 – Top 10 of topic 1 where we found our Clinical Trial

Doc2Vec was another technique used in this work to indicate records of patients similar to the clinical trial, we can see through the Ranking that some documents were chosen as more simulated.

The result was shown below with the 5 most similar documents using the Doc2vec model (Figure 11).



```
Doc2vec Ranking

Clinical Trial - Document number: 266 - Record Number: CT_NCT03986073.xml -  Patient Name: Clinical Trial

Similiarity of the documents per model using Word2ve Doc2Vec(dm/m,d50,n5,w5,mc2,s0.001,t3):

* MOST SIMILAR        =>  (157, 0.5628523230552673) - file_name : 280.xml  - Patient Name: «ZASTROW  IRENE»

* SECOND-MOST SIMILAR =>  (57, 0.5561850666999817) - file_name : 162.xml  - Patient Name: «Quijano  Bayle»

* THIRD-MOST SIMILAR  =>  (40, 0.5350175499916077) - file_name : 283.xml  - Patient Name: «HOLCOMB DENNIS»

* MEDIAN              =>  (108, 0.27954474091529846) - file_name : 392.xml  - Patient Name: «Adair  HelenMR»

* LEAST SIMILAR       =>  (41, -0.014131680130958557) - file_name : 268.xml  - Patient Name: «Garrison Sexto»
```

Figure 11 – Ranking result of Doc2Vec

5. Conclusion and future work

Although the effectiveness of these models depends on an evaluation by a recruiting specialist, Clinical Trials Recruiter, it is something outside the scope of this work that aims to show that it is possible to facilitate the recruitment process using Artificial Intelligence and the models available today. Each technician has a different approach and one of the models can better meet the recruitment process described in this work.

I believe that more research can be done in this area so that the process is more and more assertive and helps both professionals to find patients for their research, and mainly to help patients find a cure for their pathologies.

For future work, he intends to use a different approach in terms of the classification of clinical trials, that is, a supervised approach, in LSTM with NLP, to classify those that have similar eligibility criteria.

The idea would be a database of clinical trials of a specific pathology, extracted via web scrapping, with the data treated and classified by eligibility.

6. Bibliographical references

1. Matching patients to clinical trials using semantically enriched document representation
2. References [1] PM Spieth, AS Kubasch, AI Penzlin, BM Illigens, K. Bar-linn, T. Siepmann, Randomized controlled trials - a matter of design, Neuropsychiatric Dis. Treat. 12 (2016) 1341–1349
Hamed Hassanzadeha, Sarvnaz Karimib, Anthony Nguyena.
3. CA Umscheid, DJ Margolis, CE Grossman, Key concepts of clinical trials: a narrative review, Postgrad. Med. 123 (5) (2011) 194–204.
4. The majority of these Clinical Natural Language Processing (NLP) data sets were originally created at a former NIH-funded National Center for Biomed-ical Computing (NCBC) known as i2b2: Informatics for Integrating Biology and the Bedside. Based at Partners HealthCare System in Boston from 2004 to 2014.
5. Study of TQ-F3083 Capsules in Subjects With Type 2 Diabetes Mellitus (https://clinicaltrials.gov/ct2/show/NCT03986073?recrs=a&cond=Diabetes+Mellitus&draw= 2 & rank = 42)
6. https://medium.com/swlh/latent-dirichlet-allocation-lda-eff969bda284
7. https://www.machinelearningplus.com/nlp/gensim-tutorial/#2whatisadictionaryandcorpus
8. http: / /www.tfidf.com/
9. NLP: Word Embedding Techniques Demystified. Rabeh Ayari, PhD
10. https://www.ibm.com/products/clinical-trial-matching-oncology
11. NCT04228484 - The Insulin Response to the Gut Hormone GIP After Near-normalization of Plasma Glucose in Patients With Type 2 Diabetes (GA-16) - https://clinicaltrals.gov/ct2/show/NCT04228484?recrs=a&cond=Diabete+Type+2&draw=3&rank