Application of information extraction over clinical studies related to cancer diseases.

Giorgia Nidia Carranza Tejada (i6231006)

May 2020

1 Introduction - Carola and Giorgia

The project aims to apply text mining techniques over clinical studies in the cancer domain. The benefit of the operation is to draw attention to patterns in different domains.

In the first phase, it was identified the collection of data to apply mining techniques. The pre-processing is done to get a complete dictionary with the terms and their positions in the documents. One of the obtained results from the process highlights a group of relevant words over the documents proposed. Then, brief consideration is presented over the lexicon and the POS tag instrument used. To conclude, it was discussed the Named Entity Recognition to extract information about drugs, symptoms, treatment, etc over the clinical studies, analyzing the state of the art, and presenting first results. In the second phase, the characteristics and properties of the datasets were examined by different techniques like topic modelling, entity recognition and relation extraction. The non-negative matrix factorization topic model is applied to the document-term matrix for the extraction of six topics where each topic is represented by the most relevant words. This automatic determination of themes covered in texts of which you are not the author makes easier their investigation and analysis giving a real contribution to the people's work. Moreover, after filtering the categories of entities obtained by the named entity recognition, the phrases containing more than two terms of distinct groups were used to extract the possible relations. To conclude, the knowledge retrieved was represented through some visualization tools to comprehend the distribution of semantic types' terms over the documents and the creation of a graph from the triples.

Carola Roubin and Giorgia Nidia Carranza Tejada have worked on this project.

2 Dataset

2.1 Data sources - Carola

The project aims to identify the relationship between entities, drugs, and other factors, and cancer disease. The data required to develop the analysis consists of clinical medical records, including free text format to report diagnoses containing symptoms, intervention, drugs, and outcomes. This type of information was provided by the resource ClinicalTrials[1].

The Clinical Trials is a Web-based resource that provides information about medical studies. The Web site contains interventions, such as the medical product, behavior, or procedure on human volunteers and their outcomes. It is maintained by the National Library of Medicine (NLM) at the National Institutes of Health (NIH). The data from this website is split by study and, four documents, with around 5000 studies each one, are downloaded.

Instead, the definitions of terms regarding cancer were provided by the National Cancer Institute(NCI) Thesaurus[2]. The NCIt provides a vocabulary for clinical care, translational, and basic research. From the NCI website, the downloaded file Thesaurus.txt on 2020-04-07 includes all the names and synonyms associated with cancer using the following fields: code, concept name, parents, synonyms, definition, display name, concept status, semantic type.

2.2 Data Extraction - Giorgia

The NCIt provides access[3] to several collections of terminology. The set used is Thesaurus_20.03e, the latest version released on 2020-04-07 in the format of plain text of 157441 terms.

Besides, ClinicalTrials supplies an API to access all published knowledge on investigation records data. The result of the expression "cancer" to the full studies query returns 83706 with 320 different field information in XML format. The research takes into consideration 20000 records with the following entries:

[1] `https://clinicaltrials.gov/`
[2] `https://ncit.nci.nih.gov/ncitbrowser/`
[3] `https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/`

the official title The corresponding title of the protocol.

status The recruitment status of the study.

start date The date to open the recruitment or the day of the selection of the first participant.

completion date The date of the examination of the last participant.

condition list The name of the diseases or conditions of the clinical study.

study type the type of investigation realized. It can be observational, interventional, or expanded access.

intervention list the information about the intervention's type, name, and description.

outcomes module The description of the outcomes, including the title, a brief description, and the time frame of assessment.

condition mesh The corresponding condition or disease descriptor from NLM's Medical Subject Headings (MeSH) controlled vocabulary.

brief summary A description, including hypothesis, of the clinical study.

detailed summary An extended description of the protocol, which may include registry procedures and other quality factors.

A Python script was written to manage the HTTPs response and to extract the XML fields. The ClinicalTrials API handles the response of only 100 studies at a time, consequently, the code executes 200 requests to obtain 20000 records. The output of the API is an XML format containing all the 320 fields. So, the script extracts the information of the selected field and transforms the format to plain text to simplify the following procedures. In conclusion, the results are 4 files txt, each of them containing approximately 5000 studies.

3 Lexicon, Part-of-speech tagging & Negation handling

The lexicon is the dictionary of the language. The lexicon entry for each term includes orthographic, morphological, and syntactic information. For this project,

it was used the SPECIALIST Lexicon[3], since it comprehends in addition to the English lexicon, also many medical terms.

The part-of-speech tagging is the method of defining the part of speech of a word in a text. In this case, the documents analyzed cover clinical trials, and, frequently medical terms are used.

Consequently, a suitable tool to process the text is the MedPost/SKR POS Tagger[7]. The tagger was trained on the MEDLINE corpus and, it provides a tag set of 60 part-of-speech tags, which is derived from the Penn treebank tag set.

Both the instrument, previously presented, are included in the program MetaMap. The software will be later discussed in section 5, to explain the choice and employment.

An example of the results of the lexicon and POS tag of the following portion of a sentence: *with inoperable lung cancer for mutation analysis.*

world/term	lexicon	POS tag
with	with	prep
inoperable	inoperable	adj
lung cancer	lung cancer	noun
for	for	prep
mutation	mutation	noun
analysis	analysis	noun

To deal with the negation phrases, it was used the NegEx class supported by the software MetaMap. It indicates if a clinical condition is negated or possible and determines the terms which are negated. The outcome was provided in the machine output version, which collects the list of information about the type of negation, negation trigger, and negation concepts.

4 Information Extraction

Treating medical texts with related scientific terminologies is one of the major problems of the project relative to information extraction. This section will discuss the techniques of named entity recognition, the extraction of the relation, and the negation handling.

4.1 Named Entity Recognition

The named entity recognition from health text required knowledge of scientific terms as well as a different approach of the sentence, to be able to recognize the difference of diagnoses or treatment.

The first issue faced was finding a suitable tool for the analysis. The common libraries used to realize NER perform optimally in narrative text, such as spacy[4] and NLTK, also applied in section 3, recognize frequently entities: person, GPE, location. On the contrary, none of these entities covers medical information.

Another option adapts to the biomedical text was BioBERT. However, the alternative was discarded due to the limited size of the documents, and the absence of devices able to support its execution.

Spasić et al.[8] introduce text mining applications over cancer domains, and it is highlighted the frequent utilization of the application MetaMap over the biomedical domain. Reátegui et al.[4] realized a comparison of two tools for entity extraction in clinical notes, cTAKE, and MetaMap. The paper regards the extraction of obesity comorbidities, and it results in a good performance in terms of precision and recall of both the instrument, slightly better performance of cTAKE over MetaMap. Even though Rodríguez-González et al.[5] draws attention to cTAKE, which typically operates better on laboratory or test results, or locating rare symptoms, but raises in most cases the number of false positives.

Another library available for NER is scispaCy[5], which contains spaCy models for processing biomedical, scientific or clinical text. The number of categories of entities recognized is lower than MetaMap, and it reflects that it has no difference in Therapies or Symptoms.

The choice to use MetaMap is given not only by the frequent citation over previous papers but also by the facility of the employment through the batch function, which in addition to improving the general speed of execution, allows its use without minimum device requirements. The list of entities[6] recognized by the MetaMap refers to UMLS Semantic Network[7], and it includes types such as Sign or Symptom, Therapeutic or Preventive Procedure, Patient or Disabled Group, Body Location or Region, etc.

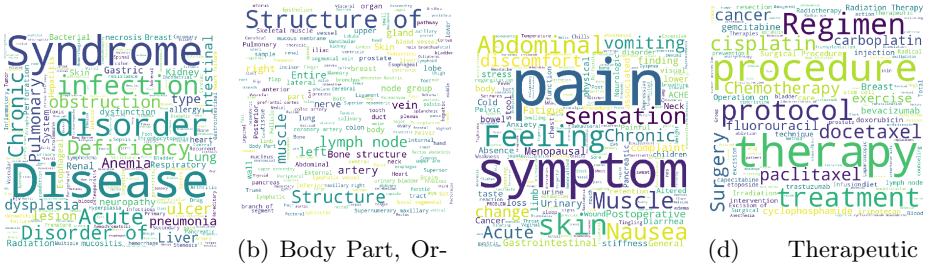
MetaMap performs the analysis over clinical study two times, to obtain two

[4] https://spacy.io/api/annotation#named-entities

[5] https://allenai.github.io/scispacy/

[6] https://metamap.nlm.nih.gov/Docs/SemanticTypes_2018AB.txt

[7] http://wayback.archive-it.org/org-350/20130703151851/http://semanticnetwork.nlm.nih.gov/Download/index.html



(a) Disease or Syndrome

(b) Body Part, Organ, or Organ Component

(c) Sign or Symptom

(d) Therapeutic or Preventive Procedure

Figure 1: Four semantic type of named entity recognition

different output formats and information. The first result returns a PROLOG format through the output option -q. It delivers information per sentence about the semantic types, processes the lexicon and pos tagging explained in section 4. The result of this operation is 4 files, with the same division explained in section 2.2, each document with a size of 2.1 GB.

The second output enriches the information of the semantic type and it is a format easily readable to a user to check the validity. As before, the resulting outputs are 4 files, each with a dimension of 850 MB.

```
 Meta Mapping (761):
593   SINGLE (Singular) [Quantitative Concept]
593   ARM (Upper arm) [Body Location or Region]
797   Radiation treatment [Therapeutic or Preventive Procedure]
```

The result of MetaMap proposed different candidate for a term, since it list different concepts that can be connected to the same term.[1] So, the outcome obtained was filtered into 4 semantic types of interest: Disease or Syndrome (dsyn), Body Part, Organ, or Organ Component (bpoc), Sign or Symptom (sosy), and Therapeutic or Preventive Procedure (topp). The figure shows the first impact of the named entity recognition through the use of word clouds. The methodology was not inserted in section seven since it can be not considered as a type of data visualization because it does not bring any additional information but just the quote of words.

4.2 Relation extraction

An important part of the information extraction consists of the extraction of the relations since it allows us to understand the interaction between different

entities such as drugs and disease.

The three different approaches to build relation extractors by hand-written patterns, by supervised machine learning or by semi-supervised and unsupervised machine learning.

Hand-built patterns can be customized to a specific domain, and since they are constructed by humans, they tend to have a high-precision. On the contrary, it involves a low-recall, and human participation limited the number of possible patterns.

Then, supervised machine learning can extract the relation between two entities and classify the relation category. The method involves the use of NER, dependency path, and classifiers such as Naïve Bayes, SVM, etc. To have a high level of accuracy in this approach, it will require a large amount of annotated data for the training and test phase.

A semi-supervised approach like distant learning, combine bootstrapping with supervised learning, by starting with several seeds, it will create a lot of features from all these examples to finally combine in a supervised classified. The method will require a large amount of data, but not labeled and will not involve human iteration to widen the number of patterns. Finally, unsupervised, such as OpenIE, will extract relations from the web without using any training data.

In this project, the relation extraction involves the same problem as the named entity recognition, that is the instrument commonly used is not suitable to be used in this domain. So, to select a proper system, several factors were taken into account. Firstly, it should deal with the absence of annotated data, since the output obtained from MetaMap is a different document, which includes not only the original phrases from the raw text, but also other information, previously discussed, in PROLOG format. Secondly, the limited time remaining from the previous operation and also taking into consideration the next step of visualization, exclude the creation of a tool from scratch as well as, laborious work of data annotation required in some of the methodologies. Finally, the system should be compatible with medical terms such as drugs, symptoms, disease, and so on.

Popular libraries as Spacy or NLTK tends to not support this functionality related to medical relations. As an example, another tool that supports the extraction, based on the knowledge graph is Pikes. Its application on some phrases extracted from the original document results to be unclear. It happens that the system erroneously recognizes entities or some other information about

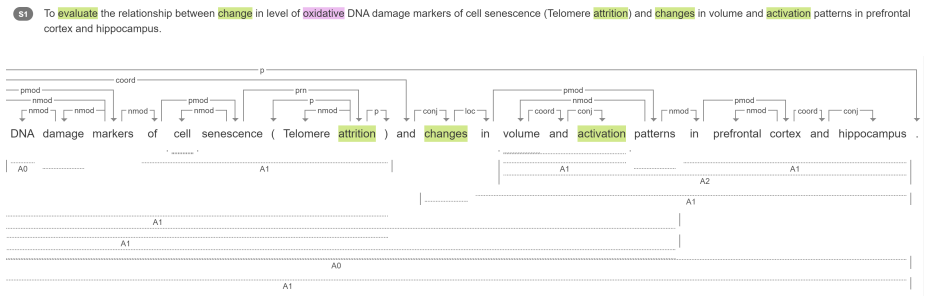


Figure 2: Phrase analysis evidence the followed entities.

different types such as diagnoses, therapy, or drugs are lost.

An example of its performance is shown the figures 2 and 3, which analyzed the phrase:

```
To evaluate the relationship between change in level of oxidative
DNA damage markers of cell senescence (Telomere attrition) and
changes in volume and activation patterns in prefrontal cortex and
hippocampus.
```

The phrase contains the followed entities:

DNA marker is a genetic marker.

Prefrontal Cortex and hippocampus are part of the body.

Cellular Senescence is the final stage of cellular differentiation, characterized by the inability to grow, divide, or repair damaged cell components, leading to cell death.

There is a relation between the "DNA marker" and the entities' "Prefrontal Cortex" and "Hippocampus" by the verb "Evaluate".

Pikes erroneously evaluates DNA Marker by taking into consideration in the graph only the word DNA, implying a different mean, and the analysis of the phrase does not evidence any of the entities previously defined.

Another approach was realized by analyzing the dependencies of the sentences using a similar algorithm proposed by [6]. In the previous section, we recognized several types of entities from the four documents of 20000 studies. Instead, in the current paragraph, we reduced the size of the text by analyzing only the phrases containing entities of the semantic types: dsyn, topp, and sosy.

Figure 3: The phrase analysis corresponds to the graphs.

To obtain the dependency of the phrases, it was used the library Spacy. Even if, it was previously mentioned that the library doesn't perform well to extract relations from the medical text eventually, it was still used since it performs dependency parsing of a sentence independently from the domain. The intention was to investigate the phrase, by easily explore the dependency tree to obtain the couple subject-object. Unfortunately, the approach didn't accomplish the objective since, the structure of the phrases was complex and it may cause the incorrect result of the algorithm.

Several papers cover the topic of relation extraction in the medical domain were analyzed. Most of the publications available proposed innovative approaches but, they did not release any source to their implementation, or the link to the tool suggested doesn't exist.

Liu et al[2] compared two systems to extract treatment relations from clinical notes. The result shows the outperformance of SemRep over MEDI.

So, the software used for the entity extraction was SemRep, a natural language processing application, targeted the extraction of semantic relationships in the biomedical text through a rule-based approach. The system relies on the domain knowledge of Unified Medical Language System (UMLS), and consist of 54 predicates between 134 entity types comprehensive the set previously filtered.

The output obtained from SemRep from the phrase " *There are no standard therapy options shown to prolong survival for patients with progressive disease on first-line docetaxel-based regimens for men with metastatic castration resistant prostate cancer (CRPC).*":

```
|relation|C0040808|Treatment Protocols|resa,topp|topp|||USES|C0246415|
docetaxel|orch,phsu|phsu||

|relation|C0246415|docetaxel|orch,phsu|phsu|||TREATS|C0025266|
```

```
Male population group|popg,humn|popg||

|relation|C0246415|docetaxel|orch,phsu|phsu|||TREATS(INFER)|C0936223|
Prostate cancer metastatic|neop|neop||

|relation|C0936223|Prostate cancer metastatic|neop|neop|||PROCESS_OF|C0025266|
Male population group|popg,humn|humn||

|relation|C1335499|Progressive Disease|fndg|fndg|||PROCESS_OF|C0030705|
Patients|podg,humn|humn||
```

Instead, the phrase "*Difference in Acupuncture Response Effect of 12 versus 24 acupuncture treatments examined for symptoms of chronic, chemotherapy-induced peripheral neuropathy in cancer patients and survivors.* ".

```
|relation|C0013216|Pharmacotherapy|topp|topp|||CAUSES|C0031117|
Peripheral Neuropathy|dsyn|dsyn||

|relation|C0031117|Peripheral Neuropathy|dsyn|dsyn|||PROCESS_OF|C0206194|
Survivors|podg,humn|podg||

|relation|C0031117|Peripheral Neuropathy|dsyn|dsyn|||PROCESS_OF|C1516213|
Cancer Patient|podg,humn|humn||

|relation|C0394664|Acupuncture procedure|topp|topp|||ISA|C0087111|
Therapeutic procedure|topp|topp||
```

The result obtained from SemRep does not limit the entity to the set previously defined, sosy, dsyn, and topp but it can extract all the 134 entities available. From the original phrases, the triples extracted were 20422 although, the number includes relations between semantic types not selected.

To filter the output, the triples considered must have the three types, previously mentioned, or as subject or objects. From this operation, the number of relations was reduced to 3284.

5 Precision and Accuracy

The precision and the recall evaluate the quality of the outcomes. A high level of precision denotes an accurate result with a low percentage of false-positive instead, a high rate of recall indicates the recognition of the majority of positives results. The formulas used to compute the precision and the recall are the following:

$$P = \frac{T_p}{T_p + F_p} \quad R = \frac{T_p}{T_p + F_n} \tag{1}$$

5.1 Named Entity Recognition

To compute the precision and accuracy of the methodology applied to recognize the entities, it required the use of annotated data to compare the MetaMap results. The project, as mentioned in section 2, is based on the clinical studies, and the text is not labeled. Therefore, the document should be manually marked but, the terms annotated by the software were quite specific, and required the knowledge to differentiate between treatments, symptoms, etc. The operation should involve a doctor or a figure specialized in the biomedical area. So, to evaluate the two values, it will be referred to previous studies that applied the same system in a similar domain.

Rodriguez-González et al[5] asses MetaMap over different diseases obtaining a precision of 93% and a recall of 67%. Besides, Reátegui et al[4] proposed the estimation of diabetes clinical studies. It will be taken into consideration only the first experiment, that did not involve any aggregation of terms. The overall precision was of 0.91 an the recall of 0.82.

6 Data visualization

The data visualization help to better understand the information extraction discussed in the previous section.

6.1 Named entity recognition

Named entity recognition identified entities from 127 types designate by the software MetaMap. From all the categories, only 4 of them were taken into consideration for the visualization. Disease or Syndrome(dsyn), Body Part, Organ, or Organ Component (bpoc), Sign or Symptom(sosy), and Therapeutic or Preventive Procedure (topp) were selected as the most representative of the

topic. Each class was collected into a list, including the term and its occurrence over the 20000 studies.

The first visualization, shown in the figures 5, 4, 7, 6, illustrates the distribution of the 15 most relevant terms through bars.

The second visualization, shown in the figures 8, 9, 11, and 10 consider the top 40 elements of each type into squares, and higher is the occurrence, larger is the dimension of the square

6.2 Relation extraction

Several approaches were tried to visualize the triples extracted. The first data visualization was the lowest accurate to expose the relation. It consists of nodes which, represent subjects, predicates, and objects, and similarly, edges to connect the triples. To display the result, it was used the library networkx. The figure 12 shows a confused outcome, which seems to be unclear and doesn't help to recognize information from an outside user.

The second visualization used the graph platform Neo4j. Neo4j is an open-source, NoSQL, native graph database, and it efficiently executes the property graph model down to the storage level.[8] The software was chosen not only for visualization purposes but also allows us to easily question the built system through Cypher query language. The first tentative, introduce the triples to the program through two files CSV, corresponding to the 1147 distinct entities and the 3248 properties. The problem, presented in the figure 13, with this approach to use only a document containing all the types, The employment of only one document to collect all the types deteriorate the visualization since all the nodes are visually similar, and the kind of relationship is not perceived. The second tentative, separate the distinct entities in 6 CSV file, corresponding to dsyn, topp, bpoc, sosy, Pharmacologic Substance (phsu) and the remaining types. This approach will help to have a better visualization since Neo4j confers different colors to distinct labels. The figures 14 and 15 shows an example of retrieval based on keywords by using a Cypher query and besides, Neo4j's interface helps to expand the visualization of other relations, starting by the result of the question, by double click the interested node. Even if, this methodology better represents the relations between different types of entities, however, Neo4j limited the visualization to only 300 items.

[8] `https://neo4j.com/developer/graph-database/`

A relative clear visulization of most of the quantities of entities is presented in the figure 16.

7 Conclusion - Carola and Giorgia

The project applied text mining techniques to medical texts concerning cancer disease. The intention was extracted information from several clinical studies as well as, categorized the topics covered in the documents, to give the interpretation of the datasets to understand their contents and potential patterns. Therefore, the entire process was divided into multiple parts correlated to the different methodologies used, and it was accomplished by two persons.

Firstly, text mining techniques are applied to reduce the number of words and useless symbols in the text getting a good performance. After, with more mathematical tools, the detection of possible patterns using topics is developed. Although, the smaller values of the topic quality, some of the created topics are enough coherent. Instead, the concern is for the topic rivers' results, in the future, time will be dedicated to them to check possible mistakes and to test other methodologies.

Besides, to extract information, some terms were categorized into types such as Sign or Symptom, Therapeutic or Preventive Procedure, Patient or Disabled-Group, Body Location or Region, etc, through Named entity recognition. Following, it was looked for connections between the entities retrieved. Most of the difficulties in both the approach concern the identification of suitable software or libraries able to process medical terms, and produce good results. So, MetaMap was applied for named entity recognition, and SemRep for relation extraction. Finally, visualization methodologies were applied to the result to improve comprehension.

Some future works may involve the realization of a system of question-answering since an approach of keyword search was proposed derived from the application of Neo4j to the information extracted.

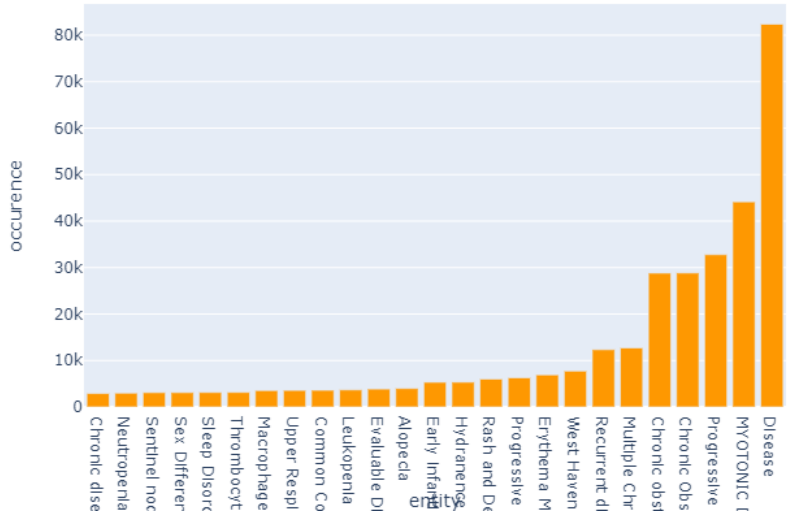


Figure 4: The distribution of disease or syndrome terms using bars.

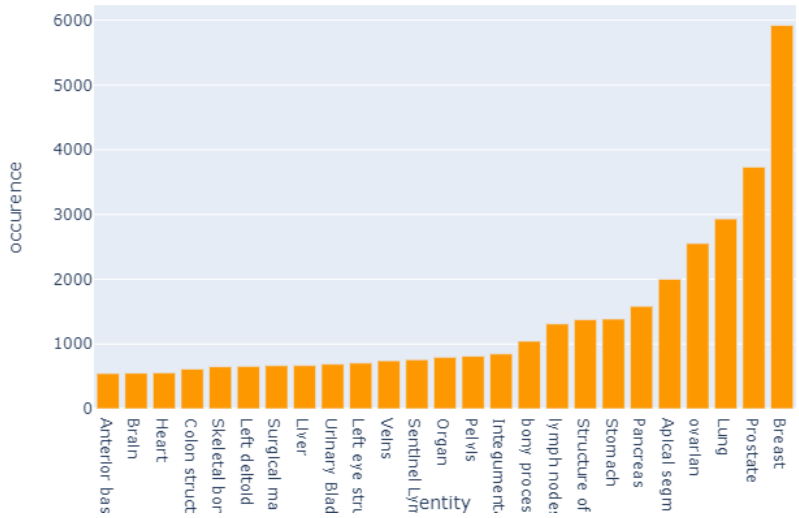


Figure 5: The distribution of body part, organ, or organ component terms using bars.

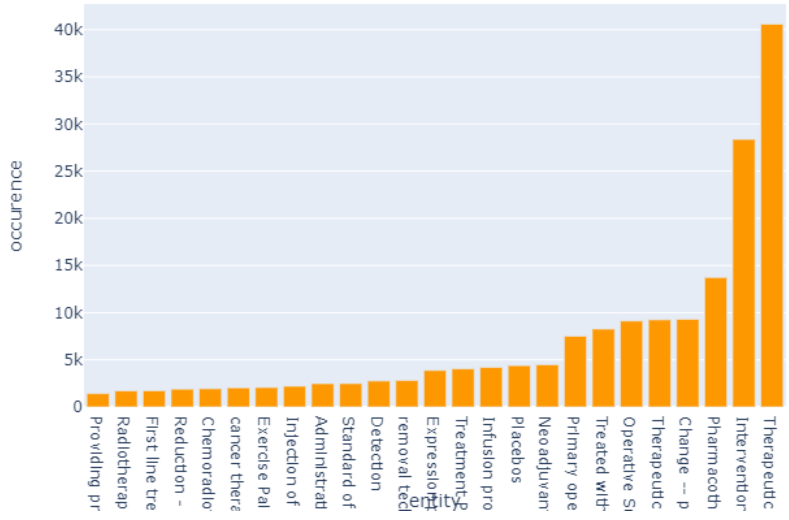


Figure 6: The distribution of therapeutic or preventive procedure terms using bars.

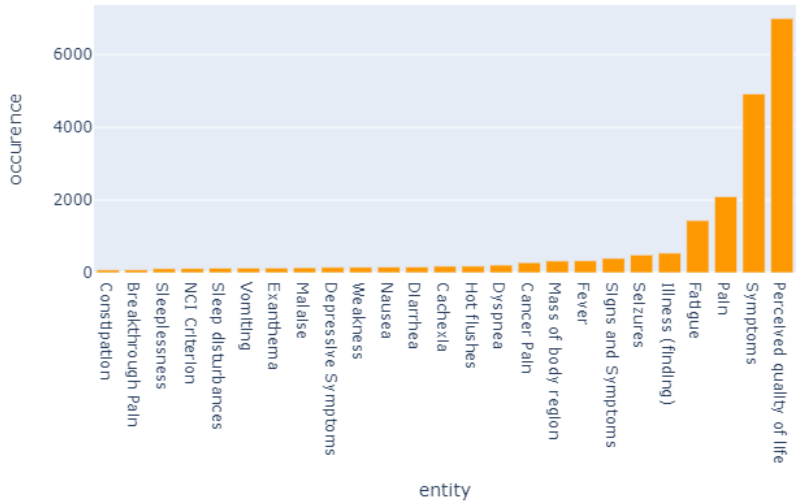


Figure 7: The distribution of sign or symptom terms using bars.

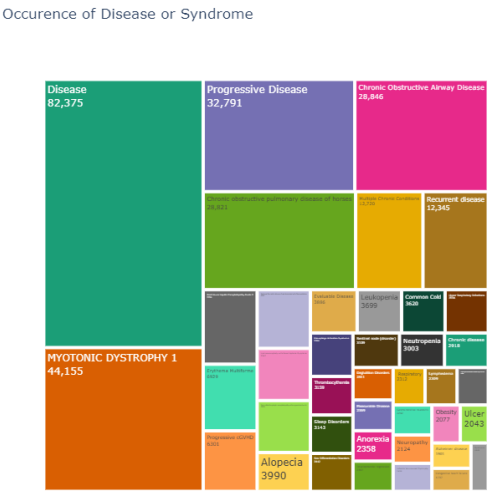


Figure 8: The distribution of disease or syndrome terms using squares representation.

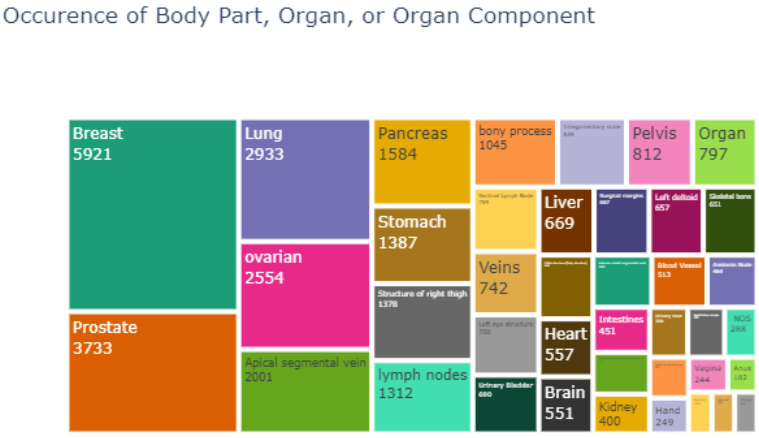


Figure 9: The distribution of body part, organ, or organ component terms using squares representation.

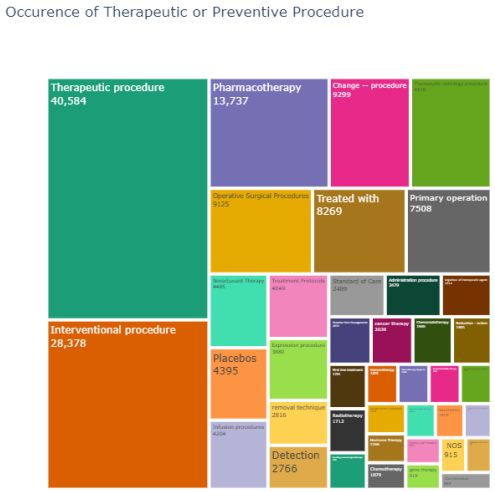


Figure 10: The distribution of therapeutic or preventive procedure terms using square representation.

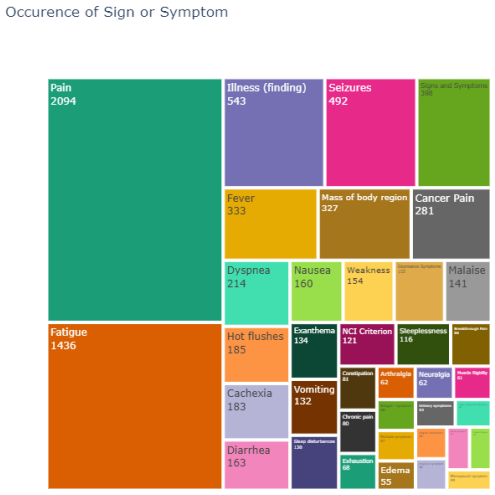


Figure 11: The distribution of sign or symptom terms using square representation.

Figure 12: Relations visualization through networkx.

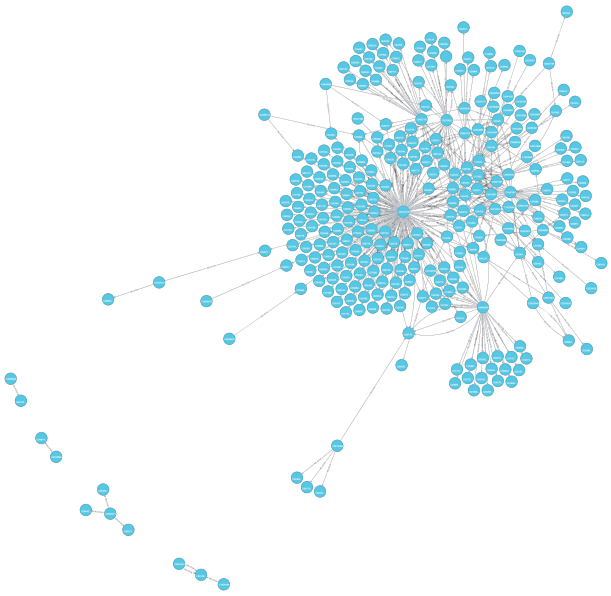


Figure 13: Visualization of 300 nodes using only 1 label for entities, and using Neo4j.

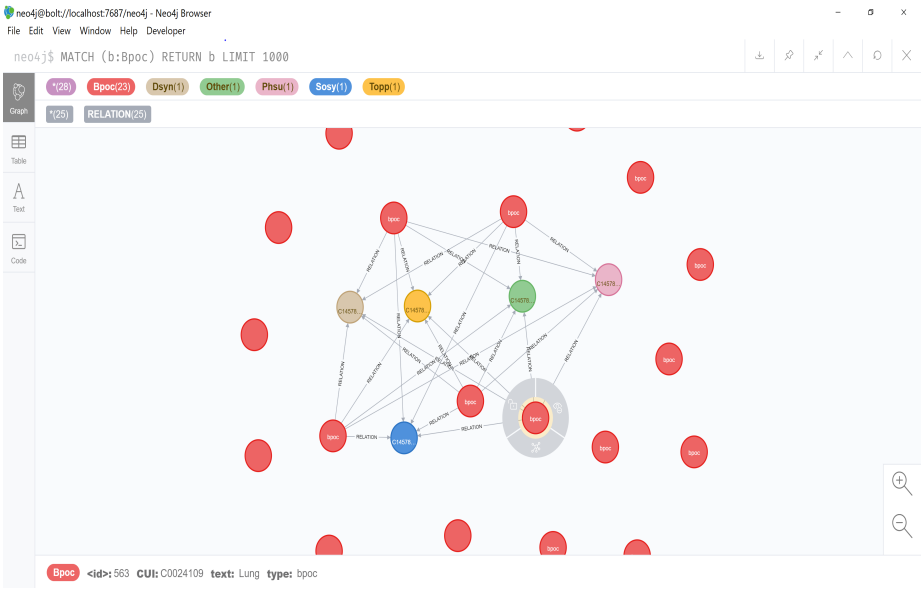


Figure 14: Visualization of 300 nodes using 6 labels, starting from the query of all the bpoc entities, and using Neo4j.

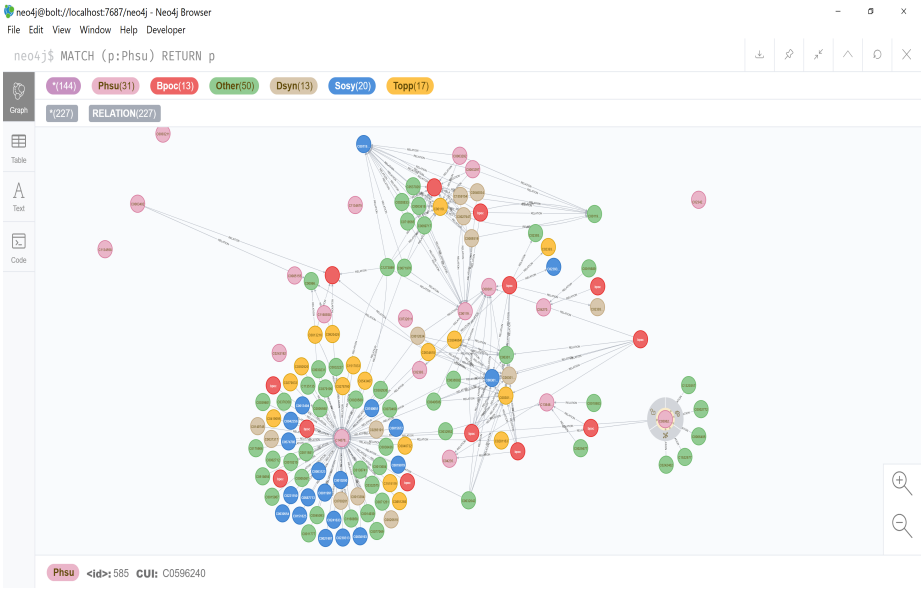


Figure 15: Visualization of 300 nodes using 6 labels, starting from the query of all the phsu entities, and using Neo4j.

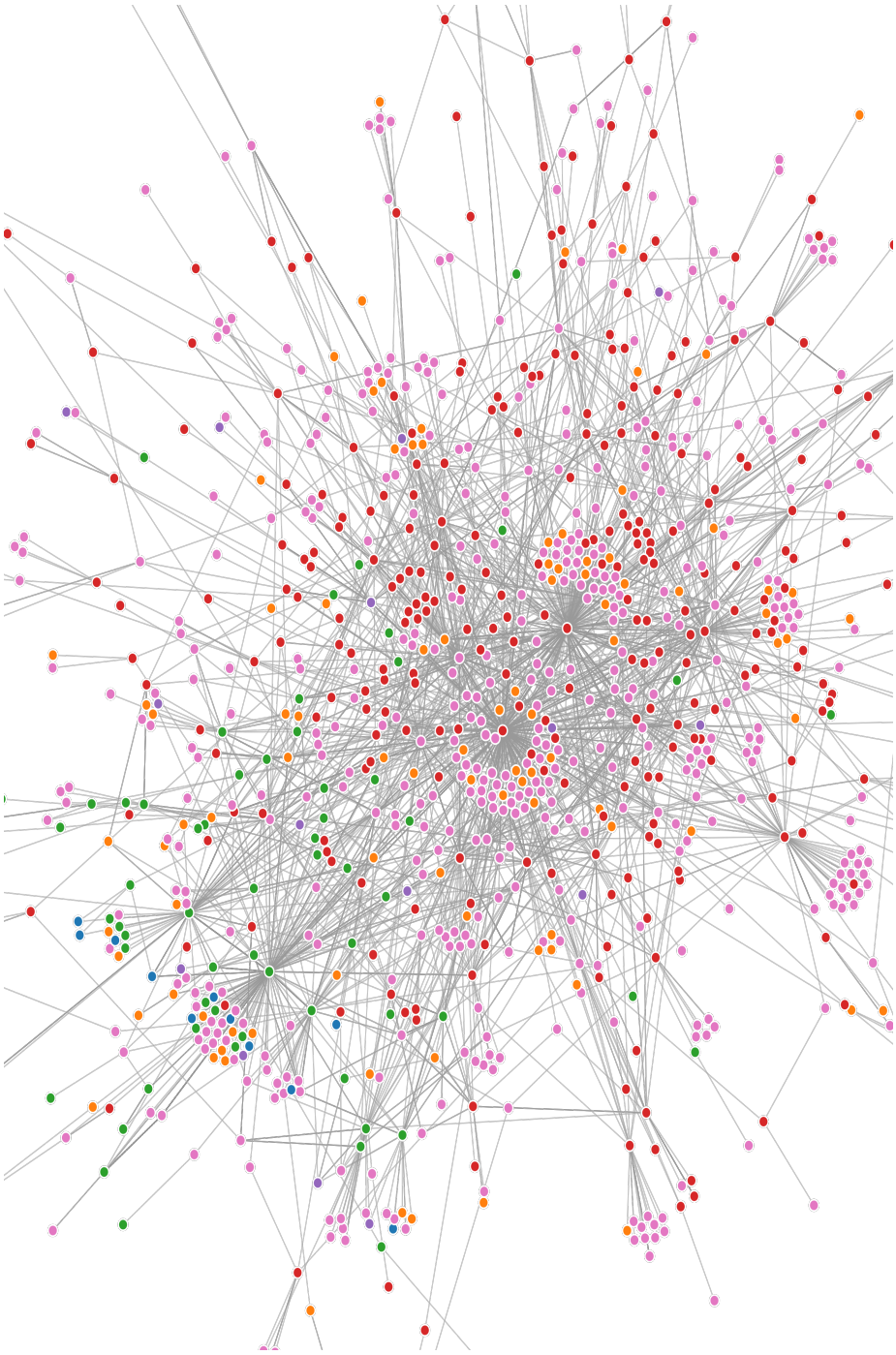


Figure 16: Visualization of the graph

A Appendix of Visualization images

References

[1] Asma Ben Abacha and Pierre Zweigenbaum. "Automatic Extraction of semantic relations between medical entities: Application to the treatment relation." In: *Semantic Mining in Biomedicine*. 2010.

[2] Ying Liu et al. "Using SemRep to label semantic relations extracted from clinical text". In: *AMIA annual symposium proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 587.

[3] Bethesda (MD): National Library of Medicine (US). "UMLS Reference Manual [Internet], SPECIALIST Lexicon and Lexical Tools". In: Available from: https://www.ncbi.nlm.nih.gov/books/NBK9680/. Sept. 2009. Chap. 6.

[4] Ruth Reátegui and Sylvie Ratté. "Comparison of MetaMap and cTAKES for entity extraction in clinical notes". In: *BMC medical informatics and decision making* 18.3 (2018), p. 74.

[5] Alejandro Rodriguez-González et al. "Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches". In: *Current Bioinformatics* 13.6 (2018), pp. 573–582.

[6] Delia Rusu et al. "Triplet extraction from sentences". In: *Proceedings of the 10th International Multiconference" Information Society-IS*. 2007, pp. 8–12.

[7] L. Smith, T. Rindflesch, and W. J. Wilbur. "MedPost: A Part-of-Speech Tagger for BioMedical Text". In: *Bioinformatics* 20.14 (Sept. 2004), pp. 2320–2321. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth227. URL: https://doi.org/10.1093/bioinformatics/bth227.

[8] Irena Spasić et al. "Text mining of cancer-related information: review of current status and future directions". In: *International journal of medical informatics* 83.9 (2014), pp. 605–623.