Alma Mater Studiorum · Università di Bologna

Physics and Astronomy Department
PhD Thesis in Applied Physics

Implementation and optimization of algorithms in Biological Big Data Analytics

Supervisor:

Prof. Daniel Remondini

Correlator:

Prof. Gastone Castellani
Prof. Armando Bazzani

Presented by:

Nico Curti

Session 2019/2020

Chapter 1

Biological Big Data - CHIMeRA project

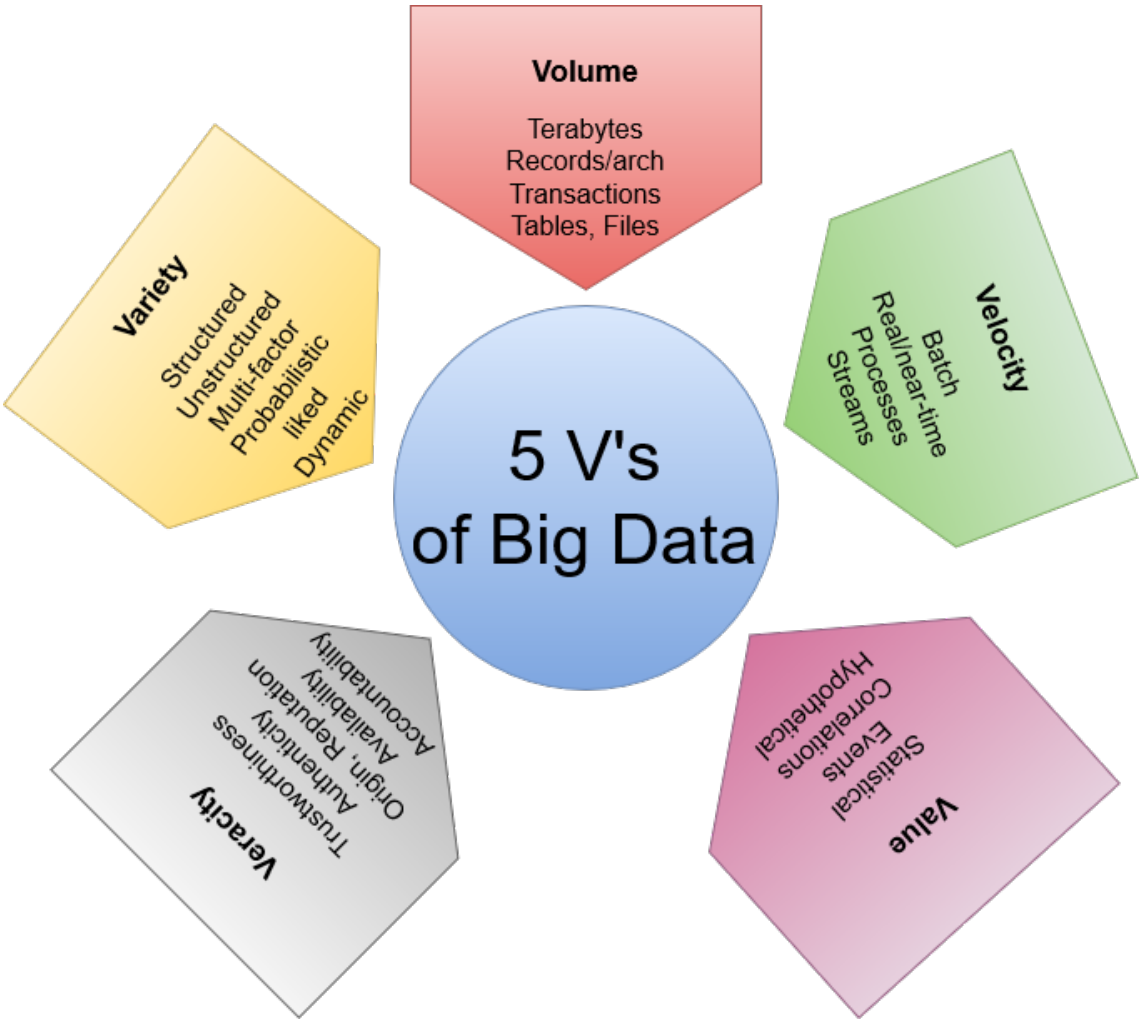

Figure 1.1: Big Data 5 V's

Every second a large quantity of data are produced and shared through Internet and Web-pages. Data are collected by social networks, messages, video streaming and images. Everyone, in fact, can easily create new data sources and share or put them in Internet

pages. The growth of this data is not limited to multimedia data but it involves many different fields. This is one of the most important feature of the contemporary time, the so-called Big Data era: this huge volume of data has created a new field in data processing which is called Big Data Analytics that nowadays positioned among top ten strategic technologies (Gartner Research, 2012).

It is still difficult to provide a definition of what exactly are the Big Data and we can find many slight different nomenclatures and categories which aim to formulate it. Moreover, Big Data does not define a particular data type but more than we normally think sources can be labeled as it. The *International Journal of Computer Applications* defined them as "$[\cdots]$ a collection of large and complex datasets that cannot be processed and analyzed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology". This definition certainly involves many aspects of Big Data processing but it does not provide any definition about their nature. Moreover, it is easy to identify them as "big" and thus difficult to analyze, but they are all around us every day and just using the Internet connection every smart-phone or laptop can extract and visualize our web queries so could be not properly correct to define them in this way. However it is certainly sure that the standard computing techniques have to be reviewed to face on this vast amount of data and a even more important attention has to be payed on the algorithm implementations.

While a global definition of them is evidently difficult we can however describe them using some of their "essential" features. One of the most common and used set of labels is given by the so called 5 V's of Big Data: volume, velocity, variety, veracity and value. Despite the first twos are quite obvious (the Big Data are certainly *big* in volume and they are produced very *fast*), the remaining three need a particular attention. Moreover, we have already treated problems about the volume of data (ref. Chapter**??**) and the need of very fast processing and algorithm optimizations (ref. Chapter**??**). Now in this chapter we want to focus on the remaining three characteristics of Big Data Analytics.

As pre-announced there many different sources able to provide data and this feature describes the extreme heterogeneity and variety of them. We can however broadly classify this variety into three global classes: *structured data*, *semi-structured data* and *unstructured*. A dataset is *structured* if we can easily manage the informations in it or, in other words, if it is described using the standard formats of data and thus can it can be "quearable". On the other side we have the completely *unstructured* dataset in which the data are disorganized and we need one or multiple pre-processing steps before use them or we have to develop different techniques to handle them. The intermediate step is given by the *semi-structured* data in which only a part of them could be handle with standard techniques or can be easily brought to their structured version. The data organization has been a crucial task on this work of thesis and we will return on this topic in the next sections.

The fourth essential characteristic of Big Data is the *veracity* of them due to data inconsistency and incompleteness. The data are shared very fast using Internet and there could be found some ambiguities and/or deceptions between different data sources. If we want to merge and aggregate different kind of informations we have also to face on this kind of problems. The final task of every Big Data Analytic application is, in fact, to process these large quantities of data and obtain a unique answer which can not vary in relation to the portion of data or dataset used.

The last and probably most important feature is certainly their *value*: it is certainly good to have access to large amount of data but unless we can turn them into value they are useless. In this vast amount of data only a small part of them can be considered as informative and it is always harder to extract this core of informations from them. Moreover, we have to take into account also the difficulties about the management of these data and their more or less complex structure. However, also in this case, it is hard to

generalize this property on all the amount of data contained in Internet: every day we see a large quantity of useless information on the web and it is hard to figure out that some of them can be useful for research applications. A key role is played by the *questions* which we ask: for every data source there is always an appropriated question which can be answer using it and vice versa. In this way also the seemingly useless datasets can acquire importance for an appropriated research project.

In this chapter we will discuss about one of the latest project developed during my PhD and which is still in work in progress: the CHIMera (*Complex Human Interactions in MEdical Records and Atlases*) project. The project is founded on Big Data applications and it is accidentally born as separated branch of the INFN FiloBlu project (ref. next sections and Appendix E for further informations about the project) which financed my last PhD year. CHIMeRA aims to create a unified database of bio-medical records using Natural Language processing techniques. The final purpose is to merge multiple data sources available on-line into a single network structure which highlights the relevant interactions between bio-medical informations, i.e starting from diseases to the biological agents involved into their causes and consequences. The realization of the first version of CHIMeRA required a lot of time and the development of novel pipelines of data processing. The project does not still achieve its conclusions but in this chapter we will cross through the main key points which allowed its construction.

1.1 The CHIMeRA project

The increasing availability of large-scale biomedical literature under the form of public on-line databases, has opened the door to a whole new understanding of multi-level associations between genomics, protein interactions and metabolic pathways for human diseases via network approaches. Many structures and resources aiming at such type of analyses have been built, with the purpose of disentangling the complex relationships between various aspects of the human system relating to diseases [8, 3, 4]. All these data come from different kind of studies performed by independent research groups who want to prove their theory about a particular aspect of biological agent interactions. Modern biological analyses perform very capillary studies on biological agents: in this way we can easily study the deeper relationships between them but we completely ignoring what are around them. This approach is certainly extremely efficient for the detection of the minimal causal agents of the problem but it tends to loose the global and complex[1] environment and prospective in which the process occurs. This is also the main problem of complex systems, i.e a system composed of multiple components with a mutual interactions between them. The study of an individual aspect, in fact, could give us only a partial overview of the system but we have to take in account the interactions between the multiple components for a global description.

The network structures are acquiring even more importance on this kind of studies. Complex System and System Biology researches have proposed multiple models about the dynamical and evolutionary interactions of the human system agents aiming to study the hidden relationships between them using network models. A network model, in fact, is able to highlight and quantify the non-trivial correlations between the system components. The main problem of this kind of approach is certainly the increasing dimensionality of the involved data: a network model could be described via its adjacency matrix, i.e a matrix $(N \times N)$ in which each row/column identifies an agent/node of the underlying problem and each entry (i, j) quantifies the importance of the interaction between the agent/node i and the agent/node j. In real data applications we can often reasonably assume that a wide amount of the matrix entries are null, i.e the interaction between the involved agent is

[1] From a physical point-of-view.

quite sparse, and thus we can used the important properties related to the sparse matrices to manage our network. However, when the amount of data increase also the management of a such sparse matrix could be difficult. More efficient solutions are provided by the modern Database formats and languages (e.g MySQL, SQLite, InfluxDB, $\cdots$) which store all these informations into a binary format and they allow to submit queries to extract the desired portion of data. A global visualization of these huge amount of data is, in fact, without practical-sense and none valuable informations can be extracted from the global representation of the system. The most important feature of network model is, in fact, the definition of a hierarchy of the interactions: the relationship between two nodes is given by the amount of connections which link them or, in other words, by their path. Starting from a node its nearest neighbors are given by the set of nodes connected to it: re-iterating this concept we can explore all the network structure[2]. In this way we can study the interactions of each node at different precision orders and causalities.

In light of these considerations we started to develop the CHIMeRA project (*Complex Human Interactions in MEdical Records and Atlases*) in which we aim to merge the state-of-art studies and databases about biomedical researches into a unified network structure. A key role on our network structure is played by the diseases: the major part of biomedical researches are focused on causes and consequences of a given disease and thus the corresponding databases involve the interactions between them and other biological factors. The diseases are also the most bigger manifestations of biological malfunctions and a large part of biological researches are financed on their study looking for their fine grain causes. Thus a disease could be a valid "bridge" between multiple data sources: in each database we can find the associations between a disease and a series of multiple biological aspects related to it which can be merged together using disease nodes.

The crucial point of this project was, in fact, the merging of different kind of informations provided by multiple distinct data structured. As told above, the major part of researches have focused on a partial aspect of the problem and they provide an independent result from the others, reducing the possibility of interactions between the outputs. Moreover, a lot of time is always spent for the creation of a practical visualization of the results using web pages and on-line services which drastically affect the real usage of these informations when we want to combine multiple sources. The CHIMeRA project started from these independent sources and it aims to maximize their overlap and thus the communications between them.

A final attention has to be payed about the format of these data: in physics we are friendly with numerical data but in these context we have to work with words. The above told databases include only the research outputs and thus the performed interpretations of numerical data studied. For example, if a numerically significant correlation was found between a disease and a gene we would found an association between them into a database and thus we can model it as a link. The only information available into this database is a link between two words, the disease name and the gene name, without any numeric value. While numbers have a unique representation (the number 2 is always 2) we can use multiple periphrases, i.e set of words, to identify the same concept. The biomedical community, in fact, has not yet provided a unified standard for disease identification or, at least, it has not yet provided a rigid standard as for other kind of data as genes or SNPs. So, if the diseases could be an efficient way to link together multiple data sources since all the studies prove correlations between them and other biomedical informations, they have the payback of an extreme variability in their nomenclature. The CHIMeRA project tried to overcome this issue using a Natural Language Processing (NLP) approach.

In the next sections we will discuss about the multiple steps which bring us to the formulation of our unified CHIMeRA database. We will start from the preliminary studies

[2] We assume that our network structure has not isolated nodes and undirected connections.

performed into the INFN FiloBlu project which allow the creation of the SymptomsNet structure, i.e a "small" network based on Italian words which connects diseases to related symptoms. Then we will briefly introduce the most common NLP techniques, also used into the CHIMeRA pipeline and finally we will show the main developed features of CHIMeRA.

1.2 How to find the data - Web Scraping

The INFN FiloBlu project was developed by the collaboration between the Physics Department of the University of Bologna and the INFN group of the Sapienza University of Rome. The project aims to implement a NLP pipeline to process messages with a medical theme. The critical issue of this project is about the message language: the project was financed by the Lazio region and it was developed to work at the Sant'Andrea Hospital of Rome so all the project involves Italian words. This constrain drastically affect the data availability which are very hard to find on-line. We tried to overcome this issue with a synthetic generator of phrases. To this purpose, a large set of medical words have to be provided and a valid interactions between symptoms and diseases can be useful to create reasonably phrases. The development of the NLP pipeline goes out from the scopes of this thesis (see Appendix E for further informations about our contributions about this task) but our first contribution to the INFN FiloBlu project concerned the realization of an Italian disease ontology. If we have a sufficient disease ontology we can use it to train our synthetic generator with reasonable phrases.

The English is becoming the predominant language in the research community and it is really hard to find (enough) data in other languages: everyone who wants to share his data and informations via Internet have to provide them in English if he wants to increase its visibility and availability. The Italian constrain posed by the project drastically limits the data sources and no public database was found. We would stress that as database we consider a public available set of structured data which can be downloaded and easily used.

Surfing on Internet many web pages can be found about diseases and their interactions with symptoms and causes, the so-called *on-line doctor*[3] (or Medical Services) pages. An on-line doctor is querable Internet service which allow to provide an auto-diagnosis of the user based on the information provided. The validity of the informations inside these tools is only partially guaranteed by the service provider and thus it can not be considered as a scientific method for medical diagnosis. However, the amount of informations collected by these applications is very interesting and it can be used to simulate reasonable medical queries, needed by our projects. Also in this case is important to notice that despite the availability of these public informations the data are structured according to the web page needs and moreover there is not an immediate download of the raw data.

So, how can we obtain these useful informations and re-organize them into a structured data format? The answer is given by the web-scraping techniques. With the term web-scraping we identify the wide set of algorithms aim to extract the information from a website, or more in general from the Internet. All the Internet pages are intrinsically pieces of codes written in different programming languages (HTML, PhP, ·). The major part of websites are written in HTML, an extreme verbose language, with more or less JavaScript supports. Write a code can be compared to an art form and the way chosen to reach the desired result is left to the programmer: in these way we do not have a rigid standard (excepted by the programming language constrains) and under each website underlies a potential completely different ensemble of code lines. Thus, the realization of a web-scraper

[3] Famous English applications are SteadyMD, MDLIVE, Sherpaa, LiveHealth Online and so on. Each service provides slight different informations and the choice of the best one vary according to the user needs.

poses several issues to the programmer who has to find the underlying patterns inside the web page to get the information stored.

A web-scraping algorithm is made by a series of multiple step which has to be performed without automatically. First of all, the algorithm has to recognized the unique website structure which can be summarized as the parsing of the underlying HTML code. Inside the large amount of code lines[4] are stored the set of informations useful for our application. So, the algorithm should be able to detect the relevant and interesting part and filter them. At this step we can easily reorganize the informations into a usable data format and save them.

There are multiple way in which all these steps could be performed and multiple open source libraries provide user friendly interfaces for the creation of own web-scraper. The most common one (and also used in our applications) is the BeautifulSoup [6] Python package. This package provides a very powerful Python library designed to navigate and read website source codes. The integration of this library with other pre- and post-processing techniques allow the extraction of the desired informations from websites and moreover their reorganization into a structured and usable data format.

1.3 SymptomsNet

The relation between symptoms and diseases is a crucial point for medical research and can be used to see analogies and co-occurrences of different pathology, including morbidity and co-morbidity. The construction of a unique and consistent database of these kind of data is an open problem for the research and a crucial task for many actual projects. The main problems arise from the complexity and heterogeneity of the available data and from the many nomenclatures used by different public databases. In fact, in many cases it is not so clear how to infer about the association between symptoms and diseases, and, in addition, different data sources provide different relations. These information are provided as sentences and periods, of variable length and we have to face on the problem of the different synonyms and periphrases used to describe the same concept.

In our work we used large-scale public on-line databases to construct a bipartite network of human symptoms-diseases. We used common tools of natural language processing to clean and uniform the data to maximize the overlap between the different sources. After its construction, this network is used to establish a score for the different words based on the centrality measure of the node.

This complex map of association can further be used to connect other data sources and enrich the diseases description from other medical points of view.

Many on-line databases offer auto-diagnosis tools and search engine in which insert a list of symptoms or diseases and obtain the corresponding "diagnoses". While many international databases are quite consistent and provided by medical/biological research groups, the available data in Italian language are quite scarce.

Using the Italian version of public *on-line doctor* websites we can obtained the needed informations. We applied a set custom web scraping pipelines[5] to the different web pages to extract the medical informations and we mainly focused on sites which highlight the relationship between symptoms and diseases. As discussed above the Italian data sources are quite fewer than the English ones so only three web pages were involved into our analysis: My PersonalTrainer, SaniHelp and Sapere.it. All these three sites provide an organized series of tables which associate a disease to the corresponding symptoms. These databases are certainly not robust from a scientific point-of-view and their vulnerability is

[4] Very large if we consider a pure HTML web page.

[5] We would stress that the extremely variety of the websites requires an equally varied set of web-scraping algorithms. Thus for each web page taken into account an appropriated web-scraper was developed.

shown also by a non-rigid labeling of the two classes: in multiple cases we can find a disease as symptom of an other one and in many cases there is not a perfect agreement between the three data sources. However, we can ensure that the most part of symptoms-disease association was validated by "authoritative" agencies in our databases selection.

The data extracted from the three websites cover a wide range of possible diseases and from each of them we obtained a network with a size of few thousand of nodes, our SymptomsNet. The overlap of the single words contained in the "disease-sentences" is quite low so a robust pre-processing was needed. The nodes were processed by standard natural language processing techniques and from each name the word stem was extracted to maximize the merging between the different sources. If two diseases show different symptoms we decided to concatenate the list of edges to not lose information.

The processing outputs generated a network with 2285 nodes and more than 29k links (only the 1% of the total number of possible links). The final SymptomsNet obtained by the merging of the three database sources is reported in Figure1.2 in which the size of nodes is proportional to the degree (Tab.1.1 for the top ranking links).

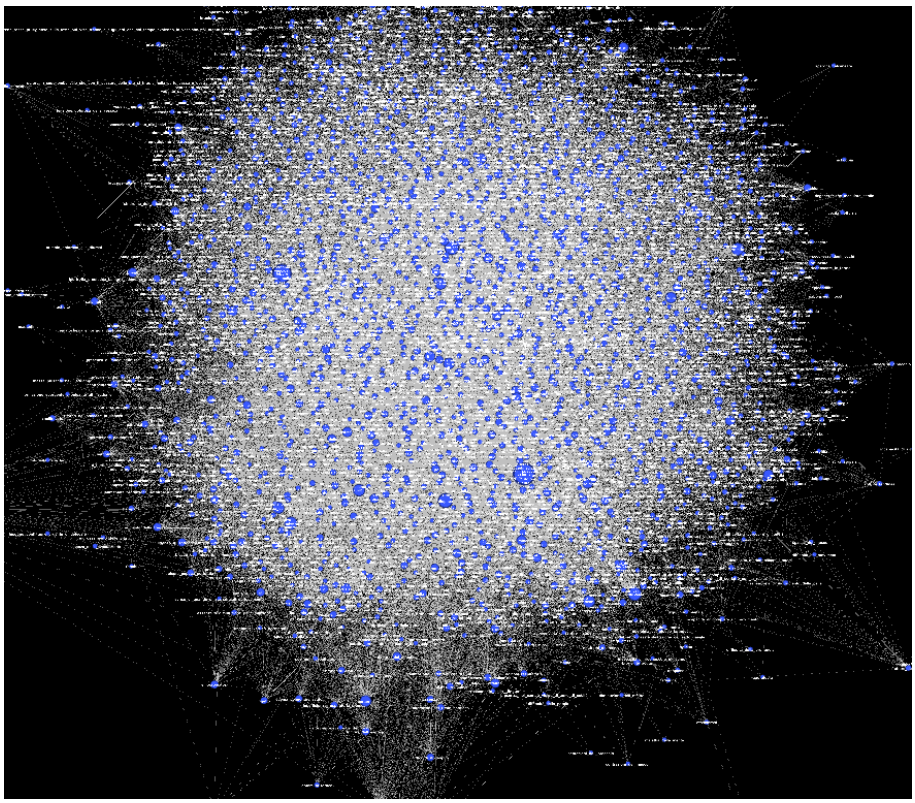

Figure 1.2: Symptoms-disease network generated by the merging of three public Italian web-pages of auto-diagnosis search engine. The network connects symptom and disease words according to validation agencies. The network comprise 2285 nodes and more than 29k links. In the figure the node size is proportional to its centrality (degree score). In this way the most common symptoms/diseases represent the biggest nodes.

In this simple example we can already notice as the most central (big) nodes are associated to most common symptoms-diseases, as expected. This result can be already interpreted as a validation of the performed processing. There are however some conceptual repetitions into the node list: if it could be a problem for the use of this network structure for theoretical analyses, it is a strength for our application to the FiloBlu project. In fact, this kind of occurrences allow us to consider a wide range of possible synonyms in the scorer attribution and so they can enforce the synthetic text generation.

Disease/Symptom	degree
Astenia	384
Febbre	313
Dispnea	225
Nausea	222
Anoressia	201
Ematemesi	193
Vomito	182
Debolezza	176
Affaticamento	176
Esaurimento	172
Mancanza Forze	168
Edema	158

Table 1.1: Top ranking link in the Symptos-Diseases network. Also in this case we can notice "periphrases/synonyms" associated to the same symptom as *Debolezza* and *Mancanza Forze*

.

We can conclude that from this very simple and preliminary work we are able to propose a novel symptoms-disease network based on Italian public databases and far as the author knows no other equivalent results are reported in literature. This works allowed also the realization of a novel database obtained by the union of the most common available data. The centrality measures extracted by this network can be used as floating point scorers or weight for the corresponding word in the diseases association and despite the still there issues it could be consider a valid input for a toy-model text generator.

This results encourage us in the usage of this kind of techniques for further and deeper applications and these ideas brought us to the development of the CHIMeRA project. If the above results could be reasonably good for the FiloBlu project purposes we could perform better using English data sources and with a better tuning of our natural language processing pipeline. In the next section we will discuss about what natural language processing means in the modern researches and we will describe the pipeline and databases using into the CHIMeRA project.

1.4 Natural Language Processing

Natural Language Processing (NLP) is a quite novel research field driven by the increasing availability of textual data (ref. Fig. 1.3). As told in the previous sections the incoming of Internet world exponentially increase the amount of data shared by people and the major part of them are textual data, i.e data composed by words, phrases and more in general texts. The NLP joins together techniques coming from the linguistic, computer science, information theory and artificial intelligence researches and it concerns the interactions between human languages and computers, or in other words it studies how a computer can analyze a huge amount of natural language data and how it could extract numerical informations from them. This is a very hard task to perform since it is not straightforward to teach to a machine how humans communicate between them so a key role is played by the artificial intelligence researches in the developing of new algorithmic techniques. The final purpose of the NLP is, in fact, to read, decipher, understand and make sense of the human languages extracting valuable and numerical results.

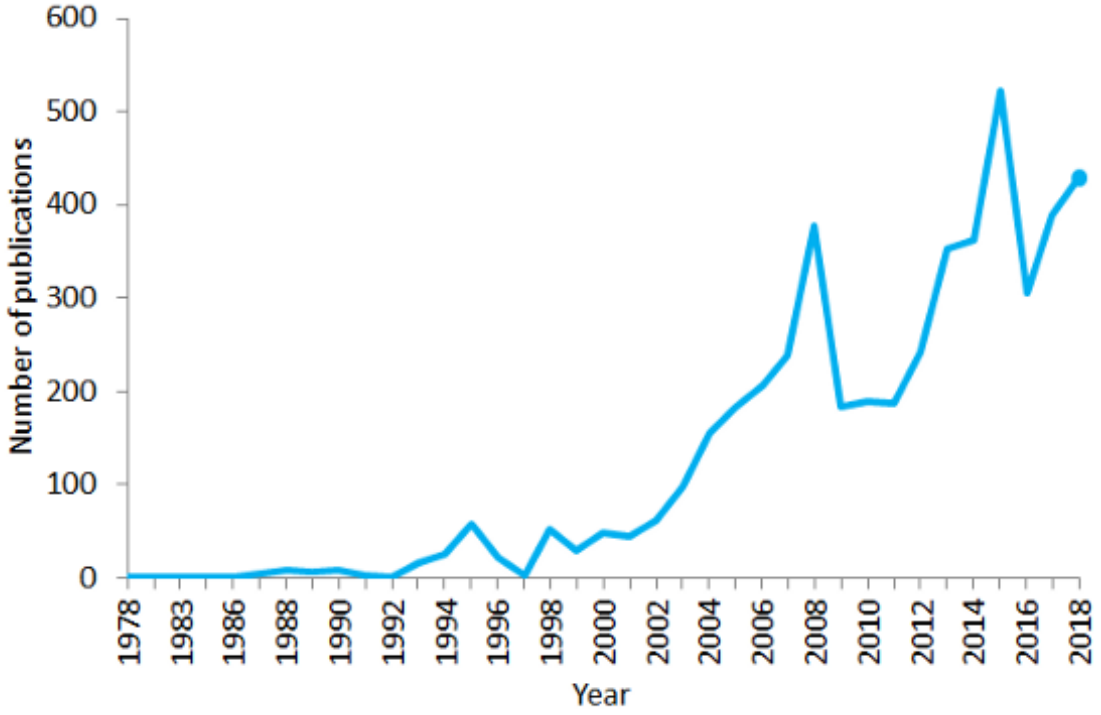

Figure 1.3: Number of publications containing the sentence "natural language processing" in PubMed in the period 1978–2018. As of 2018, PubMed comprised more than 29 million citations for biomedical literature.

Most of the modern NLP techniques are based on a Machine Learning approach to the problem and thus we can find statistical methods against deep learning neural networks trained to face on these kind of problems. A first step has to be performed to convert the human speech into a machine readable input; then the audio signal is converted into a string text and only at the end the text can be analyzed from the machine. Applying this work-flow in forward and reverse mode we can perform a communication between a human and a machine and vice versa. In this section we will ignore how the conversion from human voice to numerical inputs could be performed and its related problems and solutions but we will focusing on the last part of this pipeline, i.e in the description of the most common techniques to process a string text into numerical values. This is also the case related to our CHIMeRA project, in which we have a huge amount of names and strings related to medical terms and we want to standardize and increase their overlap.

First of all we have to take care that each human language has its own characteristics and thus it is harder to create to a pipeline ables to process all the languages at the same time while it is easier to tune an algorithm on a particular language. In our work we were focused on the Italian language (SymptomsNet) and on the English language (CHIMeRA Network). Since the SymptomsNet project was developed as simple proof of concepts, the developed Italian pipeline was really naive and for sake of brevity we will focus only on the CHIMeRA pipeline, i.e the English one. We would stress that in our application we were not interested on the understanding of the words meaning but we want to minimize the word heterogeneity maximizing their overlap. In this way we can ignore the semantic meaning of the strings and we could focus only on their syntaxes.

Syntax refers to the arrangement of words to make them grammatical sense. In this way we can create group of words applying grammatical rules: the grammatical rules have to be converted into algorithms which take in input a word and they give in output a

processed version of the same word. In this case there is not a numerical output but just a reorganization of the string letters and words. The most common techniques involved in the syntax analysis are:

- **Lemmatization:** it reduces the inflectional forms of a word into a single form.
- **Morphological segmentation:** it splits words into individual units called morphemes.
- **Word segmentation (tokenization):** it splits a large set of continuous text into units.
- **Part-of-speech tagging:** it identifies the grammatical part of speech for every word.
- **Parsing:** it provides the grammar analysis of the provided sentence.
- **Sentence breaking:** it divides a continuous text into sentences placing boundaries.
- **Stemming:** it cuts the inflected words to their root form.

Combinations of these algorithms can be find in everyday applications starting from email assistants or website chat box to the more advanced sentimental analyses and fake news identifiers. NLP pipelines could be used also in biomedical applications and the modern multinational factories like Amazon, IBM or Google are financing different kinds of research on this topic. Amazon Comprehend Medical is a NLP service developed by Amazon to extract disease conditions, medications and treatment outcomes from patient notes, electronic health records and other clinical trial reports. At the same time also companies like Yahoo and Google based their filters and email classifiers on NLP algorithms to stop spam. Also the fake news hot topic of the these years is faced on by NLP pipeline and the NLP Group at MIT is developing new tools to determine if a source is accurate or politically biased based on analyses of texts.

In our applications we constructed a custom pipeline based on part of these algorithms. In the following sections we will describe in detail our pipeline which was tuned for our case study: we would stress that the efficiency of our pipeline could not be generalized to other datasets since our purpose was to obtain the best result for our applications. In other words we can say that we had over-fitted our data. Moreover we have to clarify that our pipeline is not fully-automatic but it was made according to a semi-supervised approach: we customize the work-flow following the issues showed by our applications.

1.5 CHIMeRA datasets

We have seen how we can extract useful informations also from unstructured websites using a web-scraping pipeline. The *on-line doctor* web pages could be very useful for a toy model application like the SymptomsNet one[6] but if we want produce scientific relevant results we have to take care about the validity of the data. Since the English dataset availability is easier than the Italian one we moved to more "robust" databases.

As told in the previous sections, there are a lot of studies performed on the disease associations to other biological compounds and in many cases the resulting datasets are public available on Internet. This is the case of the DisGeNET [1] or DrugBank [5] datasets which contain the relations between a large number of diseases with genes/variants and drugs, respectively. The DisGenet web-page allows the download of the datasets already

[6] Also because no other databases were provided for the Italian language!

stored into a well structured network format (sparse adjacent matrix) while the DrugBank poses more issues to the treatment of data: the DrugBank database was designed to provide a large set of informations related to each drug inserted using its own website and thus it needs a pre-processing of the JSON dataset structured to highlight all the possible network associations. Using the DisGenet we can connect the diseases to the related genes and variants. From the reviewed format of the DrugBank, instead, we can link each disease to the consequential drugs. Associated to each drug we have also a list of gene and SNP targets which can be merged to the informations provided by DisGenet. Moreover, we can also connect food treatment to each drug and take care to the interactions between drugs (their synergy or not). We would stress that despite the trivial overlaps between the same data sources (genes, diseases and SNPs up to now), just using the rearrangement of these pair of datasets into a network structure, we can already provide a possible extrapolation of the underlying informations using the paths between nodes. Starting from a disease inside the DisGenet, using this single database we can study the "causality" relation to the connected genes. Using a multiple databases approach we can map that disease to other kind of informations like drugs or foods: in this way we can also hypothesize a direct relation between that drug or food with the above told gene passing through the disease node (2-step connection). In other words, the network structure allows the inference of missing connections using node contraction and this can be achieved only by merging multiple data sources.

To enlarge our informations about a disease we looked for other on-line data sources. A very interesting database is given by HMDB [7] (*Human Metabolite Data Bank*) which comprises a vast amount of metabolite and metabolite-pathway with the associated drugs and disease. The interconnection with the previous discussed datasets is straightforward but in this case the data are not public available with but we had to apply a web-scraping algorithm to get its informations. An analogous procedure was applied to extract the data included into the RXList database. RXList is an on-line website very similar to the previous discussed auto-diagnosis tools in which we can find associations between diseases and drugs. In this case we have a further distinction between diseases: we have diseases related to drugs and diseases connected to other caused-diseases. This further association can be modeled using directional links[7].

All these informations can enrich our database and the description of a given disease but we have to face on the problem of data merging. As previously discussed we do not have a unique nomenclature for diseases and thus we can find analogous names (periphrases or synonyms) which identify the same concept (disease). A useful tool to overcome these issues could be given by a synonym dictionary: a powerful example is given by the CTD [2] (*Comparative Toxicogenomics Database*). Thus we mined using another web-scraping pipeline also this database in which we could find associations between diseases and synonyms added to a list of phenotypes associated to each one. Using the CTD jointly with the SNAP [9] (*Stanford Large Network Dataset Collection*)[8] database we could enlarge the number of synonyms associated to each disease name.

A full list of the informations collected by our web-scraping and rearrangement pipelines is shown in Fig. **??**

We remember that the crucial point of our merging procedure is given by the disease nodes since they are the node type shared between all the databases. The help given by the synonym dictionaries certainly increase the overlap between the mined datasets but we chose to maximize it using a pre-processing NLP pipeline. So we started our pipeline using a word *standardization*, i.e converting all the words into their lower cases and replacing

[7] For sake of clarity, we encountered the same condition also into the DrugBank dataset in which we had internal connections between drugs.

[8] Extracted using another web-scraping pipeline.

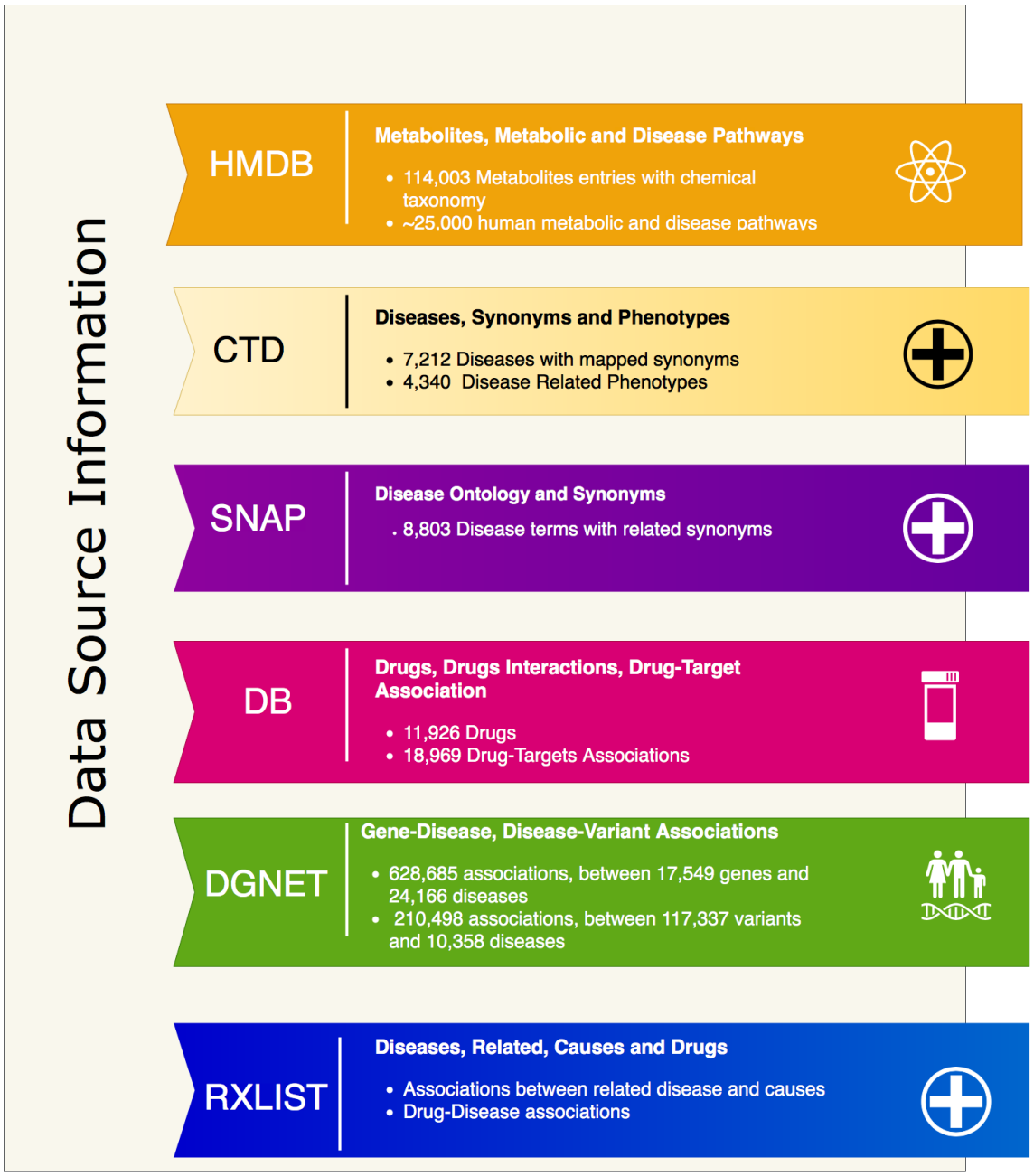

Figure 1.4: Description of the data mined by the CHIMeRA project before merging. The datasets were collected using custom web-scraping pipelines and by a rearrangement of the public data.

all the punctuation characters with a unique one[9]. Then we noticed that a not negligible part of words involved into the disease names was useless for the description: words like "syndrome", "disease", "disorder", "deficiency", $\cdots$ are not descriptive and so we can filter them. Now we can split the disease name into a series of token according to the list of words which compose it (*tokenization*). Each list of words is then sorted. To further increase the overlap we cut the inflected words to their root form using a *stemming* algorithm: the stemmer strength has to be tune according to the desired result so a first processing was performed using a Lancaster stemmer (more aggressive) but if the resulting output was too short to be compared with other names we changed it with a Porter Snowball stemmer (less aggressive). The stemmer algorithm choice is a very crucial task for NLP because using it we drastically loose word informations. Other processing steps were performed for critical cases encountered during the analyses: these steps constrain the developed pipeline and they were lost in generalization to other datasets. The work-flow outputs include multiple false-positive matches (ref. Fig. 1.5) and to remove them we can compare them to the original occurrences to evaluate a score match. This can be achieved introducing the standard word metrics: a common distance between words can be evaluated using the *Levenshtein Distance* which follows the equation

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} d_{a,b}(i-1,j) + 1 \\ d_{a,b}(i,j-1) + 1 \\ d_{a,b}(i-1,j-1) + 1_{(a \neq b)} \end{cases} & \text{otherwise} \end{cases}$$

where a and b are two strings of length $|a|$ and $|b|$ respectively. The *indicator function* $1_{(a \neq b)}$ is equal to 0 when $a_i = b_j$ and 1 otherwise. In this way the Levenshtein distance between a and b can be computed evaluating the distance between the first i characters of a and the first j characters of b.

A scheme of our custom pipeline is shown in Fig. 1.5.

[9] An unexpected issue arise in this step: different databases use different enumeration system. In some entries we found disease names associated to numbers which identify their multiple types. An example could be "Polyendocrine Autoimmune Syndrome type 1" but at the same time in a second database the same disease could be represented by "polyendocrine autoimmune TYPE I". Despite the global differences between the two names, given in this case by upper- and lower-cases of some letters and the deletion of some words, a very critical odds is the enumeration style. The performances of our pipeline dramatically increased using a roman_number_converter algorithm.

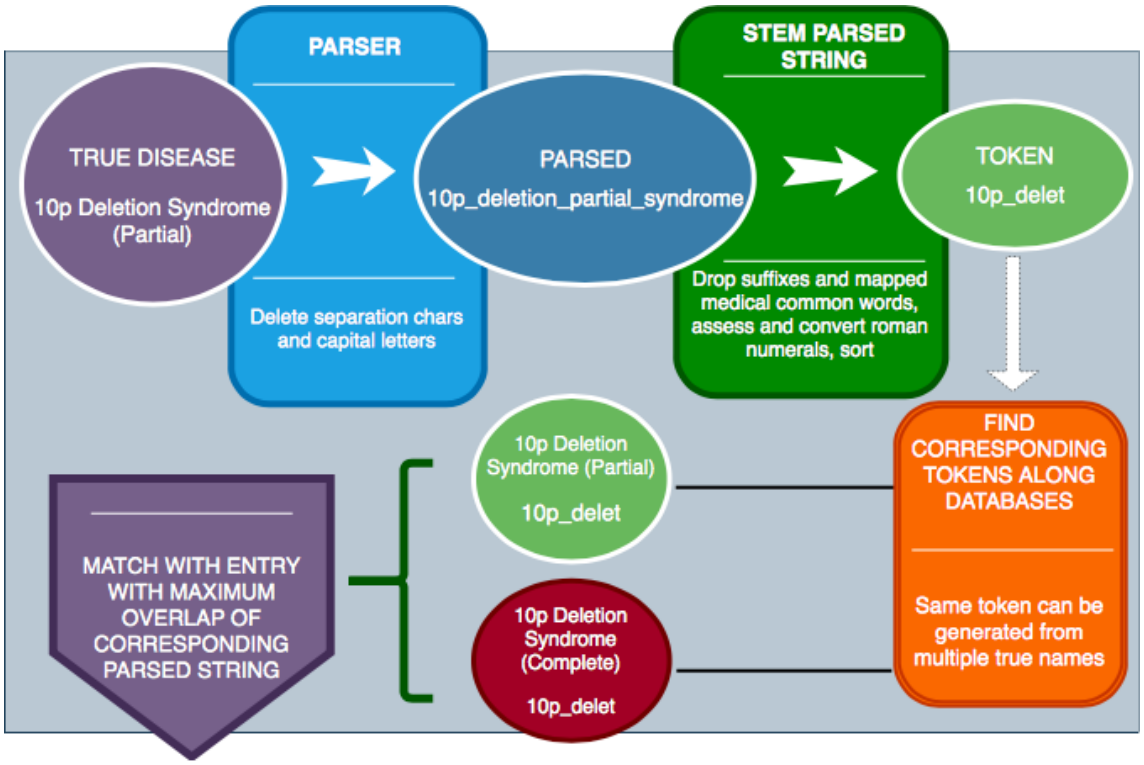

Figure 1.5: Scheme of the NLP pipeline developed in the CHIMeRA project. The disease words are processed in multiple step as showed in the example.

Bibliography

[1] G.-S. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, 10 2016.

[2] C. J. e. a. Grondin. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research*, 47(D1):D948–D954, 09 2018.

[3] C. A. e. a. Hidalgo. A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, 5(4):1–11, 04 2009.

[4] M. Lee, K. Lee, N. Yu, I. Jang, I. Choi, P. Kim, Y. E. Jang, B. Kim, S. Kim, B. Lee, J. Kang, and S. Lee. Chimerdb 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Research*, 45, Jan 2017.

[5] A. e. a. Maciejewski. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 11 2017.

[6] L. Richardson. Beautiful soup documentation. *April*, 2007.

[7] D. S. e. a. Wishart. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617, 11 2017.

[8] X. Zhou and et al. Human symptoms–disease network. *Nature Communications*, 5.

[9] M. Zitnik, R. Sosič, S. Maheshwari, and L. Jure. BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata, Aug 2018.