Disease understanding: Dealing with complex and unstructured big data in biomedical domain

Exercises proposals: 2024-25

This document contains a list of proposed exercises to be done by the students as part of this seminar. The students must conform groups of 3-4 persons (preferably 4) and select one of the proposed exercises. You can also propose another exercise if you want, but it should be approved first by the professor (contact Alejandro Rodríguez <alejandro.rg@upm.es> and Lucía Prieto <lucia.prieto.santamaria@upm.es>).

The proposed exercises are mainly based on the use of DISNET[1] platform. The platform is still under development, and we are aware that it can have some bugs and inconsistencies. Please let us know if you find any problem with the platform. You will help us to improve 😊.

The exercise consists in writing a **report** of the work done of a maximum of 3-4 pages. If there is additional material (e.g.: code, data, …), you can also include it in the file (zip, rar, …) to be sent, but **any result should be shown in the report to be delivered**. The deadline to do the exercise is March 26th, 2025. The exercises should be delivered using the available task in Moodle (in the case of those students with access to Seminars subject in Moodle). If you don't have access, please send the material to Lucía Prieto Santamaría (lucia.prieto.santamaria@upm.es) by email. In the subject you should write: "[DS Seminars] DISNET exercise".

The evaluation of the assignment depends on the specific task that you choose. We are expecting submissions where you provide some value associated to the effort of the subject (0,5 ECTS – about 10h of practical work).

Ad-hoc similarity measures (Theoretical-*Practical work*)

Imagine the following: we have an article about disease "D1" where is stated that this disease includes in its symptoms fever. On the other hand, an article about disease "D2" states that this disease includes in its symptoms hyperthermia or hyperpyrexia. The difference between the three terms from a "classification" perspective its clear (different codes, different names). However, from a semantic perspective, both terms are referred to a temperature that is about the normal range. In a similarity context (applying classical measures such as cosine, Jaccard or Dice), the three terms would be completely different, and no similarities will be found. However, we can certainly state that there is no so big difference. Can you describe a new model (from a mathematical perspective) to calculate similarities between two diseases based on their manifestations, considering this detail?

[1] http://disnet.ctb.upm.es

Can we find a better way or source to obtain the "total number of catalogued diseases"? (*Theoretical work*)

DISNET is currently using two sources to obtain the list of diseases to be processed. In Wikipedia, it is using a query against SPARQL endpoint of DBPedia and obtaining all the Wikipedia articles categorized as disease and filtering the results by keeping only those articles which contains "medical information" in specific sections. In PubMed, only the results that are within a specific set of categories are returned. Probably there are better ways of retrieving a list of Wikipedia articles that match with the required content. Can you provide a better one?

New sources of information and methods to retrieve it (*Practical work*)

DISNET is currently processing three sources of information: Mayo Clinic, PubMed and Wikipedia. However, there are plenty of sources out there that might be processed. Can you provide an example of a source that can be used to get this information and a code that allows us to extract from the source such content? (e.g.: MedlinePlus, CDC, …).

New mathematical measures in the biological layer (*Theoretical work*)

Could you propose a better mathematical measure than the explained ones to compute the similarities between diseases? Consider that two diseases could be very similar just by sharing a gene or a protein, although each of them are associated to other features. You can also consider in the formula some measures such as the disease-gene association score or the DSI (Disease Specificity Index) and DPI (Disease Pleiotropy Index) gene measures.

How sound is ChatGPT in recognizing entities (Practical work)

We show a small case of chatGPT to recognize symptoms on a text and returning their associated UMLS code, but we didn't check if those codes were correct. A possible assignment will be to do this exercise: send several texts from some diseases (for example, 5 text from diseases in Wikipedia) to chatGPT and evaluate if it has detected all the entities or not, and if the linking to UMLS is correct. Also, we aim to evaluate the determinism of the system. It would be interesting to evaluate its effectiveness by asking several times the same questions (on different moments to not influence with the previous context), and see how the results change.

How sound are LLMs in prioritizing drug repurposing opportunities (Practical work)

Choose a disease of interest and try to make drug repurposing predictions with ChatGPT or other LLM models. Ask the model to explain why it has arrived at each prediction to enhance its ability to reason over medical knowledge. Try with different prompts. Which one results in better predictions? You can use RepoDB information to test the predictions (https://unmtid-shinyapps.net/shiny/repodb/) and report the best model/prompt/approach. Could your methodology be automatized?

New statistical methods for determining sex-biased in ADRs (Theoretical-Practical work)

During the seminar, we commented that to determine the sex-biased direction of each of the ADRs (Adverse Drug Reactions) associated with cancer drugs we used a Fisher's Exact Test. This test was applied to check if there were differences between the number of occurrences an ADR was associated with each sex and to determine its direction. In addition, it was explained that the Odd Ratio associated with each of the ADRs by sex was calculated to know the probability/risk of suffering that ADR. Could you suggest an alternative method to obtain the same results?

Exploring registries of drugs from different regulatory agencies looking for sex-biased (Practical work)

In our project, we obtained data related to the adverse effects of each of the drugs from the FDA database. However, there are many other drug regulatory agencies such as the European Medicines Agency (EMA) or the Pharmaceuticals and Medical Devices Agency (PMDA). We propose that you search for the ADR registry in one of these regulatory agencies (or others that interest you), analyse the new dataset, check if there is information on the sex of the patients who have presented ADRs, and finally, propose a method to link all the information found with DISNET. For this, you will have to see which is the most appropriate vocabulary to perform this task (SNOMED, MeSH,…).