nLp-AttaCK

Team charter

Team Members:

Arun Narayanan	*Renewable Energy, Optimization*
Claire Chambers	*Bayesian Modelling, Predictive modelling, Data analysis*
Karsten Leonhardt	*Optimization, Regression & Classification, Data analysis*
Luis Vela	*Data wrangling, Regression & Classification, Data visualization*

Problem Statement:

The vast number of current research topics makes it difficult to identify a single study trend to guide future drug research and development (R&D). Hence, a system capable of extracting valuable insights from unsorted, unstructured text will be very valuable. To solve this problem, we will analyze trends in disease study using public databases.

Scope:

In the present project we will focus on the disease areas that are currently being studied, how they have changed over time, and what can be said about the future. We will however, restrict ourselves to the GEO Series studies on *Homo Sapiens* from the NCBI website (45,000 records). If time permits, we would like to expand this project to classify samples collected during experiments into healthy and unhealthy.

Team's Customers and Needs:

Astra-Zeneca want to gather information on disease topics in existing literature in order to direct and inform future research.

Success Metric:

The aim of this project is exploratory, and so assessing performance is a challenge. However, we can and will check components of our pipeline using standard methods such as comparison with ground-truth data (disease topic tagging) and we will assess how well our models generalize (cross-correlation of temporal regressor).

Specific Objectives and How Measured:

Specific Objectives:

- ➔ Disease tagging: Analysis of disease areas that are being studied in the GEO database using topic modeling from abstracts
- ➔ Evolution over time: Analysis of the evolution of extracted topics over time
- ➔ Predictive models: models of the future based on time series of topics

How measured/Performance indicators:

- ➔ Disease tagging: We will construct a Confusion matrix to measure the accuracy of our model and compare to ground truth data
- ➔ In the preliminary stages: We will count the number of unclassifiable entries and use it as a baseline to compare future results
- ➔ Predictive models: We will see if our model generalizes well with data that has never been seen before by the model

How you will work together:

- ➔ Attend scrum meetings
- ➔ Ask for help, if needed
- ➔ If deadlines are set, try meeting them, else ask for help/state the problem. Communicate in time, especially if your work is important for another person
- ➔ State any unavailability as soon as possible
- ➔ Major decisions regarding the project plans, directions, and goals should be decided by the team together and not by individuals alone

Plan:

Stakeholders:

- ➔ Astra-Zeneca

Initial team duration:

- ➔ 5 week duration of S2DS program

Activity backlog/tasks:

- ➔ Building the webscraper
- ➔ Researching NLP techniques

Meeting schedule:

- ➔ Daily scrums
- ➔ Updates on slack
- ➔ Extra meetings on zoom as needed

Milestones:

- ➔ **Week 1** get data
- ➔ **Week 2** extract topics using NLP, map to standard disease names
- ➔ **Weeks 3-5** generate descriptive statistics, build and test predictive models, visualize results

Communication expectations:

- ➔ Weekly meetings with company mentor
- ➔ Updates on slack
- ➔ Mid-term presentation to Astrazeneca and S2DS
- ➔ Final presentation to Astrazeneca and S2DS

As a … I will … when … so that … :

- ➔ As a team member I will give positive or constructive feedback when we discuss about each individuals progress so that the progress of the project always remains the top priority
- ➔ As a team member, I will listen carefully, and learn and improve when given constructive (or other) feedback so that the team's objectives and goals remain top priority
- ➔ As a team member, I will acknowledge that every member has a different skill set and be kind, nice, and supportive when interacting with team members, e.g., in team scrums, so that both the team member and the team gain value from the interactions