Part 2: Clinical NLP in Practice

Vlada Rozova

THE UNIVERSITY OF MELBOURNE

Centre for Digital Transformation of Health

CONNECTING DIGITAL INNOVATION TO HEALTH

Hands-on practice

Link to the dataset:

Link to the Jupyter Notebook:

Building a named-entity recognition (NER) tool for clinical text data

- Where to start and what to look out for?
- What are the common approaches and what to expect?
- How do they compare on real-world data?

THE UNIVERSITY OF MELBOURNE

Centre for Digital Transformation of Health

CONNECTING DIGITAL INNOVATION TO HEALTH

Let's first take a look at the data

Link to the dataset:

What is this dataset made up of?

- A .csv file with report metadata
- A folder with .txt files — reports themselves
- A folder with .ann files — annotations

Folder Navigation: <base>

Name	Size	Modified
annotations		
reports		
LICENSE.txt	2.5 KB	2023-10-20
SHA256SUMS.txt	48.6 KB	2023-11-02
annotation.conf	727 B	2023-01-26
chifir_metadata.csv	9.1 KB	2023-09-30

THE UNIVERSITY OF MELBOURNE

Centre for Digital Transformation of Health

CONNECTING DIGITAL INNOVATION TO HEALTH

Now time to head over to the Jupyter Notebook and explore

Link to the Jupyter Notebook:

Outline

- Some background on the CHIFIR dataset
- Exploratory data analysis:
  - How many reports? What do they look like?
- How to annotate data yourself?
  - Where to start? What to look out for?
- Three types of NERs:
  - How are they different? What works better on CHIFIR?
- Other tips & considerations