TC3 Information Retrieval

Master 1 AI, Upsay
T4, March to April

Kim Gerdes
Lisn

Last time

- Kaggle
- terms
- patents
- semantic relations
- term detection
  - how is it going?

Planning

1. 10/3 gentle introduction
2. 17/3 big dataset, binary evaluation
3. 24/3 improvements: embeddings
4. 28/3 project presentation, technical terms, semantic structure
5. **31/3 exploring existing tools: relations, Prodigy**
   NO 7/4!
6. 14/4 work on projects, discussions
7. 21/4 project presentation

Today

- relations
  - presenting dependency
  - presenting Spacy tools we need
  - build a baseline
  - spend time together on Notebook 5 (not to be handed in)
- you alone:
  - try to get Prodigy running for Notebook 4
  - prepare submission for next Friday

Project

for your patent domain:

1. term detection
    - rule-based baseline (dependency-parsing-based is not really rule-based)
    - train statistical model (based on spacy)
    - improve annotation using prodigy
    - evaluate
2. relation detection
    - rule-based baseline
    - improve annotation using prodigy
    - find tool(s) to extract relations
    - imagine new ways of improving the annotation quality and visualization
    - …

- discussion **April 14**
- presentation of the current state **April 21**
- It should be clear what you have tried and what remains to be done and what you won't do. I'll provide a sample presentation.
- submission **May 5**

Project

details of required work

- Think of it as a guideline for fellow students: How to build a knowledge graph from a technical text?
- one notebook, with
    - approximately 1 whole A4 page
      ~500 words of textual explanation in MD
    - many comments, in particular each function needs one
- annotation guidelines:
    - what did you call a term?
    - what did you call a relation?
- How did you start? (bootstrapping)
- What and how to compare? (baselines, partial matching…)
- visualization of the annotations?
- overview of global results of your annotated dataset
    - can you type/group the terms and the relations?
    - do you have un-related terms/sub-graphs?
- comparison with existing work?
- If you had more time, you would…

Last notebook

Finish notebook 4 including the Prodigy part on the G06F corpus for next Friday April 7

- use our Discord group to get help to get things running on your system

active learning

- Prodigy https://prodi.gy/
- https://en.wikipedia.org/wiki/Active_learning_(machine_learning)
  - optimal experimental design

Merci de votre

- attention
- considération
- intérêt
- écoute
- présence
- curiosité
- question

!