Improving tool mention clustering

Mapping the impact of research software

Daniel Garijo - Universidad Politécnica de Madrid
daniel.garijo@upm.es

The problem(s)

Tool names are too similar for clustering techniques!

1 - **Similar** (but different) mentions are currently grouped together

Panda, panda tool

Pandas, Python data analysis library

Group 1 (jaro-winkler)

2- **Exact match** mentions may be referring to different tools

?

fude was used…

fude: Utilities for Fude Polygon

Provides utilities to facilitate handling of Fude Polygon

3 - The **same tool** may be in different registries (pypi, wikidata, scricrunch)

What we'll do at the hackathon

Can we **improve groups** (clusters) and reconciliation?

- Detect errors once clustering results are obtained? (in CZI mention dataset)
  - E.g., two clusters resolve to different tools

- Describe a software mention not only in terms of name, but mention metadata
  - Paper topics, domain, name
  - Other tools mentioned in the paper

- Use tool metadata (reconciliation/disambiguation)
  - Wikidata, pypi, cran, github, etc.

Available resources

- CZI dataset (PubMed central analysis)
- Softalias KG (based on CZI + Wikidata)
- Somesci/someNLP dataset (includes alignment with Wikidata)
- Wikidata, cran, pypi, etc.