Machine learning examples on Invenio

Samuele Kaplun

IUGW2017

Disclaimer

- Presenting INSPIRE use-cases for treating HEP Literature
- INSPIRE has implemented or is using several tool that had to be trained on multi-GPU machines
- Some models we provide out of the box

Text classification

- Magpie: deep learning tool for multi-label text classification
  - title/abstract -> subject
  - title/abstract -> keywords
  - title/abstract -> experiment
  - …
- At INSPIRE we have trained for the above use cases, but others are possible (need powerful Multi-GPU for training)

Author disambiguation

Be
Bibliographic Entity Automatic Recognition and Disambiguation

- https://github.com/inspirehep/beard
- Based on scikit-learn
- Clusters together signatures of authors within papers:
  - Affiliations
  - Co-authorship
  - Subjects
  - ...

- Used at INSPIRE to prepare author profiles

CERN

INSPIRE HEP

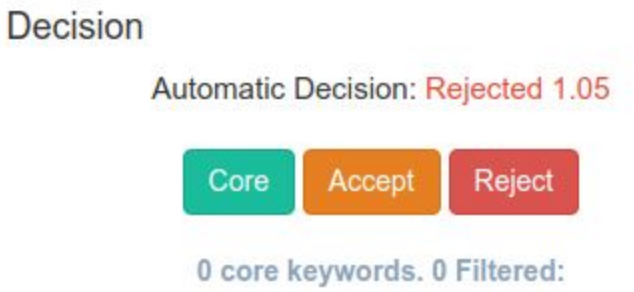
Metadata extraction from PDF

- https://github.com/kermitt2/grobid
- Parses PDF and extract all possible metadata from front-page or references

- This is a generic off-the-shelf tool that can be deployed locally and offers REST interface

Ad hoc

- At INSPIRE we are also having trained a ML model for content selection:



That has learned from catalogers

Conclusion

- Machine learning to help curation and content selection
- Many Open source solutions available

Magpie: https://github.com/inspirehep/magpie

Beard: https://github.com/inspirehep/beard

Grobid: https://github.com/kermitt2/grobid

CERN

INSPIRE HEP