OPEN FLOOR. DISCUSSION ABOUT TEXT MINING, MACHINE LEARNING, A DATABASE WITH 28 LANGUAGES

Daniela Ceccon
Manager of Databases
WageIndicator Foundation, Amsterdam – CLARIN, Utrecht
danielaceccon@wageindicator.org

Huub Bouma
Developer of the Collective Agreements Database
WageIndicator Foundation, Amsterdam

Stefano Ceccon
PhD, Data Science Lead

Developed within the SSHOC project.

Supported by the European Commission, DG Employment, Social Affairs and Inclusion, VP/2018/004/0008
Developed within the SSHOC project.

1440 Collective bargaining agreements (CBAs) texts from the WageIndicator CBA Database (since 2012)

WageIndicator.org/cbadatabase

28 languages
50+ countries
Annotated: answers to 249 labour rights related questions on 9 topics (e.g., Employment Contracts, Gender Equality Issues, etc.) + clauses selected

1. Dataset TEXTS (.csv dump with all CBA texts in html)
2. Dataset CLAUSES (.csv dump with all clauses assigned to a question (= 'bind'))

AIM OF THIS WORK

To ease future CBA texts annotation by finding the parts of texts where a question is answered = assign a 'bind' to paragraphs in new CBA texts

DATA PROCESSING:

1. We parse texts in paragraphs and create a 'paragraphs dictionary' with languages as keys, containing all the paragraphs for each language.
2. For each clause selected, we check whether it is contained in a paragraph. If that is the case, then the bind is assigned to the paragraph in a new data frame with 1 or 0 identifying whether the bind is assigned to that paragraph or not.
3. We can only do the training on binds that have a sufficient number of assigned paragraphs: we decide for 5 as a minimum.
4. We perform cleaning (tokenisation - lemmatisation - stop words removal) using NLTK tools (WordNetLemmatizer for English, Snowball Stemmer for other languages).
5. We add a column with cleaned paragraphs to our data frame.

Python script

Dataset PARAGRAPHS (all cleaned paragraphs with 1 or 0 for each bind)
MODELS for CBA ANNOTATION

- RELATIVE FREQUENCY*
  - BAG OF WORDS and Term Frequency (TF) / Inverse Document Frequency (IDF)*

- UNIVERSAL SENTENCE ENCODER

- PAR2VEC

- BERT

*Simpler and based on word frequency

Developed within the COLBAR-EUROPE project.