D5: Databases and Information Systems
Information Extraction, WS 2019/2020
Simon Razniewski & Cuong Xuan Chu
Lab #07: Open Information Extraction

max planck institut
informatik

Problem 1 (Open IE).
In this lab, we are working on open information extraction: given a sentence, extract all statements, S-P-O. We provide a dataset[1], which includes:

- `./oie_corpus/sentences.txt`: 1601 sample input sentences
- `./oie_corpus/extractions-groundtruth.oie`: The corresponding extraction groundtruth statements. Each line represents a single Open IE extraction, in a tab separated format:
  `SENTENCE [tab] PREDICATE_HEAD [tab] FULL_PREDICATE [tab] ARG1 [tab] ARG2 [tab]....`

You can use any tool to pre-process the data, like POS tagging, dependency parsing, etc. You may also use pretrained word embeddings like word2vec or BERT. However, you are not allowed to use any existing open information extraction or semantic role labeling tools.

Your program, called `run.py`, takes a file (e.g. `./oie_corpus/sentences.txt`) as the input and returns the output in a file (e.g. `./output/results.txt`) that has the following format:

```
sentence_1
id_of_sentence_1 [tab] "subject_1" [tab] "predicate_1" [tab] "object_1" [tab] 0
id_of_sentence_1 [tab] "subject_2" [tab] "predicate_2" [tab] "object_2" [tab] 0
...
```

The last column represents the confidence score, please supply always `0`.

Note: Our baseline method extracts all triples of nominal subject, verb, direct object from sentences (your output file should have the same format as the baseline output file, `./output/baseline_results.txt`).

How to run: `python run.py input_file output_file`
Example: `python run.py ./oie_corpus/sentences.txt ./output/baseline_results.txt`

How to evaluate: The evaluation script takes the output file as input, computes precision and recall (P/R) scores which are saved in a `.dat` file, and return the maximal F1-score.

```
./eval.sh output_file
```
Example: `./eval.sh ./output/baseline_results.txt`

To run and evaluate: `./run_eval.sh input_file output_file`

Your submitted files must include all necessary code and files, especially the main program file `run.py`. If you used any external libraries, please indicate them in a README file.

Please submit all necessary files, which are compressed into a zip file named:
Lab07_MatriculationNumber_Name.zip
to the email address: **cxchu@mpi-inf.mpg.de** with title of the email: **[IE]Lab07_MatriculationNumber_Name**

Deadline: 23:59 07.12.2019 (Saturday)

[1] Link on course website, taken from Stanovsky & Dagan: Creating a Large Benchmark for Open Information Extraction, *EMNLP 2016*.