Dictionary-based text tagging of articles related to COVID-19

Jesper Laurell & Jennie Karlsson

Project idea

- Covid-19 Pandemic
- Kaggle Task
- Open research dataset

Tagger

Dictionaries

- Virus
- Disease
- Symptoms

Input

- Json
- Parsing
- Kaggle database

```
covid19
covid 19
ncp
coronavirus disease 2019
corona virus disease 2019
coronavirus disease19
```

Tagger

Output - Json


```
{"cord_uid":"PMC7159299","sourcedb": "PMC", "sourceid": "n/a",
"div_id": 0, "text": "Clinical features of patients infected
with 2019 novel coronavirus in Wuhan, China. ", "denotations":
[{"id": "JJ_dict_v1_virus", "span": {"begin": 44, "end": 66},
"obj": "Virus_SARS-CoV-2"}]}
```


PubAnnotation

Clinical features of patients infected with Virus_SARS-CoV-2 2019 novel coronavirus in Wuhan, China.

Tagger - the algorithm

Idea

- Sort dictionaries
- Regex - finditer
- Loop article for each dict

Considerations

- Hyphens
- Capital Letters

Evaluation

- Gold standard
- Compare
- Performance assessment crucial
- Identify weaknesses
- Risk: optimize tagger for gold std

Evaluation scores

- Recall
- Precision
- Micro
- Macro
- F1 score
- Runtime

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Results

Dictionaries	Precision	Recall
Disease_COVID-19	70%	70%
Virus_SARS_CoV-2	40%	15%
Symptom_COVID-19	68%	61%

Total	Precision	Recall	F1-score
Micro	67%	59%	63%
Macro	59%	49%	53%

Results

Class	TP	FP	FN	True Total
Disease_COVID-19	32	14	14	46
Virus_SARS-CoV-2	2	3	11	13
Symptom_COVID-19	17	8	11	28

	Runtime [s]
Tagging gold std	1.5
Evaluating gold std	0.002
Tagging subset_100	120

Final version

Assessment

- Runtime
- Compact code
- 2/3 dicts acceptable results

Future improvements

- Improve dictionaries
- Use more dictionaries

THANK YOU!

QUESTIONS?