Distinguish companies from noise.

Alexey Osipov

16.03.2020

Statement of the problem.

Given: the file with 21332 rows with data on GLE claims. It has column Claim_Name. Two cases:

- It corresponds to the company, e.g. Grangl GmbH
- It corresponds to the noise, e.g., DO NOT USE Sweco Lietuva UAB

How to distinguish these two cases?

Regular expressions approach.

- turn to lowercase
- If it contains az, ag, inc, lp, corp, gmbh, ltd, a.s., sp., s.r.o., llc, l.p., p.c., p.a., b.v., then it is a company.
- If it contains dr, accident, prof, year, do not use, tba, then it is a noise.
- Manual validation (knownnotcompanies.csv).
- The approach was used to prepare a labelled data set for the ML approach.

ML approach, part 1

- Feature extraction: hasComma, hasDigit, hasBracket, numberOfWords, averageWordLength, numberOfLines,...
- Metric is F1-score, hold out set (20%), train-test split, 3-fold cross-validation
- Candidate algorithms: random forest, decision tree, gradient boosting, logistic regression, SVN.
- The best was gradient boosting, the second was random forest.
- F1-score was good, but there were problems on the unlabelled dataset.

ML approach, part 2

- Only the best features:

```
In [12]: pipeline.showColumnsByImportance(X_model.columns, model.feature_importances_)
```

Out[12]:

	columns	importances
5	Claim Security Indicator	0.527086
3	numberOfWords	0.194896
4	averageWordLength	0.119226
0	hasDigit	0.069908
1	numberOfLines	0.068502
7	Source System	0.007228
6	Type of Agreement	0.006932
2	hasBracket	0.006221

- Model: Decision tree.

$$F1_{CV} = 0.9062, F1_{test} = 0.8905, F1_{rep} = 0.932.$$

About 66% of data corresponds to companies.

- Solution in R and Python.