Classifying enterprises by economic activity

Cedefop/CRISP/ESSNet

21 March 2018, Milan

Outline

- What is this about
- Few small presentations (~7 min)
  - UK – Fero Hajnovic
  - Germany – Chris Gabriel Islam
  - France – Maxime Bergeat
  - Belgium – Thomas Delclite
- Discussion

What is this about?

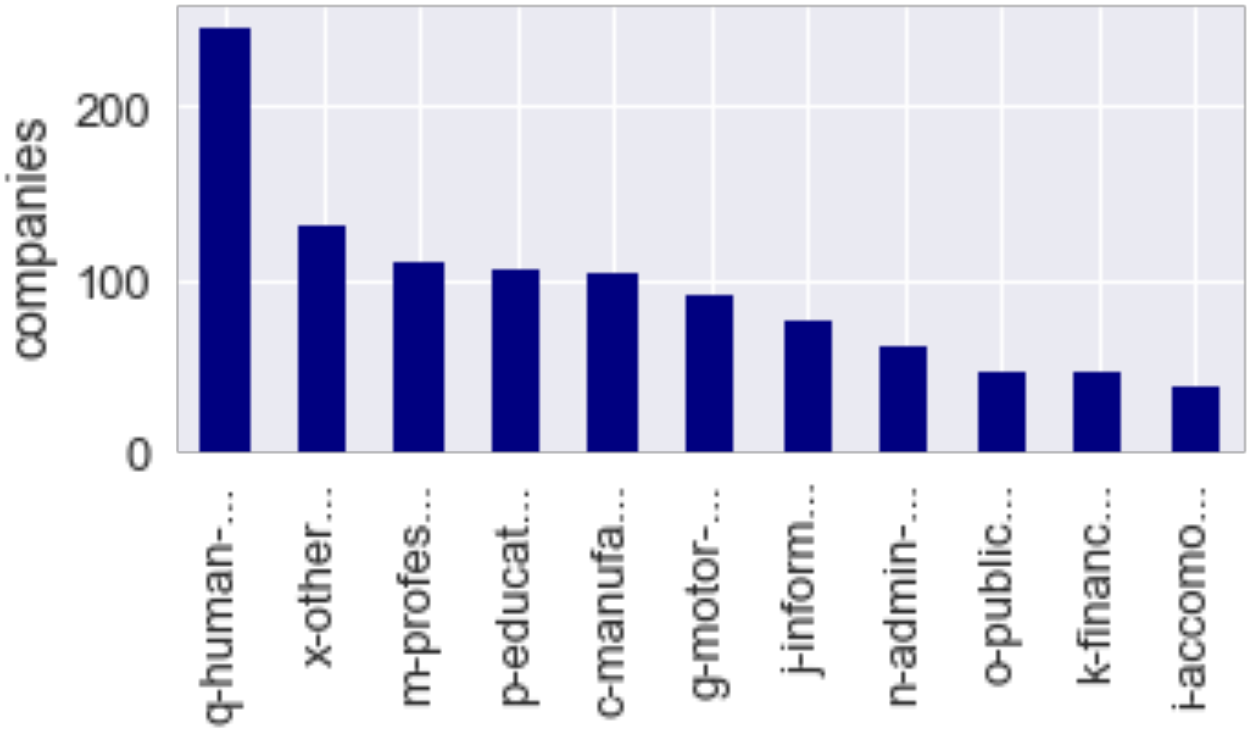
- Methods for extracting economic activity from job ads
- Assessment of the algorithms (accuracy, speed, applicability)

UK

Fero Hajnovic

Data

- Adzuna, 10th of Jan. 2018
- Survey (has SIC)
- Matching



	company_name	sic_class
0	Charters School	p-education
1	Disabled Living Foundation	q-human-health
2	Skylon Restaurant	i-accomodation-food-services
3	Multimatic	c-manufacturing
4	Marketing Sciences	m-professional-scientific-technical

	Ads	Comps
Before matching	~ 1 000 000	~45 000
After matching	~ 10 000	~ 1000

Machine learning

1. NLP

```
The Skills People Group are
the largest provider of...
```

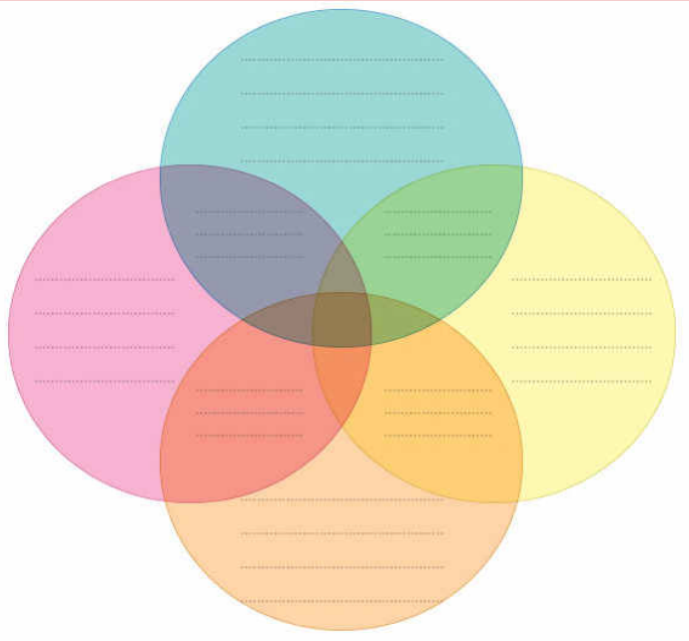
```
skills people group largest
provider funded commercial...
```

2. TF-IDF

```
1051x27808 sparse matrix with
385703 stored elements
```

3. Word selection

- low IDFs
- 2000 words
- in some…
- ...but not all SICs

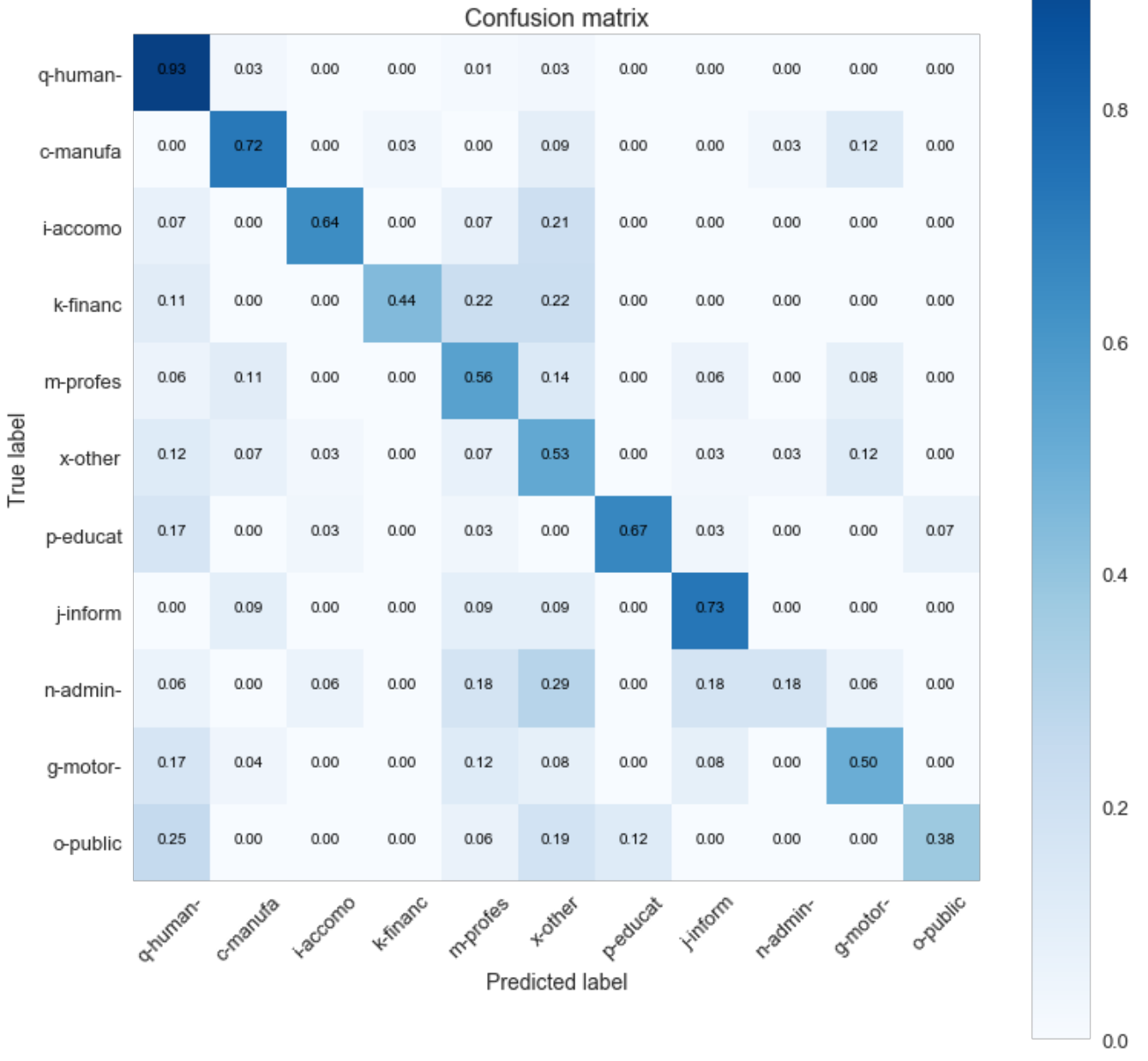


4. Classifier + grid search

- Multinomial NB
- SVM

Results

- Accuracy
  - 66% (NB)
  - 60% (SVC)
  - 55% (Combined - NB + Random forest)
- Similar recall
- Todo
  - More data
  - Equal distribution
  - Word selection



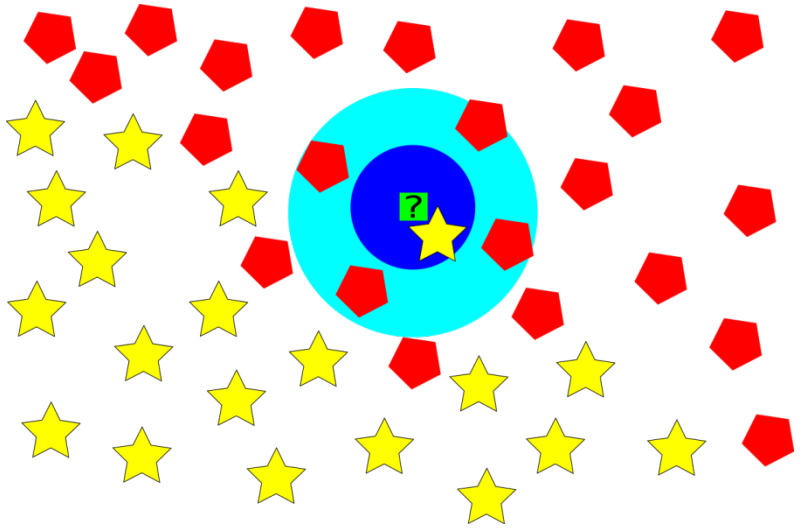
Germany

Chris Gabriel Islam

Using R on CEDEFOP data

- Used Package: Class
- Algorithm: K-Nearest-Neighbour, Accurarcy: 78 %

	A	C	D	F	G	H	I	J	K	M	N	Q	R	
A	317	1	0	0	0	0	0	0	0	4	2	0	0	98%
C	5	758	0	5	183	1	367	0	0	50	2	2	0	55%
D	0	0	0	3	0	0	0	0	0	0	0	0	0	0%
F	1	3	1	1804	7	44	1	0	0	27	48	1	0	93%
G	0	109	0	26	806	22	131	1	12	13	506	3	0	49%
H	0	3	0	10	13	4626	7	0	3	19	83	0	0	97%
I	0	366	0	3	171	5	181	1	0	7	67	2	0	23%
J	0	0	0	4	2	4	0	29	1	7	30	0	0	38%
K	0	0	0	0	9	0	0	0	39	9	54	0	0	35%
M	2	77	0	47	24	32	8	10	8	940	224	15	1	68%
N	1	4	0	57	358	112	50	10	34	92	2999	42	0	80%
Q	0	0	0	0	0	0	1	0	0	11	17	482	0	94%
R	0	1	0	0	1	0	0	0	0	1	0	0	66	96%
	97%	57%	0%	92%	51%	95%	24%	57%	40%	80%	74%	88%	99%	



Source: Wikipedia, Adrichel

Using Python on Federal Employment Agency data

- Used Packages: NLTK and Sklearn
- Algorithm: Multinomial Naive Bayes Classifier, Accuracy: 85 %

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
A	446	0	19	0	0	21	48	14	17	1	0	1	5	33	1	0	4	1	2	0	0	73%
B	0	9	16	0	0	14	2	6	0	0	0	0	3	4	0	0	0	0	0	0	0	17%
C	15	0	8866	0	1	539	597	98	221	172	0	1	95	728	0	0	48	0	7	0	0	78%
D	2	0	24	51	0	16	1	0	2	11	0	0	21	6	7	0	0	0	0	0	0	36%
E	0	0	24	0	221	37	16	45	1	4	0	3	10	30	11	0	1	0	0	0	0	55%
F	1	0	231	0	3	6711	88	52	11	14	0	8	90	220	0	0	8	0	1	0	0	90%
G	21	0	673	0	10	294	8598	261	223	254	1	6	112	400	0	4	53	2	23	1	0	79%
H	4	0	52	0	1	69	152	3103	23	40	1	0	115	201	0	1	9	1	1	0	0	82%
I	2	0	22	0	0	5	39	8	5747	3	0	1	18	29	4	1	67	5	8	0	0	96%
J	0	0	79	0	0	28	46	8	11	1879	1	0	44	40	9	0	6	1	3	0	0	87%
K	0	0	5	1	0	4	22	0	3	32	478	2	29	30	2	0	5	1	3	0	0	77%
L	1	0	3	0	0	32	9	2	38	2	3	194	45	37	4	0	5	3	2	0	0	51%
M	4	0	435	0	12	154	399	112	97	319	5	16	4352	242	21	46	299	15	44	0	0	66%
N	55	0	524	1	11	695	260	190	240	286	9	27	267	25618	13	12	347	9	20	1	0	90%
O	1	0	3	0	3	5	2	1	4	10	0	0	1	1	954	19	53	4	5	0	0	89%
P	6	0	4	0	0	1	4	11	15	16	0	0	18	24	16	848	172	10	7	0	0	74%
Q	3	0	15	0	0	6	12	18	98	24	0	0	35	45	15	107	9747	10	13	1	0	96%
R	5	0	7	0	1	5	20	1	105	5	0	0	11	21	9	7	19	304	5	0	0	58%
S	8	0	22	0	0	29	71	15	64	27	2	2	37	60	32	111	146	6	1393	0	0	69%
T	2	0	0	0	0	3	1	1	9	0	0	0	1	3	0	0	69	0	3	12	0	12%
U	0	0	1	0	0	1	1	1	0	0	0	0	1	1	3	0	0	0	0	0	31	78%
	77%	100%	80%	96%	84%	77%	83%	79%	83%	61%	96%	74%	82%	92%	87%	73%	88%	82%	90%	80%	100%	

Outlook

- Focus on signal words or dictionaries
- ML in order to detect ads from employment agencies
- Try out other packages and tools: FastText, Tensorflow, etc.
- Cooperate with collleagues in our institute on our newly built ML platform

France

Maxime Bergeat

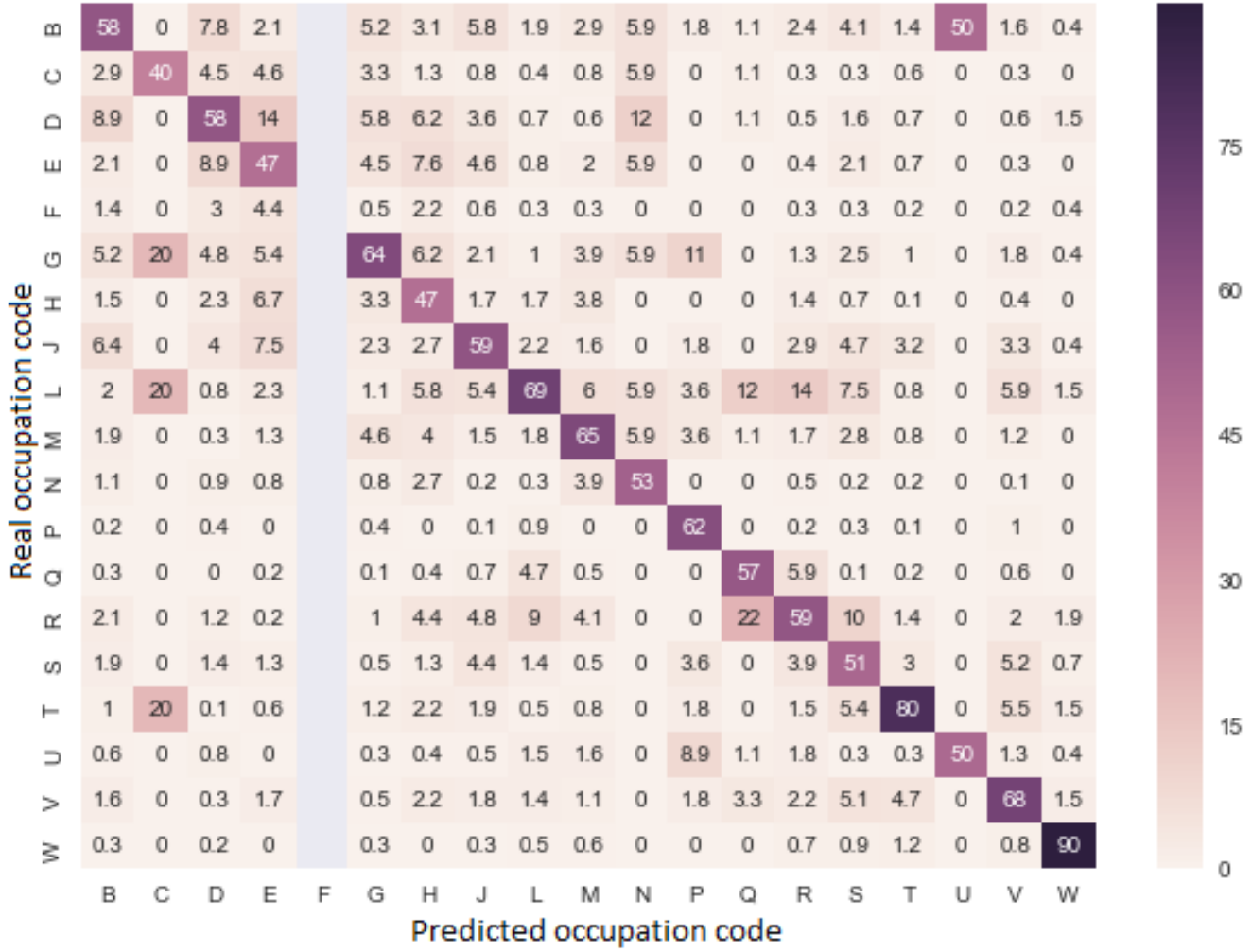
Methodology

- Data received from National Employment Agency (Pôle emploi)
  - ~140 websites
  - Subset of 100 000 job ads
  - Always including job description and coded occupation (no job title ☹)
- Variables used as predictors
  - Words from job description (input is basically a document-term matrix)
  - Selection of words is based on term frequencies of words.
- Mostly Python sklearn library
- Goal: classification of occupation (22 categories)

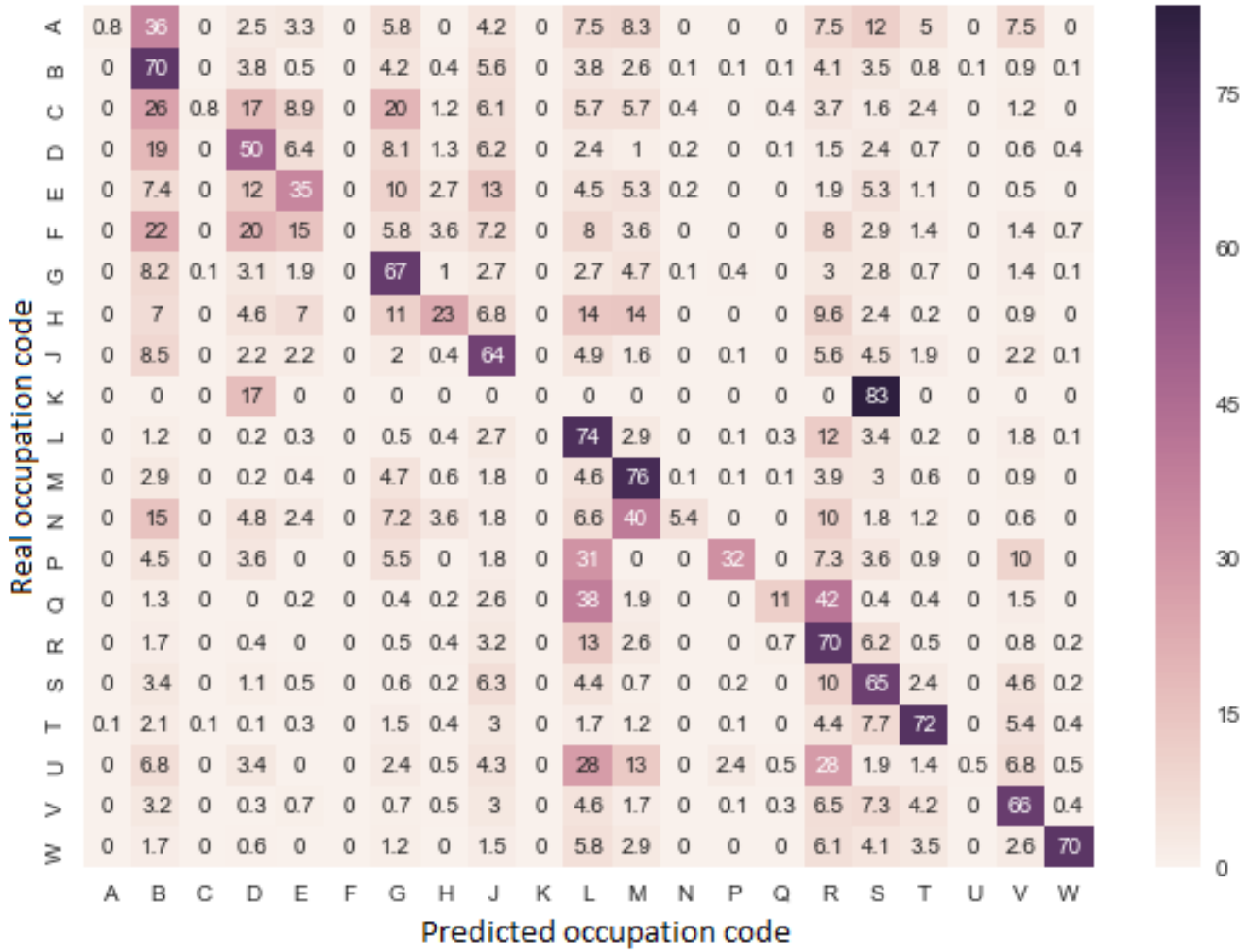
Application #1

- ~1500 words used as predictors (words present in > 0.5% of all documents)
- Different methods tried for classification
  - Logistic Regression
  - Gradient boosting
  - Random Forests with bagging
  - Basic neural network
  - Support vector machine
- Here: results presented for
  - Logistic regression
  - With L1 penalty
- Evaluation of results: train and test datasets (80% / 20%)
- Global accuracy: 70%

Application #2 – Precision matrix



Application #3 – Recall matrix



Future work

- Improve selection of words used as predictors
  - Choose words which are occupation-specific
  - Improve tuning of parameters used in predictive methods
  - Improve evaluation process -> Cross-validation

- First results seem good…
  - … Global accuracy of 80% with selection of occupation-specific words for predictors and basic tuning of parameter (for penalty) with random forests
    - (words present in > 2% of documents for at least one occupation category)
  - First experiments show that cross-validation does not improve a lot robustness of evaluation results (k-fold cross validation).

- To be continued…

Belgium

Thomas Delclite

STATBEL
La Belgique en chiffres

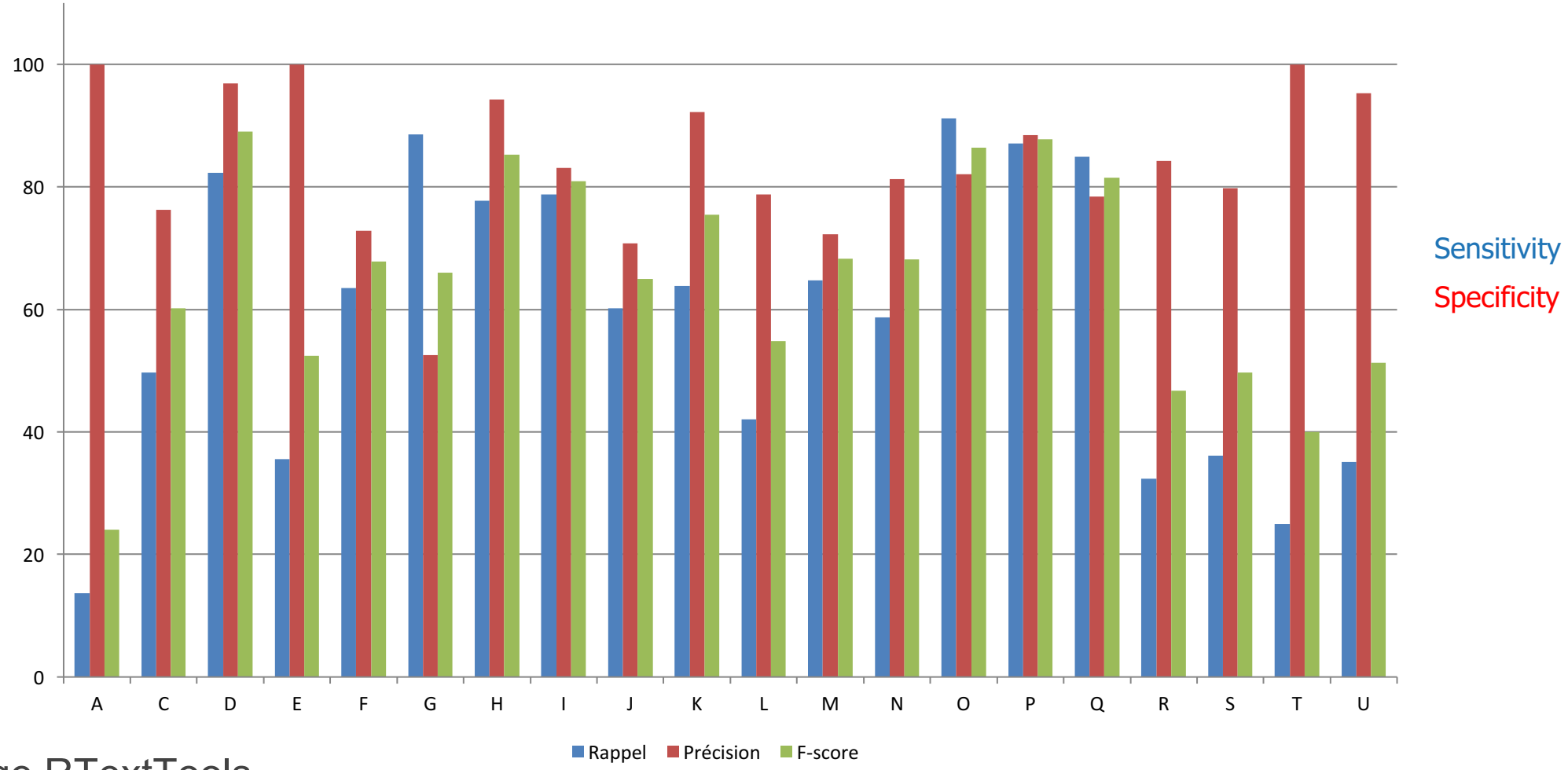
Machine Learning and prediction of NACE code by job description

ESSnet Big Data, Milan, 03/2018

economie
SPF Economie, P.M.E., Classes moyennes et Energie
.be

Using R package on administrative data

Précision de détection des NACE par Machine Learning (Actiris FR)
65210 offres (50 000 en apprentissage, 15210 en test)



Package RTextTools

http://statbel.fgov.be

STATBEL
La Belgique en chiffres

Using R package on administrative data

		NACE code by Machine Learning																			Sensitivity
		A	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
Real NACE code	A	3	0	0	0	1	16	0	0	0	0	0	1	0	0	0	1	0	0	0	0	14%
	C	0	469	1	0	41	287	2	27	16	0	1	44	18	14	2	22	0	0	0	0	50%
	D	0	0	219	0	1	16	0	0	6	0	0	7	5	7	0	3	0	2	0	0	82%
	E	0	0	0	16	3	9	0	1	1	0	0	2	3	8	2	0	0	0	0	0	36%
	F	0	4	0	0	671	200	7	21	6	1	6	45	24	42	3	24	0	3	0	0	63%
	G	0	26	0	0	39	4320	29	120	85	1	2	118	37	29	10	44	1	18	0	0	89%
	H	0	4	0	0	11	213	1499	3	18	0	1	25	5	57	0	89	0	4	0	0	78%
	I	0	10	2	0	6	355	1	1879	3	0	1	23	19	40	3	37	2	4	0	0	79%
	J	0	14	2	0	14	386	6	11	1096	1	3	152	40	61	4	28	0	5	0	0	60%
	K	0	0	0	0	1	138	0	1	23	524	8	73	22	24	1	5	0	1	0	0	64%
	L	0	2	0	0	12	118	1	21	4	1	249	48	20	32	8	69	1	6	0	0	42%
	M	0	44	0	0	45	876	10	38	134	29	10	2847	130	103	14	106	2	14	0	0	65%
	N	0	18	1	0	49	699	14	35	106	8	8	313	2116	108	12	107	2	6	0	0	59%
	O	0	4	1	0	5	51	4	12	6	2	12	33	20	6950	362	156	1	6	0	0	91%
	P	0	3	0	0	7	59	2	10	7	0	1	57	14	295	5141	276	8	21	0	1	87%
	Q	0	8	0	0	10	167	9	56	12	0	4	42	60	375	187	5499	6	45	0	0	85%
	R	0	3	0	0	2	103	0	7	9	0	2	22	7	91	26	158	213	15	0	0	32%
	S	0	6	0	0	3	193	3	16	13	1	8	86	58	232	39	389	16	601	0	0	36%
	T	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	1	0	0	2	0	25%
	U	0	0	0	0	0	12	3	2	4	0	0	4	2	5	0	2	1	2	0	20	35%
Specificity		100%	76%	97%	100%	73%	53%	94%	83%	71%	92%	79%	72%	81%	82%	88%	78%	84%	80%	100%	95%	

- Package RTextTools

STATBEL
La Belgique en chiffres

Issues

- Predictions for three languages
- Application to webscraping data
- Quality of webscraping data to assess variations of job vacancy

Discussion

- SIC classification in Cedefop data?
- Why is this useful?

Thank you