Unsupervised Keyword Extraction from Polish Legal Texts

Michał Jungiewicz [1,2] and Michał Łopuszyński [1]

[1] Interdisciplinary Centre For Mathematical and Computational Modelling, University of Warsaw

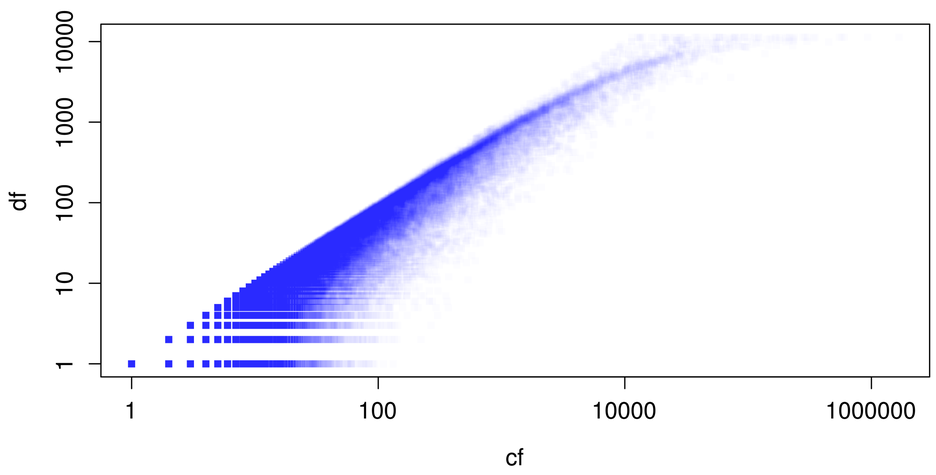
[2] Faculty of Electronics and Information Technology, Warsaw University of Technology

How & Why?

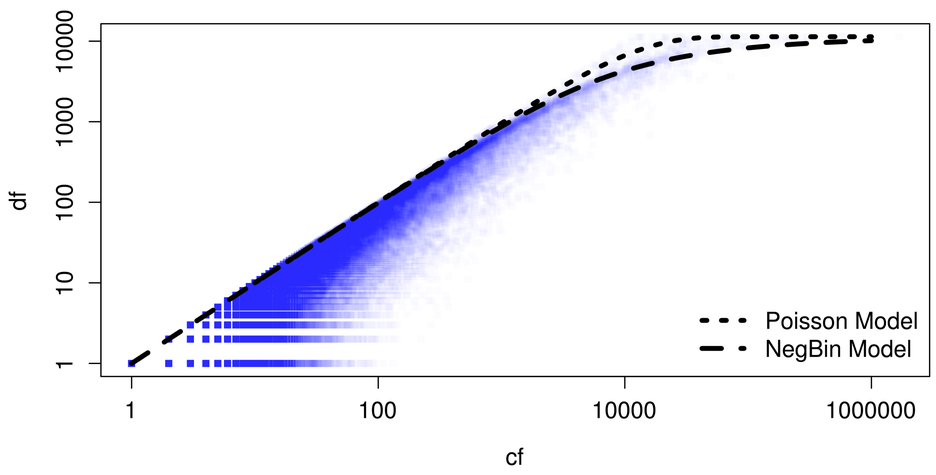
- We look for KE method which will be **unsupervised**, suitable for **Polish legal texts** and potentially extensible to **other languages**
- Obtained keywords to be applied in ML & IR tasks
- Test corpus = 11k judgements from National Appeal Chamber (KIO)
- We employ unsupervised keyword extraction method RAKE [1]
- RAKE is a language independent method, the only language specific input is a set of uninformative words (stoplist)
- This is **work in progress report** on supplementing RAKE with automatic stoplist generation

Automatic Stoplist Generation

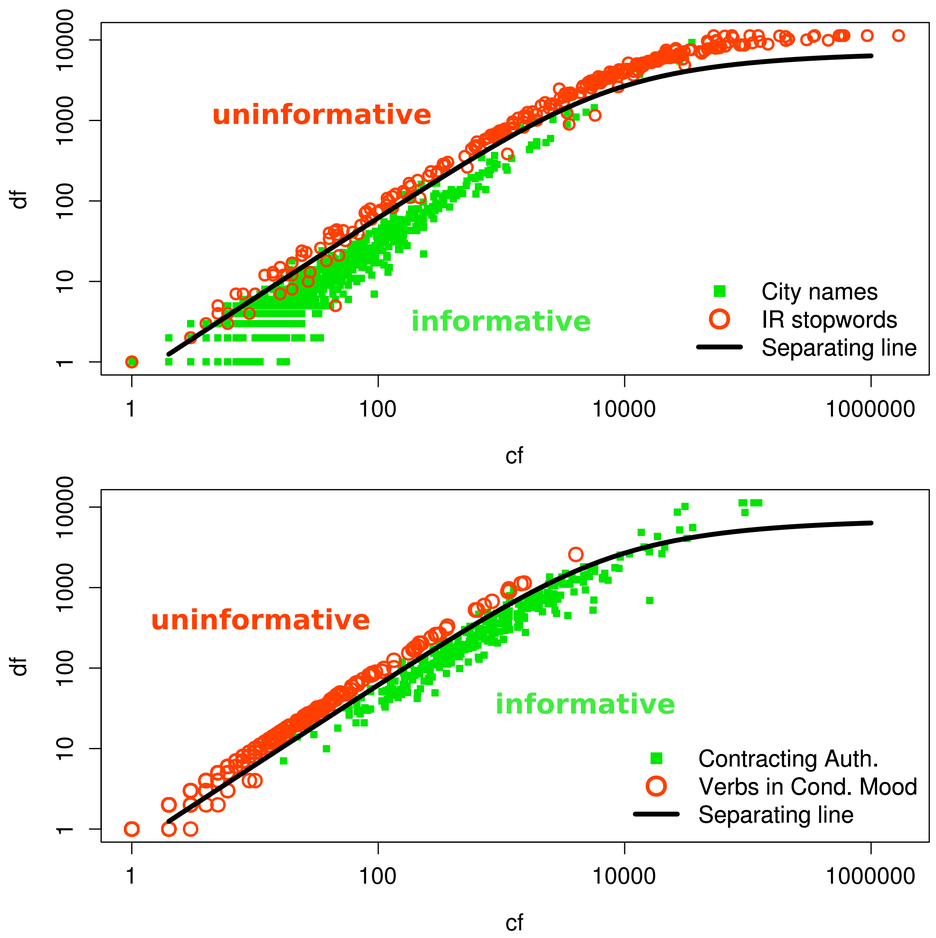
- Uninformative words can be distinguished by plotting number of documents with a given word (**df**) vs. collection frequency of given word (**cf**) [2, 3]



- Using random model of word occurrence (e.g. Poisson or NegBin) the relation between **cf** and average **df** can be computed



- Uninformative words closely follow random model, whereas informative words occur in more clustered fashion, which results in lower average **df** than resulting from random prediction [2,3]



- Words that deviate from random model $\overline{\mathrm{df}}(\mathbf{cf})$ less than certain threshold p are treated as uninformative words for RAKE

$$\frac{\overline{\mathrm{df}}(\mathbf{cf}) - \mathbf{df}}{\mathbf{df}} < p$$

Sample Results

- Top-score keywords obtained using **standard IR stoplist**

Polish	English translation
samej grupy kapitałowej dotyczącego wykonawcy Przedsiębiorstwo Usług Komunalnych Empol sp.	the same capital group concerning the contractor Municipal Services Company Empol
Dzienniku Urzędowym Unii Europejskiej 23 marca 2013 r.	(in) the Official Journal of the European Union 23 March 2013
Prezesa Krajowej Izby Odwoławczej 20 czerwca 2012 r.	Chairman of the National Appeal Chamber 20 June 2012
pierwszej kolejności Krajowa Izba Odwoławcza winna ocenić	firstly the National Appeal Chamber should judge
Krajowa Izba Odwoławcza uwzględniła odwołanie konsorcjum Sita Małopolska	National Appeal Chamber has upheld the appeal of the Sita Małopolska Consortium

☹ Many **long**, **uninformative** keywords

- All keywords obtained using **autogenerated stoplist**

Polish	English translation
Przedsiębiorstwo Usług Komunalnych Empol	Municipal Services Company Empol
przedsiębiorstwo usług komunalnych	municipal services company
zagospodarowanie odpadów komunalnych	management of municipal waste
odbieranie odpadów komunalnych	municipal waste collection
właścicieli nieruchomości zamieszkałych	residential real estate owner
konsorcjum Sita Małopolska	Consortium Sita Małopolska
Sita Małopolska	Sita Małopolska
grupy kapitałowej	capital group

☺ **Short**, **interpretable** keywords

- Top-frequency keywords in the corpus (**autogenerated stoplist**)

All keywords

Polish	English translation
roboty budowlane	construction works
robót budowlanych	construction works (different form)
konsorcjum firm	consortium of companies
ograniczoną odpowiedzialnością	limited liability
formularzu ofertowym	offer form

Four tokens keywords

Polish	English translation
PKP Polskie Linie Kolejowe	PKP Polish State Railways
Generalnej Dyrekcji Dróg Krajowych	General Directorate for National Roads
samodzielny publiczny szpital kliniczny	independent public clinical hospital
wykazu wykonanych robót budowlanych	list of conducted construction works
GE Medical Systems Polska	GE Medical Systems Poland

☺ Contain **names of companies** and **institutions**

☺ Mostly **noun phrases** as expected for manually assigned keyprhases [4], even though no POS information was used

Conclusions and Outlook

- Preliminary results seem promising
- Quantitative verification is a must, but we know of no established corpus for Polish language (esp. for legal texts)
- Quantitative verification as features in ML tasks might be an option
- Based on tests, further tuning & extensions of the method are planned

References

1. S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, in *Text Mining, Applications and Theory*, Wiley (2010)
2. K. W. Church, W. A. Gale, Poisson mixtures, Natural Language Engineering, 1, 163 (1995)
3. C. D. Manning and H. Schütze, Foundations of statistical natural language processing, MIT Press, Cambridge (1999)
4. J. Chuang, C. D. Manning, and J. Heer, Without the clutter of unimportant words: Descriptive Keyphrases for Text Visualization, ACM TOCHI 19, 19 (2012)