Detecting topics in civil service job offers using Latent Dirichlet Allocation model

Adam Bień
University of Economics and Business Poznań
Maciej Beręsewicz
University of Economics and Business Poznań
Statistical Office Poznań

WhyR?, 26-29 September 2019, Warsaw

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

Agenda

Agenda

Introduction

This presentation deals with and describes an application of Latent Dirichlet Allocation Model to assess the demand for different kinds of labour provided by workers in the civil service.

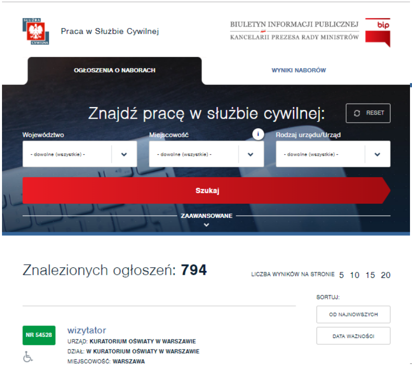
Introduction – motivation

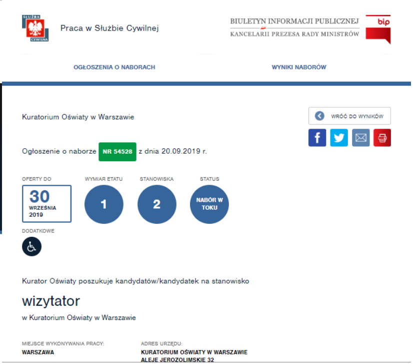
- Alternative approach to source data.
- Utilisation of Internet recruiting and information contained within job's offer description.
- Experimenting with different ways to examine labour demand, focusing on high speed and low cost of the study.
- Automation of grouping workers by the kind of labour they supply.
- Time-series analysis of demand for different kinds of labour.

Prime Minister's Office – work in civil service

- purpose – "creation of a modern nation, increasing efficiency of public administration bodies' operations and foremost satisfaction of polish citizens."
- `nabory.kprm.gov.pl` – job search engine for diverse professions in the civil service with exception for higher positions.
- Open competition recruitment procedure.

nabory.kprm.gov.pl – screenshots





Agenda

Gathering data

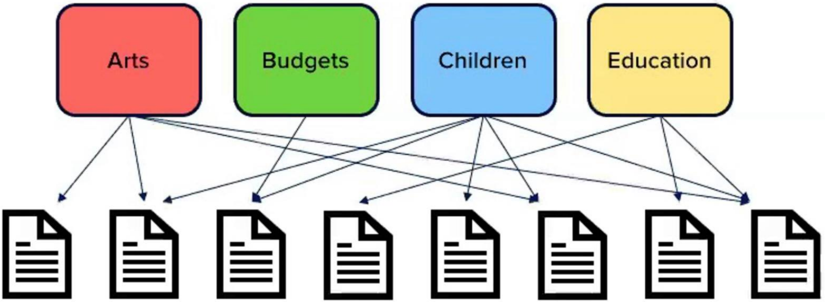
- The data was gathered via web-scraping means (automatic web-site browsing and data extraction).
- Stemming has been applied to the dataset using **morfologik** dictionary for polish language.

Dataset

- **Population**: Job offers published on `nabory.kprm.gov.pl` Web-site during 06.2016 - 04.2019 – over 42 thousand observations.
- **Variables**: ID, date of publication, deadline to send documents, title, description, requirements, work conditions, source (ongoing or archived), url.
- Not every variable was used in the research – room for improvement!

Agenda

Topic modelling



Latent Dirichlet Allocation

- Utility – identify topics (and words associated with those topics) over a set of documents.
- Proposed by Blei et al. in 2003.
- Can also be used in: sentiment analysis, object localization for images, automatic harmonic analysis for music, etc.

LDA – Assumptions

- Documents with similar topics will use similar group of words.
- Documents are probability distributions over latent topics.
- Topics are probability distributions over words.

LDA – Generative process (working backwards)

LDA assumes that new documents are generated in a following way:

- Determine number of words in a document.
- Assign a combination of topics to the document from a given set of topics (i.e 75% topic A, 0% topic B, 25% topic C).
- Generate a word in the document (for every word):
  - pick a topic (based on the document's distribution)
  - pick a specific word (based on the topic's distribution over word buckets)

LDA – Generative process example

Let the topics and the words associated with them be as follows:
Culinary (75%): {dish, food, slice, dice},
Machines (0%): {engine, wheels, tool, cog},
Animals (25%): {pig, cow, dog, cat}.
then:

- Choose length of a new document == 100.
- Choose topics (by document's distribution) and words (by topic's distribution) to fill in the document.
- Result = approximately 75 words from "Culinary" word bucket and 25 words form "Animal" word bucket (in random order).

Way LDA actually works – simplified

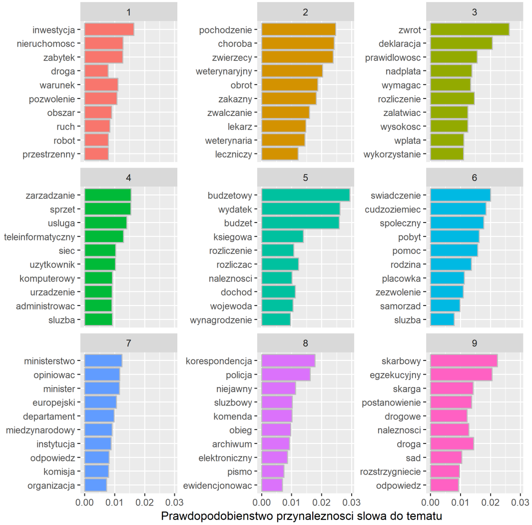
- Assign one topic (from fixed K number of topics) to each word in each document
- Select a document and assume, that every `other document` is assigned correctly. Compute:
  - Proportion of words in the selected document that are assigned to each topic ($Pr(\text{topic } t \mid \text{document } d)$)
  - Proportion of assignments to topic t over all documents, that come from word w ($Pr(\text{word } w \mid \text{topic } t)$).
- Multiply the two proportions and assign new topics to the words in the selected document based on the proportion.
- Iterate and eventually algorithm should reach a steady state in which assignments make sense.

Agenda

Research summary

- 9 topics have been discovered among the dataset.
- By analysing most crucial vocabulary for each topic I was able to name the topics as follows:
  - Infrastructure
  - Veterinary, food safety and hygiene.
  - Record and registration.
  - Informatics, programming and tech support.
  - Finanse, accounting and budgeting.
  - Social assistance and foreigners.
  - International politics.
  - National security.
  - Legislation.

Most crucial vocabulary for each identified topic



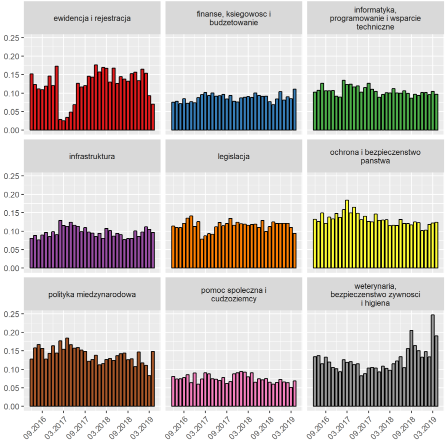
Quantitative research summary

By summing probabilities for topic assignment over all documents we can approximate topic shares for the dataset in a quantitative manner:

Topic	Number of offers	Share
Record and registration	5576	0.132
Finance, accounting and budgeting	3667	0.0867
Informatics, programming and tech support	4401	0.104
Infrastructure	4038	0.0954
Legislation	4931	0.117
National security	5457	0.129
International politics	5762	0.136
Social assistance and foreigners	3149	0.0744
Veterinary, food safety and hygiene	5340	0.126

Time series analysis

We can use previously obtained data to visualise how the demand for different kind labour changed over time:



Agenda

Conclusion

- Latent Dirichlet Allocation and other topic modelling algorithms offer identifying similarities among text documents.
- I was able to process over 42 thousand job offer descriptions, discover and name 9 topics they refer to and quantify them.
- This research gives an idea what can be done by embracing alternative data sources, with all it's pros and cons.
- There is still more room for improvement to this research, i.e analysing offers based on their status: successful, cancelled, archived etc.

Agenda

Bilbiography (selected positions)

- Ponweiser, M. (2012). Latent Dirichlet allocation in R
- Krotov, V. Silva, L. (2018). Legality and Ethics of Web Scraping
- Gładysz, A. (2014). Wykorzystanie metody opartej na ukrytej alokacji Dirichleta do automatycz-nej identyfikacji słów kluczowych w dokumentach.Logistyka, (3), 2011–2019.
- Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R. Delen, D. (2012).Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Elsevier Science.

Thank you for your time!

Appendix

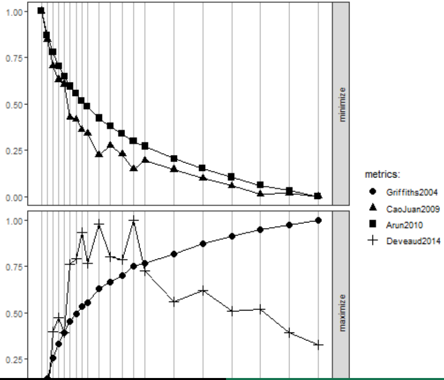
LDA – Mathematical equation

$$p(w, z, \theta, \phi|\alpha, \beta) = p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z, \phi). \quad (1)$$

where: w_i – word i in the document; z – corpus; θ_d – topic distribution over words in a document; ϕ_k word distribution over topics; α, β – Dirichlet distribution parameters.

LDA model evaluation

There are plenty of methods to evaluate this model. You can download and install `ldatuning` package to make this task easier. The package offers to compute and plot four of the most common methods:



More information

To read more about this project and how it was prepared I encourage you to read my Bachelor's Thesis and go through my code which you can find on my `https://github.com/adambien1`