Hybrid Teacher (Hybride Docent) Project

Prepared by: Vladimer Kobayashi

# Deliverables

|  |  |  |
| --- | --- | --- |
| Deliverable | URL | Delivered? |
| Github repo of the project | https://github.com/vkobayashi/teachhybridjobs |  |
| App | https://vkobayashi.shinyapps.io/analysis\_en\_vis/ |  |
| Documentation | This document |  |

# DOCUMENTATION

## Github repo

All codes associated to the project are in this Github repo:

https://github.com/vkobayashi/teachhybridjobs

The repo has 2 folders, namely analysis\_en\_vis and reports. The folder analysis\_en\_vis contains the codes of the App and the reports folder contain the reports prepared for this project.

The main R file containing the code for the data analysis is the analysis\_part1.R

## App

The app is a simple interface which helps users (e.g. teacher, policy-makers) explore opportunities for hybrid teachers. A sample use case is:

Suppose you have this profile:

1. A Nederlands teacher (dutch language teacher) at the moment
2. highest educational attainment is HBO (according to dutch educational system) or an equivalent qualification
3. Interested in jobs with permanent contract
4. Interested in full time employment, i.e. at least 32 hours per week
5. Interested in jobs located in Amsterdam region
6. Interested in jobs not exceeding 40 hours per week

You first input these details in the appropriate fields and then choose the criterion. For choosing the criterion there are two options: Average Cosine and Maximum Cosine. The default option should be Maximum Cosine since this criterion gives jobs which often give more emphasis to the characteristics corresponding to the type of teacher provided. In this case the top function job classes for Nederlands teachers are

1. Recreatie, sport en toerisme
2. Onderwijs opleiding en training
3. Verkoop en handel
4. Gezondheidszorg en welzijn
5. Engineering
6. Kunst, cultuur en media

The results perfectly match our expectation, except perhaps for the Engineering profession class, maybe because jobs associated to Engineering class often mention good command of the dutch language as a requirement or as a preferred quality. Regarding the visualization, the size of the circles corresponding to the volume of the vacancies available in our database, the X-axis is the maximum of the maximum cosine scores and the Y-axis is the mean of the maximum cosine scores. Each vacancy has a cosine score so at the aggregate level (job group or function class) we compute aggregate statistics, namely mean and maximum.

We can now select a specific function class to further see the matching jobs and why it gave us this match. Suppose we select the top matching function class, that is, Recreatie, sport en toerisme.

We see in the second chart the matching jobs and the third chart the word cloud which shows the words that were used in the matching. The words that seem to be influential to the match are

1. Communicatieve (communication)
2. Sociale (social)
3. Aandacht (may refer to attention to details)

Documentation

# Introduction

The project was done in cooperation with Hybride Docent (<https://www.hybridedocent.nl/>) and de baaningenieurs (<http://debaaningenieurs.nl/>).

Our goal is to provide a proof of concept for hybrid teachers. The reference profession is (secondary) teaching profession. Consequently, we also propose a methodology on how to perform the matching.

Our strategy is to match teaching to non-teaching jobs and identify the quality of the match. We define match as the overlap of skills, competencies, roles, knowledge, tasks and responsibilities between teaching and non-teaching jobs. We outline a procedure by which we can assess the quality of match in the sections that will follow. By doing this, we are able to identify which among the non-teaching jobs are suited for teachers and which among the non-teaching job holders show potential for teaching jobs.

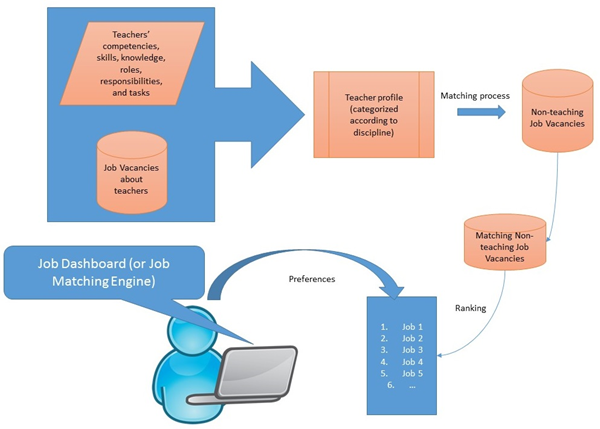
The motivation for doing this project is to remedy the problems of teachers’ burn-out and clamor for augmented salary by providing teachers with alternative jobs. Simultaneously, we address the problem  of teachers’ shortage by encouraging people from non-teaching professions to teach.

# Methodology

The methodology followed in this study is outlined below:

1. Collect data about teachers' competencies, knowledge, functions, tasks, and roles. These data will be used to define teacher profiles. Initially, we will consider the most commonly advertised teaching jobs, namely language teachers, math teachers, and geography teachers . Although teachers have shared characteristics, they also diverge on knowledge and tasks determined by the specific disciplines they are tasked to teach.  
2. Use vacancy data about teachers to enhance the profile of teachers. This will also serve to incorporate the labour market perspective about teachers such as the current demand for teachers.  
3. Based on the enhanced teacher profile, we will search for non-teaching professions that matches well with the teachers’ profiles. The data source for these non-teaching professions are online job vacancies. Here we will use a measure based from latent semantic analysis and cosine similarity to assess the extent of match. The matching can then be used for two purposes: First, matching teaching jobs to non-teaching jobs that may be recommended to teachers. Second, the job titles in the vacancies may be used to identify non-teachers who are potentially qualified to become teachers next to their current job role.

4. Create a dashboard that will help decision makers adjust thresholds for matching thereby allowing flexibility and verification of the matching results.



We discuss next the data sources used in this study.

# **Data Sources**

Two data sources were used in this project.

1. Data source compiled by teaching job experts
2. Job vacancies (for both teaching and non-teaching jobs)

## **Data Source from Teaching Job Experts**

The table below exhibits the sources where we collected information about teachers

|  |  |
| --- | --- |
| Teaching Characteristics | Sources |
| Competencies | https://ctmeter.nl/onderwijs/voortgezet-onderwijs/competenties |
| Knowledge | (Leraar Nederlands)  <https://www.examenblad.nl/examenstof/nederlandse-taal-en-literatuur-2/2017/vwo/f=/examenprogramma_nederlands_havo_vwo_2014.pdf>  (Wiskunde)  <http://www.slo.nl/downloads/archief/Examenprogramma_wiskunde_B_vwo_DEFINITIEF.pdf/>  (Aardrijkskunde)  <http://www.slo.nl/downloads/archief/Examenprogramma____aardrijkskunde____DEFINITIEF_5b1_5d.pdf> |
| Functions/ Tasks | https://onderwijscooperatie.nl/nieuws/voorstel-herijking-bekwaamheidseisen/ |
| Roles | <http://media.leidenuniv.nl/legacy/eindtermen-van-masteropleiding-leraar-voorbereidend-hoger-onderwijs.pdf> |

## **Teaching Job Vacancies**

As was mentioned, we also used teaching vacancies to augment the information we have about teachers. For this, we collected online teaching vacancies (or job advertisements). We provide next summary statistics about our teaching vacancies.

1) 9 teaching types: Informatiekunde/ICT-vakken, Scheikunde, Natuurkunde, Duits, Frans, Wiskunde, Aardrijkskunde, Biologie, Nederlands

## **Column names**

|  |  |
| --- | --- |
| **Column names** | **Description** |
| id | Unique identifier for each vacancy |
| date | Date a vacancy was posted |
| title | Job title |
| job\_location\_latitude | Job location (latitude) |
| job\_location\_longitude | Job location (longitude) |
| education\_level | Education level required |
| salary\_min | Minimum salary |
| salary\_max | Maximum salary |
| vac\_uid | Vacancy identifier |
| sector | Sector who posted the vacancy |
| org\_size | Organization Size |
| candidate\_descr | Candidate Description required for the job |
| job\_type | Teaching job type (one of aardrijskunde, nederlands, or wiskunde) |

# **Data Preprocessing**

We combined the two data sources and apply preprocessing techniques. The preprocessing dealt mostly with  removing terms/words that are irrelevant for matching.

For the text data obtained from experts we applied a text summarization technique to obtain key phrases and then applied tokenization to get keywords. We saved the keywords and later used them to understand the matching.

For the text data from vacancies we did a number of preprocessing steps to cast them into a form suitable for the application of analytical techniques. In what follows, we describe the steps we have undertaken to prepare and analyse vacancies.

## **Vacancy data cleaning and vacancy corpus construction**

For the data cleaning, we converted all letters to lower case and removed Dutch stopwords. The stopwords are shown next.

|  |
| --- |
| "de"      "en" "van"     "ik" "te" "dat"     "die"  "in"      "een" "hij"     "het" "niet" "zijn"    "is"  "was"     "op" "aan"     "met" "als" "voor"    "had"  "er"      "maar" "om"      "hem" "dan" "zou"     "of"  "wat"     "mijn" "men"     "dit" "zo" "door"    "over"  "ze"      "zich" "bij"     "ook" "tot" "je"      "mij"  "uit"     "der" "daar"    "haar" "naar" "heb"     "hoe"  "heeft"   "hebben" "deze"    "u" "want" "nog"     "zal"  "me"      "zij" "nu"      "ge" "geen" "omdat"   "iets"  "worden"  "toch" "al"      "waren" "veel" "meer"    "doen"  "toen"    "moet" "ben"     "zonder" "kan" "hun"     "dus"  "alles"   "onder" "ja"      "eens" "hier" "wie"     "werd"  "altijd"  "doch" "wordt"   "wezen" "kunnen" "ons"     "zelf"  "tegen"   "na" "reeds"   "wil" "kon" "niets"   "uw"  "iemand"  "geweest" "andere" “hebt” “bent” |

We then removed all numbers and these two strings: “aa” and “âââ”.

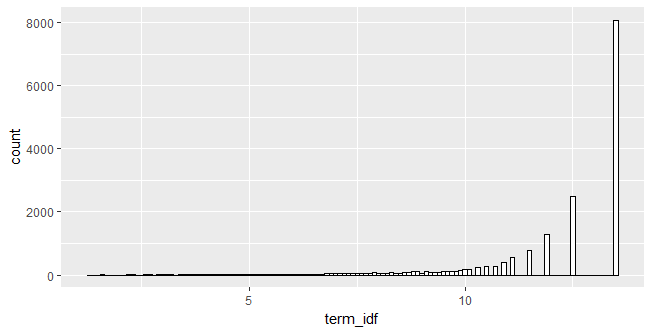
After data cleaning, we build the corpus. The final corpus has 14955 documents (which consisted of candidate descriptions from the vacancies)

# **Text transformation**

We transformed the corpus into a document-by-term matrix (dtm). This is a matrix representation for text data where the columns of this matrix are the unique terms and the rows are the documents. In the process of building the dtm we decided to remove terms having less than 3 letters (or characters).

The frequency of a term indicates its importance in a document. Another weight measure called inverse document frequency (idf) gives an idea about the importance of a term in a corpus and can be used as a measure to assess the discriminatory power of a term. Hence, we compute the idf of each term here.

The figure below shows a histogram of idfs.



We filter out terms with high idf. We set the threshold equal to the median which in this case is 12.47104.

After filtering out terms having idf higher than 12.47104, we are left with 9438 terms. The removal of some terms resulted to some documents now becoming empty. The number of documents remaining is 11347. Hence, we are left with a matrix with dimensions 11,347 by 9438 which we call the final dtm.

# **Analysis**

## **Latent Semantic Analysis**

We applied latent semantic analysis (LSA) to reduce dimensionality and to incorporate the semantics of words in the matching. We used 2 different criteria on how to select the number of dimensions to retain. The next table summarizes the choices we made for running LSA.

|  |  |
| --- | --- |
| Term specific weight | Term frequency |
| Global weight | none |
| Criteria for choosing the number of dimensions to retain | (1) Kaiser and (2) a fraction of the sum of the selected singular values to the sum of all singular values (share = 60%) |

Since we used two different criteria for choosing the number of dimensions we have two LSA models, one from using Kaiser and one from using share, denoted by LSA\_kaiser and LSA\_share respectively in our succeeding discussion. LSA\_kaiser has retained 4748 dimensions while LSA\_share has retained 1091 dimensions. From our experiments including manual examination of the resulting matches, LSA\_kaiser gives the most reasonable performance in the matching. Thus, we used LSA\_kaiser in the succeeding analysis.

# **Matching**

Using the output from LSA models specifically LSA\_kaiser. We subsequently performed the matching on non-teaching vacancies.

## **Matching using LSA\_kaiser**

We matched around 122,081 vacancies containing a mix of teaching and non-teaching vacancies. The matching was done in the following way:

1. Project each vacancy to the constructed LSA dimensions.
2. For each vacancy, compute its match to each of the 11,347 teaching vacancies using cosine similarity. For each vacancy then, we obtained 11,347 computed cosine similarities.
3. Determine the “aggregate” match of a vacancy to each teaching type (math, dutch, and geography) by computing the average of the similarities conditioned on each teaching type. Using the aggregate cosine scores we are now able to determine how likely this vacancy can be filled by teachers of each type. In computing the average score we only take into account score of at least 0.3. Also we take note of the maximum cosine score.
4. Determine the proportion of teaching vacancies in each type which have at least 0.3 cosine score.
5. The aggregate cosine score and the percentage of matching vacancies are subsequently used to assess the match in the succeeding analysis. The higher the aggregate cosine score and percentage of match the better the match.

Note that the match should be interpreted symmetrically, that is, matching teaching and non-teaching jobs means that job holders from the non-teaching jobs could potentially teach and teachers could potentially satisfy the requirements and carry out the responsibilities of the  non-teaching jobs.

# **Results**

The computed matching scores (cosine and percentage) will help us address the following points:

1. Top matching function classes
2. Top matching function roles
3. Sample job titles
4. Optimal setting for the measures (aggregate cosine score, percentage of match, and maximum cosine)

In this section we present the following information for each teaching type.

To help visualize the results we created an app for this purpose. The url of the app is: