Hybrid Jobs: Matching nonteaching and teaching jobs

Prepared by: Vladimer Kobayashi

Last Updated: July 7, 2017

# Summary of Teaching Jobs

|  |  |  |
| --- | --- | --- |
| Teacher type | Count | Proportion |
| Aardrijkskunde | 1468 | 0.10 |
| Nederlands | 7746 | 0.52 |
| Wiskunde | 5742 | 0.38 |
| **Total** | **14956** | **1.00** |

## 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 cleaning and corpus construction

We deal only with text from candidate\_descr column. 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).

## Some dtm statistics

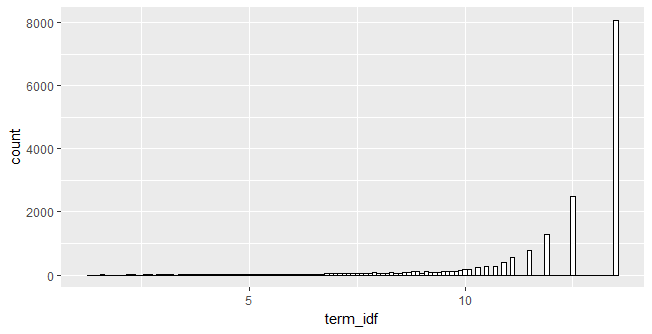
|  |  |
| --- | --- |
| Number of rows | 14952 |
| Number of columns (vocabulary size) | 17525 |
| Non sparse entries | 378025 (0.001) |
| Sparsity | Almost 100% |
| Most frequent terms (at least 2000) | "beschikt", "bevoegdheid","docent","ervaring", "graads","hbo","kennis","leerlingen",  "nederlands","onderwijs","opleiding","vaardigheden",  "wij", "wiskunde" |

## Summary of frequencies

|  |  |
| --- | --- |
| Minimum | 1 |
| 1st Quantile | 1 |
| Median | 2 |
| Mean | 24.58 |
| 3rd quantile | 6 |
| Maximum | 7431 |

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 it 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 11347 by 9438 which we call the final dtm.

# Analysis

## Topic Modeling

We applied topic modeling to the final dtm with the following parameters.

|  |  |
| --- | --- |
| Method | Latent Dirichlet Allocation |
| Number of topics | 100 |
| Estimation | Gibbs sampling |
| **Parameters for Gibbs sampling** | |
| Burn in | 1000 |
| Thinning | 100 |
| Number of runs | 1000 |

## Latent Semantic Analysis

We also applied latent semantic analysis (LSA). 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.

# Matching

Using the output from LDA and LSA models. We performed the matching on non-teaching vacancies.

## Matching using LSA\_share

Initially we matched around 122,081 vacancies containing a mix of teaching and non-teaching vacancies. The matching was done using the cosine similarity. We considered it a match if the computed cosine is at least 0.30.

## Matching using LSA\_kaiser

The matching was done using the cosine similarity. We considered it a match if the computed cosine is at least 0.30.

## Matching using LDA

The matching was done using the Shannon-Jensen divergence.

# Results

## Matching using LSA\_kaiser – average cosine score

### Aardrijkskunde

Summary of matches based on **average cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Function class** | **education** | **Function role** | **Count** | **Frequency** |
| Vervoer en transport | MBO | postbode | 467 | 0.70 |
| Overig | HBO | stage / afstudeerplaats / bbl | 373 | 0.47 |
| Verzekeringen en financiële | HBO | assistent-accountant | 233 | 0.32 |
| Gezondheidszorg en welzijn | MBO | verzorgende | 210 | 0.23 |
| Installatie, reparatie en onderhoud | MBO | Reparatie- en onderhoudsmonteur | 206 | 0.21 |
| Overig | MBO | stage / afstudeerplaats / bbl | 193 | 0.44 |
| Verkoop en handel | HBO | accountmanager overig | 175 | 0.16 |
| Administratie en klantenservice | MBO | medewerker klantenservice | 168 | 0.28 |
| Overig | HBO/WO | stage / afstudeerplaats / bbl | 165 | 0.46 |
| Gezondheidszorg en welzijn | MBO | vrijwilliger | 150 | 0.17 |

### Nederlands

Summary of matches based on **average cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Functional class** | **Education** | **Function role** | **Count** | **Frequency** |
| Overig | HBO | stage / afstudeerplaats / bbl | 817 | 0.54 |
| Administratie en klantenservice | MBO | medewerker klantenservice | 651 | 0.29 |
| Installatie, reparatie en onderhoud | MBO | reparatie- en onderhoudsmonteur | 578 | 0.21 |
| Overig | MBO | stage / afstudeerplaats / bbl | 440 | 0.41 |
| Informatie- en communicatietechnologie | HBO | software engineer | 422 | 0.11 |
| Gezondheidszorg en welzijn | MBO | verzorgende | 415 | 0.33 |
| Overig | Onbeken | onbekend | 407 | 0.51 |
| Verkoop en handel | MBO | medewerker callcenter | 404 | 0.18 |
| Administratie en klantenservice | HBO | medewerker klantenservice | 391 | 0.20 |
| Overig | HBO/WO | stage / afstudeerplaats / bbl | 343 | 0.56 |

### Wiskunde

Summary of matches based on **average cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Functional class** | **Education** | **Function role** | **Count** | **Frequency** |
| Overig | Onbekend | onbekend | 93 | 0.51 |
| Overig | HBO | stage / afstudeerplaats / bbl | 78 | 0.62 |
| Informatie- en communicatietechnologie | HBO/WO | software engineer | 74 | 0.12 |
| Informatie- en communicatietechnologie | HBO/WO | java ontwikkelaar | 73 | 0.12 |
| Gezondheidszorg en welzijn | MBO | vrijwilliger | 48 | 0.22 |
| Gezondheidszorg en welzijn | MBO | verzorgende | 47 | 0.22 |
| Overig | Onbekend | stage / afstudeerplaats / bbl | 47 | 0.26 |
| Administratie en klantenservice | MBO | medewerker klantenservice | 46 | 0.39 |
| Gezondheidszorg en welzijn | MBO | kinderverzorgende | 40 | 0.19 |
| Overig | Onbekend | algemeen medewerker | 40 | 0.22 |

## Matching using LSA\_kaiser – maximum cosine score

### Aardrijkskunde

Summary of matches based on **maximum cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Functional class** | **Education** | **Function role** | **Count** | **Frequency** |
| Overig | HBO | stage / afstudeerplaats / bbl | 63 | 0.52 |
| Verzekeringen en financiële | HBO | assistent-accountant | 40 | 0.40 |
| Administratie en klantenservice | MBO | medewerker klantenservice | 28 | 0.35 |
| Overig | HBO/WO | stage / afstudeerplaats / bbl | 26 | 0.50 |
| Inkoop en magazijnbeheer | VMBO | vakkenvuller | 23 | 0.77 |
| Installatie, reparatie en onderhoud | MBO | Reparatie- en onderhoudsmonteur | 23 | 0.24 |
| Overig | Onbekend | algemeen medewerker | 23 | 0.59 |
| Overig | HBO | onbekend | 22 | 0.18 |
| Verkoop en handel | HBO | accountmanager overig | 21 | 0.12 |
| Verkoop en handel | MBO | medewerker callcenter | 21 | 0.29 |

### Nederlands

Summary of matches based on **maximum cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Functional class** | **Education** | **Function role** | **Count** | **Frequency** |
| Overig | HBO | stage / afstudeerplaats / bbl | 1076 | 0.54 |
| Administratie en klantenservice | MBO | medewerker klantenservice | 779 | 0.29 |
| Installatie, reparatie en onderhoud | MBO | Reparatie- en onderhoudsmonteur | 685 | 0.21 |
| Overig | MBO | stage / afstudeerplaats / bbl | 587 | 0.42 |
| Gezondheidszorg en welzijn | MBO | verzorgende | 547 | 0.30 |
| Vervoer en transport | MBO | postbode | 513 | 0.46 |
| Overig | Onbekend | onbekend | 509 | 0.48 |
| Verkoop en handel | MBO | medewerker callcenter | 468 | 0.17 |
| Informatie- en communicatietechnologie | HBO | software engineer | 467 | 0.11 |
| Administratie en klantenservice | HBO | medewerker klantenservic | 438 | 0.19 |

## Wiskunde

Summary of matches based on **maximum cosine score** per function class, education, and function role.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Functional class** | **Education** | **Function role** | **Count** | **Frequency** |
| Informatie- en communicatietechnologie | HBO/WO | java ontwikkelaar | 172 | 0.16 |
| Overig | HBO | stage / afstudeerplaats / bbl | 129 | 0.41 |
| Gezondheidszorg en welzijn | MBO | verzorgende | 116 | 0.24 |
| Informatie- en communicatietechnologie | HBO/WO | software engineer | 111 | 0.10 |
| Verzekeringen en financiële | HBO | assistent-accountant | 109 | 0.32 |
| Overig | Onbekend | onbekend | 107 | 0.49 |
| Installatie, reparatie en onderhoud | MBO | Reparatie- en onderhoudsmonteur | 103 | 0.17 |
| Installatie, reparatie en onderhoud | MBO | monteur elektrotechnische apparatuur | 102 | 0.17 |
| Verkoop en handel | HBO | accountmanager overig | 97 | 0.18 |
| Gezondheidszorg en welzijn | MBO | vrijwilliger | 84 | 0.17 |

# Visualization based on the highest average cosine score

## Aardrijkskunde

## Nederlands

## Wiskund

# Job Explorer

The job explorer app allows the user to explore matching jobs based on percentages of matches, average and maximum cosine scores, job title, function role, and education level. Additionally, it shows information about job location, function type, and profession class.