Mapping the Skill Sets and Knowledge Domains of Data Science and Artificial Intelligence Program Graduates in the Netherlands

Mol, Mathijs

Mapping the Skill Sets and Knowledge Domains of Data Science and Artificial Intelligence Program Graduates in the Netherlands

Mathijs Jonathan Mol

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: 07-06-2022

Student number: s1520776

Supervisor: Dr. Zsuzsa Bakk (internal)

Abstract

Despite the growing popularity, no clear general definition of data science and artificial intelligence has been established. People are often left into the unknown when it comes to the specific definition of these fields. In this study, the first step towards defining these fields is made. Three text analyses models were used to extract the general topics from various data science or artificial intelligence related program or course descriptions. These topics were used to be able to get a grasp on what skill sets are taught to data science and artificial intelligence students. Afterwards, an analysis of posterior classification of the topics per university was performed to explore the differences and similarities between the universities on their orientation of data science and artificial intelligence programs. General and specific skill sets are uncovered and differences between the universities are described in this paper. The results of this paper might be insightful for institutes that have no clear view whether their vacancies might be fit for data science or artificial intelligence graduates.

Content

1 Introduction

The field of data science is continually growing. The total number of university programs in this field has grown substantially and governments try to stimulate the implementation of data science and statistics in various industries. Despite the growing popularity, no clear general definition of data science has been established. People are often left into the unknown when it comes to the specific definition of this field. This definition often consists of a mixture of the definitions of data science, artificial intelligence and computer science.

Problems arise when employers face difficulties because of the unclear definition. This results in discrepancies between expectation and reality when data science graduates are hired for solving specific problems.

To get a good grasp on what a complicated concept precisely entails, a knowledge map can be constructed. For example, Zins (2007) has constructed a knowledge map of the field of information science. This map is defined by having 10 basic categories, namely Foundations, (2) Resources, (3) Knowledge Workers, (4) Contents, (5) Applications, (6) Operations and Processes, (7) Technologies, (8) Environments, (9) Organizations, and (10) Users. Having knowledge over these categories gives a clear overview of what the concept means and what is important when trying to get a clear and complete understanding on the topic.

After some extensive literature review, it became apparent that no such thing had been investigated for the field of data science or artificial intelligence. A wide variety of search keys have been thoroughly investigated (Appendix A) with no relevant result. Thus, scientific theory about the knowledge map of data science and artificial intelligence is absent. Nonetheless, applied studies have shown some results. For example, Markow et al. (2017) defines key skills for data scientists within the Data Science and Analytics framework. Skills include Machine learning, Python, R and Apache Hadoop. Though, a precise framework is

not given. Furthermore, Sigelman et al. (2019) defines hybrid jobs as more complex in the sense that they require a wider variety of skill sets from different fields and states the increasing importance of these jobs. However, it does not give a clear definition of any of these hybrid jobs. It does state however that the demand of data scientists on the job market has increased by 663% from 2013 to 2018 and the demand of marketing data analysts has increased with 194% in the same timeframe.

Applied papers have been proven to be useful to be taken into consideration. In *Niet-routinematige vaardigheden in hbo-profielen* (Allen et al., 2021), the Dutch Maastricht University investigated if target non-routine skill sets were taught in certain programs by analyzing the concerning HBO profile descriptions using text mining techniques. This paper concluded that none to very few terms within the profile descriptions were impossible to group with the target non-routine skill sets. This offered insight into the teaching process of the entire program.

As a result of the absence of the scientific literature, this paper will take the first scientific step into making a knowledge map of what skill sets define the field of data.

In order to approach the complete definition of a field of science, one should look at all the possibilities and requirements within this field. In the case of data science these possibilities and requirements would for example be statistical tests, programming in different languages and being able to perform statistical inference. However, because it is not possible to state every single aspect of a field of science, we looked at a narrower approach. Since the definition of a field of science is a collection of all the knowledge and skill sets people working within this area have, we decided to gather information from the very base of knowledge: the educational system. We decided specifically to investigate what knowledge is taught within Dutch university master programs that involve either “Data Science” or “Artificial Intelligence”. One way of tackling this problem is by scraping the officially

published program and course descriptions from the university websites to extract information from these descriptions. Extracting this information can be done using a technique called "Topic Modelling" (Blei & Lafferty, 2009). In machine learning and natural language processing, a topic model is defined as a type of statistical model that derives the "topics" that occur in a collection of documents. This collection is referred to as a corpus. Topic modeling is an often-used text-mining tool to uncover hidden semantic structures within a text body. Various types of text mining exist. In this paper, we will focus on three different models, namely, Latent Semantic Analysis (Deerwester et al., 1990), Latent Dirichlet Allocation (Blei et al., 2003) and Correlated Topic Models (Blei & Lafferty, 2007). Latent Semantic Analysis (LSA) derives the latent semantical structures from a given set of documents to create topics. Latent Dirichlet Allocation (LDA) posits that each document is a mixture of topics and that each word's presence in a document is attributable to the document's topics. Correlated Topic Models (CTM) is an extension to the LDA. It allows for topics to correlate and therefore topics can cluster together. To expand on this concept, a classification of the outcome topics per university for the LDA and CTM will also be performed. Since descriptions from different universities will be used, we believe it might be insightful to see if there are major differences between universities in defining their data science- and artificial intelligence courses. An elaborate explanation about the previous mentioned forms of topic modelling will be given in the method section.

The goal of this study is to map the skill sets/knowledge domains that data science programs cover, to be able to make a distinction between core skill sets within the data science master programs and the program-specific skill sets that differ per field within data science (i.e., biology, computer science, etc.). Therefore the first research question is: "What are the skill sets within data science related master programs throughout the Netherlands and which of these skill sets are subject specific (i.e., biology, computer science, etc.)? The

second research question is: “What are the differences and similarities in the outcome topics of the LDA and CTM between the investigated universities?”.

2 Methods

In order to answer the research questions, information is gathered about the content of data science and artificial intelligence master programs throughout the Netherlands. This information is gathered in the form of descriptions of programs and courses. Thereafter, three different forms of text analysis will be performed, namely LSA, LDA and CTM. Lastly, the posteriors of the LDA and CTM have been used to calculate the inter-university differences.

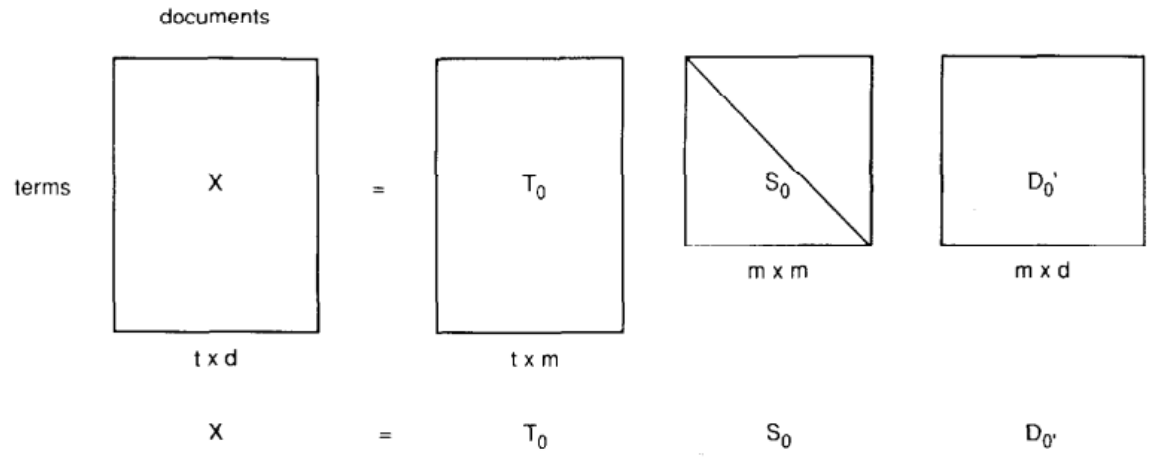
2.1 Topic models

In order to retrieve information from the descriptions, three different methods of text analysis have been chosen. Each of the three analysis strategies offers a different important insight to the research question and complement each other in different ways. A more elaborate motivation for the choice of these three analyses are given in paragraph 2.1.4.

2.1.1 Latent Semantic Analysis

The first method selected for this paper is the LSA (Deerwester et al., 1990). LSA is a widely used natural language processing technique that analyses relationships between a set of documents and their terms by producing a set of concepts relating to these documents and terms. This model assumes that words with a corresponding or close meaning will occur in similar pieces of text. LSA uses a term frequency-inverted document frequency (tf-idf) matrix as input. In a tf-idf matrix the normal term frequency matrix values are weighted

proportionally to the number of times the terms appear in each document. This upweights rare terms to reflect their relative importance and to downweigh terms that occur a lot in every document. The number of rows of this tf-idf matrix will then be reduced while preserving the similarity structure among columns by using a mathematical technique called singular value decomposition (SVD). The SVD model is shown in Figure 1.



Where T_0 has orthogonal, unit-length columns ($T_0'T_0 = I$)
D_0 has orthogonal, unit-length columns ($D_0'D_0 = I$)
S_0 is the diagonal matrix of singular values
t is the number of rows of X
d is the number of columns of X
m is the rank of X (< min(t, d))

Figure 1. Schematic representation of the (SVD) of a rectangular term by document matrix. The original matrix is decomposed into three matrices each with linearly independent components. From *Indexing by latent semantic analysis,* by Deerwester, S., Dumais, S.T., Furnas, G.W., Landbauer, T.K., & Harshman, R. (1990). Journal of the American Society for Information Science, 41(6), 391-407.

The SVD will divide matrix X into the product of three other matrixes. Resulting in the following formula.

$$X = T_0 S_0 D_0',$$

Since the singular values in matrix S_0 are ordered by size, the first k largest values may be kept and the remaining smaller values can be set to 0 in order to achieve optimal approximate

fit for a smaller matrix. The product of the resulting matrices will be a matrix $\hat{X}$ which is approximately equal to X with rank k. Since zeros were introduced in S_0, the representation can be simplified by deleting the rows and columns that contain 0 values, to obtain a new diagonal matrix S. The corresponding columns of T_0 and D_0 will be removed as well to obtain matrices T and D respectively. This results in the new following equation.

$$X \approx \hat{X} = TSD'$$

Documents will thereafter be compared by taking the cosine of the angle between the two vectors formed by any two columns of the matrix. This results in very similar documents being scored with values close to 1 and very dissimilar documents being scored with values close to 0.

In order to run the LSA, the "textmineR" package has been used. (Jones, Doane & Attborn, 2021).

2.1.2 Latent Dirichlet Allocation

The second method selected for this paper is Latent Dirichlet Allocation or LDA (Blei et al., 2003). This method is one of the most principal approaches in topic modelling. In LDA, the latent semantical context within each document is derived to reveal the statistical structure across all document. To explain the LDA model, the following terms are formally defined as defined in Blei et al. (2007).

- **Words and documents.** *Words* are considered the only observable random variables organized into *documents*. Where $w_{d,n}$ denotes the n^{th} word in the d^{th} document, which

is an element in a *V*-term vocabulary ($\{1,\ldots,V\}$). $\mathbf{w}_d$ denotes the vector of n_d word associated with document *d*.

- **A corpus.** A corpus is a collection of M documents denoted by $D = \{\mathbf{w}1, \mathbf{w}2, \ldots, \mathbf{w}_M\}$.
- **Topics.** A *topic* $\boldsymbol{\beta}$ is a distribution of the vocabulary. This can be described as a point on the $V - 1$ simplex. The total number of topics is denoted as *K* and therefore the total number of topics in the model is denoted as $\boldsymbol{\beta}_{1:K}$.
- **Topic assignments.** Each word is assumed to be originated from one of the *K* topics. Therefore the topic assignment of the N^{th} word and the d^{th} document is denoted by $z_{d,n}$.
- **Topic proportions.** Each document has a set of topic proportions, denoted by θ_d. This is a point on the $K - 1$ simplex. Thus θ_d is a distribution of topic indices which reflects the probabilities with which the words are drawn from each topic in document *d*. A natural parameterization of this multinominal is typically considered as $\boldsymbol{\eta} = \log(\theta_i / \theta_k)$.

LDA is a generative probabilistic model of a corpus. In this model, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. In LDA, the following generative process for each document **w** in corpus *D* is assumed.

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the *N* words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

The LDA model estimates each topic z as a mixture of words. This is represented in the Beta matrix, composed of k topics x V terms. In mathematical terms, this matrix is comprised of i topics containing j words, causing the matrix to be defined as $\beta_{ij} = p\ (w_j = 1 \mid z_i = 1)$. The gamma matrix contains the document-topic probabilities. These probabilities represent the proportion of words from some document **w** that are generated by topic *z*. The mathematical formula for the gamma matrix is as follows

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n\,|z_n, \beta)$$

In Figure 2 this model can be represented as a probabilistic graphical model. The figure presents the model as divided into three levels. The parameters α and β are corpus-level parameters. They are assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document. A classical clustering model would only have two levels in which a Dirichlet would be sampled

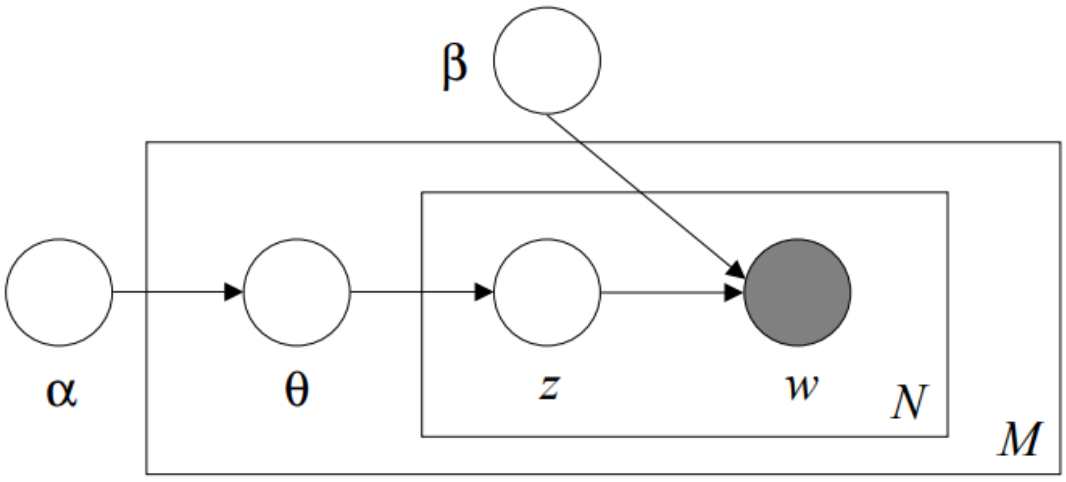


Figure 2 Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. From Latent Dirichlet Allocation, by Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). *Journal of Machine Learning Research, 3*, 993–1022.

only once for the corpus, a multinomial clustering variable only once per document and a set of words would be selected conditional on the cluster variable. This restricts a document of being associated with only one topic. Since the LDA model involves three levels and the topic node is sampled repeatedly within the document, documents can be associated with multiple topics.

In order to run the LDA, the R-package "topicmodels" has been used (Grün & Hornik, 2011). This package uses a term frequency matrix as input. A TF matrix is defined as a **w** x *w* with counts in every cell.

2.1.3 Correlated topic models

The last method selected is the Correlated Topic Model or CTM (Blei & Lafferty, 2007). The CTM is a hierarchical model of document collections. The CTM models the words of each document from a mixture model. The components of this mixture model are shared by all documents in the collection, therefore the mixture proportions are document specific random variables. The CTM allows for multiple topics with different proportions for each document. Thus, it allows to capture the heterogeneity in grouped data that show multiple latent patterns. To describe the data, latent variables and parameters within the CTM, the same terminology and notations have been used as within the LDA, as described previously in section 2.1.2.

Given topics $\boldsymbol{\beta}_{1:K}$, a K-vector $\boldsymbol{\mu}$ and a K x K covariance matrix $\boldsymbol{\Sigma}$, the CTM assumes that an N-word document is generated by the following process.

1. Draw $\eta_d \,|\{ \boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. For $n \in \{1, ..., N_d\}$:

(a) Draw topic assignment $z_{d,n}|\boldsymbol{\eta}_d$ from Mult($f(\boldsymbol{\eta}_d)$).

(b) Draw word $w_{d,n}|\{z_{d,n}, \boldsymbol{\beta}_{1:K}\}$ from Mult($\boldsymbol{\beta}_{Zd,n}$),

Where $f(\boldsymbol{\eta})$ maps a natural parameterization of topic proportions to the mean parameterization, described mathematically in the next equation.

$$\theta = f(\boldsymbol{\eta}) = \frac{\exp\{\boldsymbol{\eta}\}}{\sum_i \exp\{\boldsymbol{\eta}_i\}}$$

This formula can be illustrated in a probabilistic graphical model, displayed in Figure 3.
A probabilistic graphical model is defined as graphical representation of a collection of joint distributions with nodes denoting the random variables and the edges denoting possible dependencies between the random variables.

The CTM is an expansion of the LDA model. As comes forth from the previous model descriptions, the LDA assumes a nearly identical generative process, but the topic proportions are drawn from a Dirichlet. This Dirichlet is a computationally convenient distribution over topic proportions, for it is conjugate to the set of topic assignments. However, the Dirichlet assumes near independence of the individual proportions. This means one could simulate a draw from a Dirichlet by drawing from K independent Gamma distributions and normalizing the resulting vector. In contrast to the LDA, the CTM does not use a Dirichlet, but draws a real valued random vector from a multivariate Gaussian distribution and then maps it to the

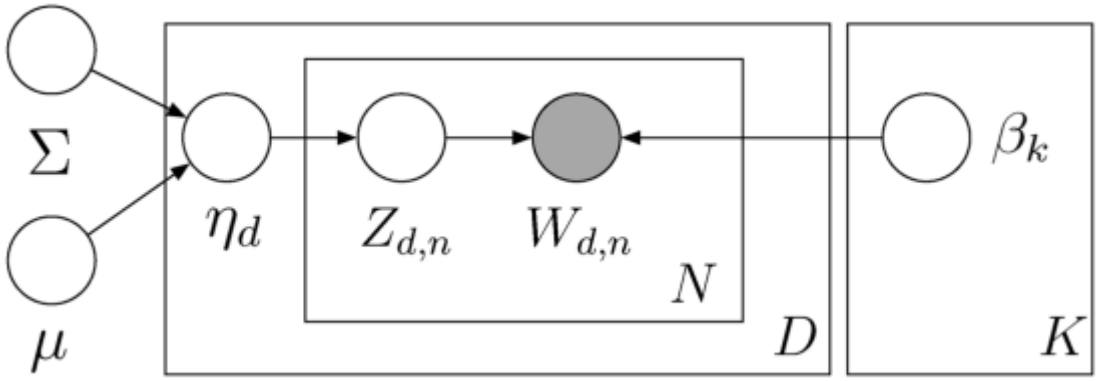


Figure 4. Probabilistic graphical model representation of the CTM. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a LDA model. From A Correlated Topic Model of Science, by Blei D.M., Lafferty J.D. (2007). *The Annals of Applied Statistics, 1*(1), 17–35.

simplex to obtain a multinomial parameter. This is the defining characteristic of the logistic Normal distribution (Aitchison, 1982; Aitchison, 1985; Aitchison & Shen, 1980). This distribution is used to model the latent composition of topics associated with each document. This means correlations and covariances are calculated between all K topics. A graphical representation of this is shown in Figure 4. With these correlations and covariances a more clustered construction of the topics is formed.

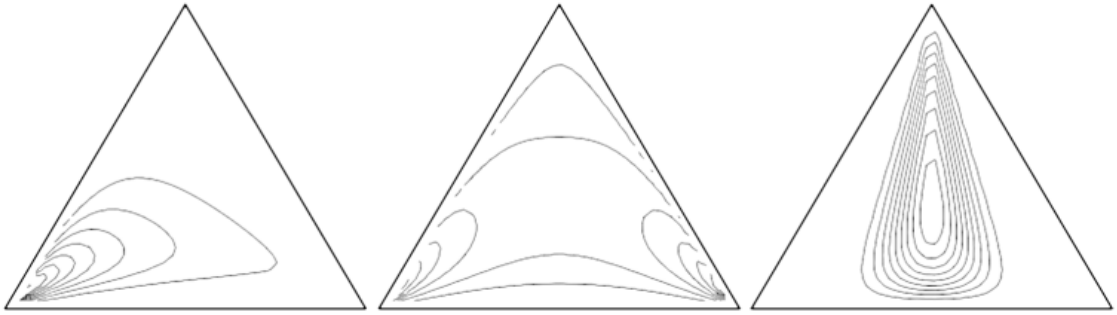


Figure 3. Example densities of the logistic normal on the 2-simplex. From left to right: the diagonal covariance and nonzero-mean, the negative correlation between topic 1 and 2 and lastly the positive correlation between topic 1 and 2. From A Correlated Topic Model of Science, by Blei D.M., Lafferty J.D. (2007). *The Annals of Applied Statistics, 1*(1), 17–35.

For the CTM, the R-package “topicmodels” was used (Grün & Hornik, 2011). This package uses a term frequency matrix as input as well. A TF matrix is defined as a **w** x *w* with counts in every cell.

2.1.4 Choice of analysis

To answer the research question, the three methods discussed in the previous paragraphs have been selected. Firstly, the LSA has been selected because this analysis gives more insight in the latent semantical structure of the texts. This means it takes into account if words are more often used in proximity to other words. It includes this information to create clusters of words and favors strong combinations of words (meaning if the word “data” is used a lot and “science” is used less, but almost always in combination with the word “data”, the model would be more likely to include the word “science” into the topic in which the word “data” is highly present). In conclusion, this analysis focusses more on the words within the topics and gives us a better insight in the content of the topics.

Secondly, the LDA has been selected because it offers a statistically strong overview of the underlying topics of the descriptions. General patterns of the content of the programs are expected to be uncovered by this analysis. The focus in this analysis is mainly on the difference between topics. The results of this analysis will show what the order from most to least present topic is in the corpus and what words are most common in these topics. For example, when considered a fictional LDA model where K = 3, topic 1, 2 and 3 could be defined by the most common words being “data”, “statistical_learning” and “artificial_intelligence”, topic 2 by “regression”, “model” and “techniques” and topic 3 by “neuro_network”, “quantum” and “biological”. The interpretation of this fictional result might be that the main subject within the corpus is revolved around data science, whereas the second topic indicates that a part of the programs is revolved around the statistical part of data

science. The third topic might suggest that a biology subdomain within the data science programs is present. Important to note is that the LDA allows the same word to be present in different topics. Therefore, considering the example, the word “model” might be present in both topic 1 and 2.

Lastly, the CTM has been selected because this technique allows topics to correlate and therefore topics can be clustered together. When considering an example where the previous example is expanded upon with two more topics, which are defined by the most common words for topic 4 being “management”, “business” and “innovation” and topic 5 being “marketing”, “financial and “market”, the CTM would be highly likely to identify these different topics as a cluster and be more prone to identify them as topics. This will cause the probabilities for the words in these topics to increase and therefore general clusters will be more likely to be uncovered.

2.2 Selecting the optimal number of K topics

The previously mentioned text analysis methods all require a predefined number of k topics in order for the analysis to be able to run (Deerwester et al., 1990; Blei et al., 2003; Blei & Lafferty, 2007). Having a number of topics that is too small might lead to the analyses not capturing all facets of the semantical contexts. However, if k is too large, interpretability might be lost and topics might not be as coherent to the human reader. There are multiple ways of determining what the optimal number of K topics is for topic models. The most common ways are by calculating the perplexity (Newman at al., 2009) for the LDA and the CTM and the coherence (O’Callaghan et al., 2015; Mimno et al., 2011) for all three analyses. Perplexity is a widely used evaluation metric for language model evaluation. It is a statistical measure that reflects how well a probability model reflects a sample. Perplexity is algebraically equivalent to the inverse of the geometric mean per-word likelihood, where a

smaller perplexity value reflects better generalization performance. In mathematical notations, the perplexity of a test set D_{test} for M documents is defined as the following.

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right)$$

where N_d is the length of document *d*, $p(d_d)$ is the probability of document *d*, generated by the model.

Coherence first measures how semantically similar high-scoring words within a topic are. These scores are later combined to form a general coherence measure. Coherence therefore helps to distinguish between semantically interpretable topics and topics that are formed as a result of the statistical inference within the text analysis. This is illustrated in the following formula to calculate topic coherence.

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D\left(v_m^{(t)}, v_l^{(t)}\right)+1}{D\left(v_l^{(t)}\right)}$$

where $D(v)$ is the document frequency of word type *v* (i.e. the number of documents with at least one token of type *v*). $D(v, v')$ is defined as the co-document frequency of word types *v* and *v'* (i.e. the number of documents containing at least one token of type *v* and *v'*). $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is defined as a list of the *M* words with the highest topic-specific "collapsed" probabilities. Lastly, the addition of 1 is introduced to avoid taking the logarithm of zero.

Four other measures have also been looked at to determine what the optimal number of K topics is for LDA and CTM. The first method investigated a measure that required to split the corpus into two matrix factors. The measure was then computed in terms of symmetric KL-Divergence of salient distributions that are derived from these matrix factors (Rajkumar Arun et al., 2010). The second method was a density-based method where the

LDA performed best where the average cosine distance of topics reached a minimum (Juan et al., 2009). The third method considered a non-symmetric measure, namely the Jensen-Shannon divergence, which is a symmetrised version of the KL divergence (Deveaud et al., 2014). The last method used a Markov chain Monte Carlo algorithm for inference to extract an optimal number of topics for LDA (Griffiths & Steyvers, 2004). All four methods selected a similar number of topics as compared to the perplexity measure. However, when implemented into this paper, the four techniques resulted in selecting different numbers of K topics and thus were not in line with either the perplexity measure and one another. Therefore, we decided not to use any of these methods and chose to rely solely on perplexity and coherence.

2.3 Posterior classification

After the main analyses, a classification of the posteriors to the different universities will be performed on the outcome topics of the LDA and CTM. This will be done by calculating the posterior values for every description per topic and assigning them to the corresponding university. The mean of the posteriors for every description per topic per university will be calculated to get an overview of how well a topic fits a university's descriptions for both LDA and CTM separately (as can been seen in Table 7 and 8). Since calculating posteriors is not applicable for LSA, only the results of the LDA and CTM will be used for the clustering.

2.4 Data collection

The total dataset consists of 1009 descriptions that have been manually collected from university websites. These universities all are part of 'De Vereniging van Universiteiten' (VSNU), which is a trade group of ten government-funded universities, three special

universities and an open university. We chose to only use programs of universities that are part of this association for reliability reasons. In Table 1 a list is shown of the universities together with the number and type of description per university. For one university, no type of course was stated at the official university website. Therefore a column has been added to the table containing course descriptions that could either be core courses or electives.

Table 1

Number of descriptions per university

	Program	Core course	Elective	Core or elective
Delft University of Technology	1	10	66	0
Eindhoven University of Technology	5	32	45	0
Leiden University	3	33	68	0
Maastricht University	5	26	76	0
Radboud University Nijmegen	2	5	13	45
Tilburg University	5	39	64	0
University of Groningen	2	22	29	0
University of Twente	6	34	87	0
University van Amsterdam	5	49	37	0
Utrecht University	2	6	49	0
Vrije Universiteit	4	24	67	0
Combination: Vrije Universiteit, Erasmus University of Rotterdam and Universiteit van Amsterdam[a]	1	20	22	0

Note. The number of program descriptions equals the number of programs selected for this study

[a]This is one program offered by three different universities and is thus a combination of institutions

2.5 Data processing

After collecting the descriptions, some altercations have been made to the text in order to obtain more optimized results. The specific changes are stated in Table 2. The three text analyses all work based on a count of words. This is defined as “Term Frequency”. Because the context is not taken into account in this method, some combination of words are replaced by the two words added together by an underscore, such as: “statistical_learning” and some

plural forms have been changed into singular form. Words that did not hold any information and numbers in general have been removed and all letters have been set to lower case.

Table 2

Manual changes to the texts

Change	Words
Plural to singular	models, systems, sets, problems, networks, games
Words combined	machine learning, deep learning, statistical learning, data science, computer science, artificial intelligence, data mining, text mining, time series, neural network, research project, distributed computing, natural language processing, probabilistic theory, distributed system, critical thinking, decision making, skill sets, ad hoc
Words deleted	will, course, courses, student, students, able, university, master, can, skills, work, new, use, used, using, also, different, learn, learning, part, master's, understand, one, two, game, topics, understanding, based, many, several, exam, make, discussed, ad hoc

3 Results

3.1 Selecting the optimal number of K topics

In order to select the optimal number of K topics, coherence has been calculated for the LSA, LDA and CTM for all models with K topics ranging from 2 to 20. The same has been done for the perplexity measure for the LDA and CTM, also with K topics ranging from 2 to 20. Models with only one topic are not possible, since this would set the topic equal to the vocabulary.

3.1.1 K topics – LSA

The coherence scores have been calculated for the LSA models with K topics ranging from 2 to 20. The scores are plotted in Figure 5. Since a model with a high coherence score

means higher semantical similarity of high scoring words within each topic, a model with a low amount of topic is preferred. However, since too little topics will give a low amount of information, a K of 4 has been chosen. To investigate if a less parsimonious model would still give coherent and interpretable results, we compared the K = 4 model to a model where K = 7 and decided after which model will be chosen as the main model. This will be described in section 3.2.1.

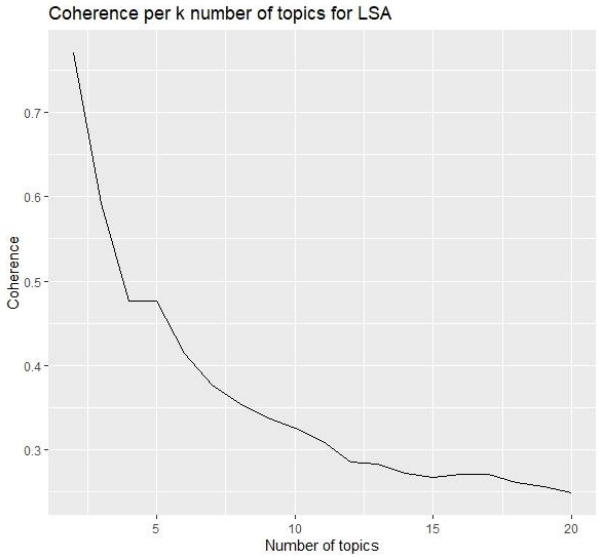


Figure 5. Plot of the coherence scores per number of K topics for the LSA

3.1.2 K topics – LDA

Selecting the optimal number of topics for the LDA has been done by investigating both the coherence and perplexity scores. Figure 6 shows that coherence favours either a model of 13 or 15 topics and perplexity favours a more parsimonious model. However, since a complex model complicates interpretability and since coherence scores are practically equal for LDA models ranging from 7 to 12 topics, the 7 and 13 topic model were chosen to be inspected.

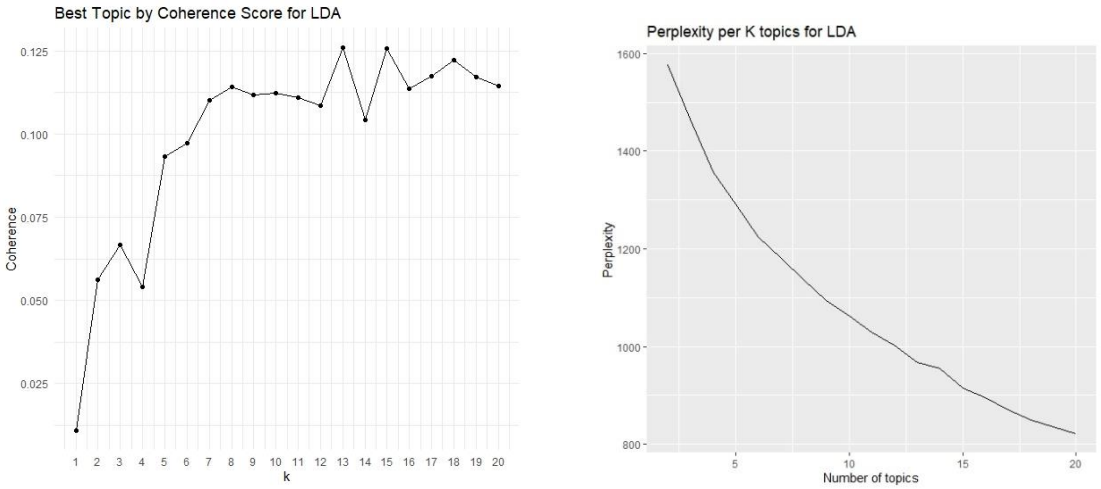


Figure 6. Plot of the coherence scores per number of K topics for the LDA (a) and Plot of the perplexity scores per number of K topics for the LDA (b)

3.1.3 K topics – CTM

The coherence and perplexity scores have also been calculated for the CTM and results have been plotted in Figure 7. Similarly to the interpretation of the selection of the number of K topics for the LDA, the model with 7 and 13 topics have been selected to be investigated. The optimum for the perplexity measure for the CTM also lies at a more complex model. A model of 7 and 13 topics has been chosen to be inspected.

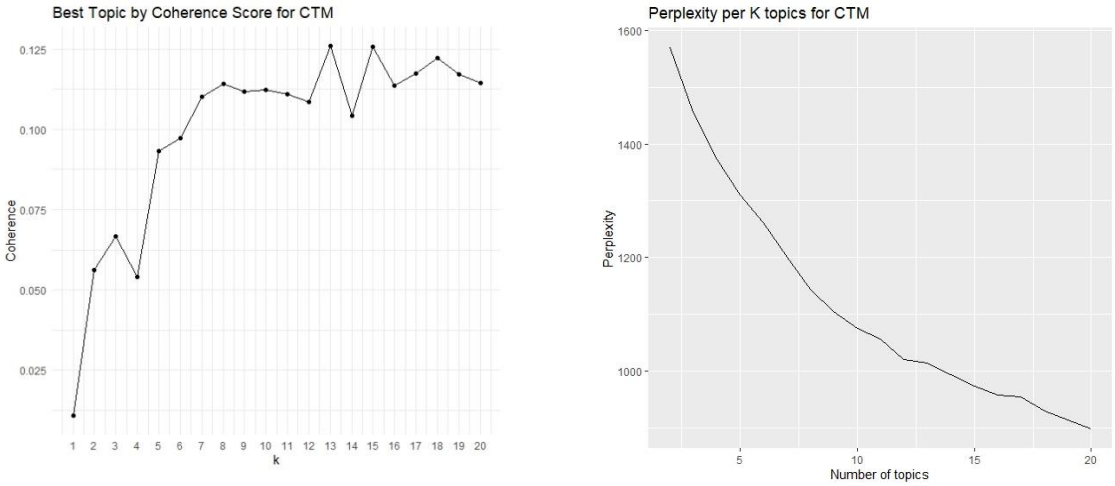


Figure 7. Plot of the coherence scores per number of K topics for the CTM (a) and Plot of the perplexity scores per number of K topics for the CTM (b)

3.2 Topic modelling results

After specifying the parameter K, the topic models have been run for the LSA, LDA and CTM. After running the models, the results were investigated and some extra words were removed after. As stated in the method section, some words have been deleted to optimize the results. A special case of this were the words “ad” and “hoc” (Appendix B). This resulted in the final models, which are discussed below. To summarize the interpretation of the words within the topics, key terms have been manually added to Table 3, 4 and 5 in a row named “Core”.

3.2.1 Results LSA

The results of the LSA are shown in Table 3. Based on the latent semantical contexts, the first three topics mainly give information about how to reach the main objectives in the programs, whereas the fourth topic is partly about deep learning and partly about network search. Overall, the LSA where K = 4 does not offer a very insightful overview of the specific skill sets of data science graduates. Therefore, the LSA where K = 7 has also been investigated.

In the added three topics, some additional domains are revealed (see Table 4). The LSA tells us the 5th through the 7th latent semantical spaces revolve around data mining/processing, subdomains and a topic specifically about language processing and deep learning. The specific subdomains stated in topic 6 of the LSA are Astronomy, Law and Research. This indicates that skill sets of data science graduates include a very versatile mixture of various fields of science. Because this model offers a substantially better overview of the corpus, the LSA model where K = 7 has been chosen as the main model.

Table 3

Top ten most common words per topic for the LSA where K = 4

	t_1	t_2	t_3	t_4
1	aims_contentin	topic	programme	multimedia
2	check_information	topic_teacher	master	multimedia_search
3	contentin_longer	topic_teaches	semester	web
4	goal_objectives	teacher	international	search_recommendation
5	information_check	dm	science	deep
6	information_scheduled	teaches	data_science	model
7	longer_goal	data_topic	thesis	image
8	objectives_information	dpv	maastricht	learning
9	approaches_core	mining	year	search
10	areas_phonetics	process_mining	research	network
Core	*Signal words main objectives*	*Learning*	*Program layout*	*Network search / deep learning*

Table 4

Top ten most common words per topic for the LSA where K = 7

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
1	aims_contentin	topic	programme	multimedia	process_mining	web	image
2	check_information	topic_teacher	master	multimedia_search	event	science	deep
3	contentin_longer	topic_teaches	semester	web	process	data_science	language
4	goal_objectives	teacher	international	search_recommendation	data_science	leiden	processing
5	information_check	dm	science	deep	science	astronomy	language_processing
6	information_scheduled	teaches	data_science	model	mining	law	natural_language
7	longer_goal	data_topic	thesis	image	event_data	web_data	natural
8	objectives_information	dpv	maastricht	learning	model	research	deep_learning
9	approaches_core	mining	year	search	process_model	based_information	network
10	areas_phonetics	process_mining	research	network	customer	information_system	learning
Core	Objectives	Learning	Program layout	Network search / deep learning	Data mining / processing	Subdomains (law, astronomy and research	NLP / deep learning

3.2.2 Results LDA

The results of the LDA where K = 7 are shown in Table 5. Again, we see a more general first topic. We see that most topics are about the most important domains as artificial intelligence, statistics, machine learning, data analyses and deep learning. Lastly, we also see one topic (Topic 2) consisting of words about business, which is a side domain.

When the LDA where K = 13 is investigated, we see that some important concepts within data science are added, such as the process of analysis and deep learning, as well as more subdomains (see Appendix C). This includes the human brain, marketing, research, architecture and security. Subdomains seem to be described in more detail in more complex models. Since the LDA model where K = 7 gives a sufficient overview of the corpus, this model has been chosen as the main model.

Table 5

Top ten most common words per topic for the LDA where K = 7

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	data	business	system	data	data	system	model
2	research	Data	data	model	model	data	network
3	project	research	information	methods	techniques	knowledge	deeplearning
4	topic	datascience	retrieval	analysis	algorithms	design	machinelearning
5	scientific	marketing	human	techniques	analysis	software	neuronetwork
6	thesis	knowledge	multimedia	statistical	image	web	theory
7	projects	innovation	search	machinelearning	problem	information	algorithms
8	datascience	development	artificialintelligence	research	theory	techniques	methods
9	researchproject	management	ethical	naturallanguageprocessing	methods	programming	reinforcement
10	knowledge	analysis	design	language	process	security	system
Core	*Core terms*	*Business/marketing domain*	*AI/information*	*Statistics / Machine learning*	*Data analyzing*	*Information/ IT*	*Deeplearning / machinelearning*

3.2.3 Results CTM

Lastly, the results of the CTM where K = 7 were analysed and the results are shown in Table 6. In CTM, topics are allowed to correlate and can therefore cluster together. We see a large number of subdomains and skills sets being described by the CTM. Apart from the first topic describing core terms, every topic denotes a domain within data science and artificial intelligence, meaning skill sets of this field are being expounded in the descriptions. Only the sixth topic is about ethics which is a less often discussed subdomain.

When compared to the CTM model where K = 13 (see Appendix D), we see there is a lot of overlap with the CTM model where K = 7. Main differences include ethics not being denoted in the K = 13 model and more subdomains being described in the more complex model such as machine learning, natural language processing, AI/optimization, the human brain, information security and the business domain. Since the CTM model where K = 7 gives a sufficient overview of the corpus, this model has been chosen as the main model.

Table 6

Top ten most common words per topic for the CTM where K = 7

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	research	model	data	algorithms	data	system	research
2	project	Data	model	machinelearning	system	artificialintelligence	model
3	datascience	analysis	techniques	data	information	data	social
4	business	techniques	process	techniques	software	health	system
5	data	image	deeplearning	methods	business	concepts	knowledge
6	programme	statistical	naturallanguageprocessing	model	design	datascience	design
7	knowledge	theory	language	knowledge	services	ethical	network
8	scientific	methods	machinelearning	programming	web	decisions	data
9	development	linear	methods	problem	model	privacy	methods
10	thesis	computer	datascience	datamining	distributed	problem	human
Core	*Core terms*	*Statistics*	*Data processing techniques*	*Data processing techniques*	*Electronics-IT*	*Ethics*	*Research*

3.3 Posterior classification

After the main analyses have been run, the classifications have been performed for the LDA and CTM separately. The results are shown in Table 7 and 8.

3.3.1 Posterior classification – LDA

We see in Table 7 that there is quite a lot of differences in order between the general outcome and the individual universities. For example, none of the universities have the “core terms” topic as the most prominent topic. We do see the more technical universities both score high on information/IT and Data analysing. However, they do not differ with non-technical universities, since for example the Universiteit of Twente also scores high on this

topic. Additionally, we can see that the UvA, TU/e, Til and combi[1] score very high on the business/marketing domain. Please take note that these results do not mean that the university as a whole is oriented more on the business or technical domain. It simply states that the data science and artificial intelligence programs in those universities are also focussed on the mentioned side domains.

3.3.2 Posterior classification – CTM

In Table 8 the results are shown. The results seem to differ a lot from the results of the LDA posterior classification. A lot more similarities between universities are present in the CTM topics. Core terms, research and ethics are very prominent topics for example. A small cluster consisting of Rad, Til, RUG and UU which all have the first two same topics, namely research and ethics is also uncovered. However, apart from the first corresponding topics, the order of the topics that come after do differ a lot. Lastly, the two technical universities both score high on electronics/IT, which is the more technical topic.

[1] Combi is defined as one program offered by three universities simultaneously UvA, VU and Erasmus university Rotterdam

Table 7

Order of most important manually named topic terms per university based on posterior values for the LDA

	Uni	First	Second	Third	Fourth	Fifth	Sixth	Seventh
1	TU Delft	Information/IT	Data analysing	Deep learning / machine learning	AI/information	Statistics / machine learning	Business/marketing domain	Core terms
2	TU/e	Business/marketing domain	Data analysing	Information/IT	AI/information	Statistics / machine learning	Deep learning / machine learning	Core terms
3	LU	Data analysing	Core terms	Information/IT	Business/marketing domain	Deep learning / machine learning	Statistics / machine learning	AI/information
4	MU	Statistics / machine learning	Data analysing	Information/IT	Deep learning / machine learning	AI/information	Core terms	Business/marketing domain
5	RAD	AI/information	Statistics / machine learning	Deep learning / machine learning	Business/marketing domain	Information/IT	Core terms	Data analysing
6	Til	Business/marketing domain	Statistics / machine learning	Deep learning / machine learning	Core terms	AI/information	Information/IT	Data analysing
7	RUG	Statistics / machine learning	Information/IT	Business/marketing domain	Core terms	Deep learning / machine learning	Data analysing	AI/information
8	UT	Data analysing	AI/information	Deep learning / machine learning	Business/marketing domain	Statistics / machine learning	Information/IT	Core terms
9	UvA	Business/marketing domain	Statistics / machine learning	Deep learning / machine learning	Data analysing	Core terms	AI/information	Information/IT
10	UU	Statistics / machine learning	Information/IT	Business/marketing domain	Deep learning / machine learning	AI/information	Data analysing	Core terms
11	VU	Information/IT	Statistics / machine learning	Business/marketing domain	Data analysing	Core terms	Deep learning / machine learning	AI/information
12	Combi *	Business/marketing domain	Data analysing	Statistics / machine learning	Deep learning / machine learning	Core terms	Information/IT	AI/information

* Combi is defined as one program offered by three universities simultaneously UvA, VU and Erasmus university Rotterdam

Table 8

Order of most important manually named topic terms per university based on posterior values for the CTM

	Uni	First	Second	Third	Fourth	Fifth	Sixth	Seventh
1	TU Delft	Electronics/IT	Data processing techniques_4	Research	Statistics	Ethics	Data processing techniques_3	Core terms
2	TU/e	Core terms	Electronics/IT	Research	Data processing techniques_4	Ethics	Statistics	Data processing techniques_3
3	LU	Core terms	Data processing techniques_4	Research	Statistics	Electronics/IT	Data processing techniques_3	Ethics
4	MU	Data processing techniques_4	Data processing techniques_3	Ethics	Statistics	Core terms	Research	Electronics/IT
5	RAD	Research	Ethics	Data processing techniques_4	Core terms	Data processing techniques_3	Statistics	Electronics/IT
6	Til	Research	Ethics	Core terms	Statistics	Electronics/IT	Data processing techniques_3	Data processing techniques_4
7	RUG	Research	Ethics	Core terms	Data processing techniques_3	Statistics	Data processing techniques_4	Electronics/IT
8	UT	Data processing techniques_3	Electronics/IT	Statistics	Ethics	Research	Core terms	Data processing techniques_4
9	UvA	Core terms	Research	Ethics	Data processing techniques_4	Statistics	Data processing techniques_3	Electronics/IT
10	UU	Research	Ethics	Statistics	Data processing techniques_4	Core terms	Data processing techniques_3	Electronics/IT
11	VU	Research	Core terms	Data processing techniques_4	Statistics	Data processing techniques_3	Electronics/IT	Ethics
12	Combi *	Statistics	Research	Core terms	Data processing techniques_4	Ethics	Electronics/IT	Data processing techniques_3

* Combi is defined as one program offered by three universities simultaneously UvA, VU and Erasmus university Rotterdam

4 Discussion

In Table 9, a schematic overview of the manually added key terms for every main model per method have been provided to identify the most important skill sets.

Table 9

Manually added key terms describing the main topics per analysis

	LSA	LDA	CTM
1	Signal words main objectives	Core terms	Core terms
2	Learning	Business/marketing domain	Statistics
3	Program layout	AI/information	Data processing techniques
4	Network search / deep learning	Statistics / machine learning	Data processing techniques
5	Data mining / processing	Data analysing	Electronics/IT
6	Law, astronomy and research	Information/IT	Ethics
7	NLP / deep learning	Deep learning / machine learning	Research

4.1 LSA

The LSA calculated the first three topics and parts of the fourth topic to describe very general concepts. The first three topics consist of signal words, words about the teaching process and a description of how the programs are divided. This is in accordance with the expectations, since the LSA derives the latent semantical contexts of a corpus, meaning that it favours words that occur close to each other in a text when topics are calculated. Since signal words for main objectives, learning and program layout are all concepts that occur in every program and are applicable throughout, it is logical that these topics would be favoured the highest by the LSA. Secondly, the LSA selected domains of data science, indicating that skill sets data science graduates will obtain during the data science programs originate from various domains within data science. These domains are shown in Table 10.

4.2 LDA

We see that the LDA calculated the first topic to be describing a general concept as well. Since the LDA focusses on dividing the corpus into different “factors”, this method will try to find topics that do not overlap in meaning. Therefore, only one descriptive first topic is

in line with the expectations of the results of the LDA. The skill sets ought to be taught to data science graduates suggested by the LDA are also shown in Table 10. We can see topics do overlap within the list of skill sets. For example, “AI/information” and “Information/IT” are both topics given by the LDA. Even though it might seem like an overlap, this can be interpreted as a topic about the informational side of artificial intelligence and the informational side of IT. This also applies for “Statistics / Machine learning” and “Deep learning / Machine learning”.

4.3 CTM

For the CTM, it was possible for topics to correlate. The first topic was also a general descriptive topic. Subsequently, the topics again all indicated domains of data science, but differed a lot from the topics suggested by the LDA. We see a small cluster, since topics three and four overlap in describing the same concept. The domains of which the skill sets are taught to data science graduates are shown in Table 10.

Table 10

Indicated skill sets taught by data science programs by LSA, LDA and CTM

LSA	LDA	CTM
Network search / deep learning	Business/marketing domain	Statistics
Data mining / processing	AI/information	Data processing techniques
Law, astronomy and research	Statistics / machine learning	Data processing techniques
NLP / deep learning	Data analysing	Electronics/IT
	Information/IT	Ethics
	Deep learning / machine learning	Research

4.4 General outcomes

When we look at the indicated skill sets in Table 10, we can clearly see that there is a wide variety of skill sets needed to be able to complete a data science master program in the Netherlands. It displays the core concepts of data science, but also the subdomains of which

some knowledge is needed. To clarify this, a distinction has been made between the two and are shown in Table 11.

Table 11

Sorted skill sets based on manually named topics according to LSA, LDA and CTM

Data science specific skill sets	*Skill sets of subdomains*
Network search	Law
Deep learning	Astronomy
Data mining	Research
Natural language processing	Business/marketing
Artificial intelligence	Ethics
Machine learning	Research
Data analysing	
IT	
Machine learning	
Statistics	
Data processing techniques	

Skill sets that correspond to the mentioned fields in Table 11 are taught in data science related master programs throughout the Netherlands. Based on these results, it might be useful for employers to investigate these terms closer to get a better grasp on what data science graduates have to offer, in order to match the skill sets to the concerning vacancies.

4.5 Further differences

A key difference we see between the LSA and the other two methods is that the LSA has three of the seven topics defined by concepts that are generally used in descriptions and do not hold any specific information about the content of the described programs. This might be attributed to the use of TF-IDF by LSA versus the use of TF by LDA and CTM. In TF-IDF, the most and least often used words are given less priority in scoring, meaning that the centre of the distribution will get the highest scores. Since descriptions generally focus solely on the content concept that is to be described, the actual content might have been on the top of the distribution and been given less priority, which might have caused the descriptive terms to

be ranked first in the model. In LDA and CTM, a standard set of stop words are erased and this reordering of scores is not used.

4.6 Posterior classification

The topic order of the posterior classification differs quite a lot from the general topic order for both the LDA and the CTM. At first glance this seems to be a non-expected outcome, however, since the classification is performed by taking the mean of the posteriors of all descriptions per university per topic and the number of programs per university differ, the order of the posterior classifications might differ from the general topic order.

In both Figure 8 and 9, seven plots of the posteriors per university for every topic (LDA and CTM respectively) are shown. These visualisations are plotted to easily identify which universities are similar in scoring high on a topic. Again, note that when universities score high on a topic, it means that their data science and artificial intelligence programs are more oriented towards the described topics. It does not mean that the entire university is oriented towards the concerning domains.

The posteriors of the LDA (see Figure 8) tell us that TU/e, Til, UvA and Combi score similarly high on business/marketing, Til scores exceptionally high on AI/Information, Rug scores high on Statistics / Machine learning, TU Delft and VU score high on Information/IT. We can identify quite a high number of similarities between the universities in these results.

When we look at the posteriors of the CTM (see Figure 9), we see that the combi scores very high on statistics, the scores on data processing techniques 3 and 4 seem to be randomly divided over the universities, TU Delft scores very high on Electronics/It and RUG, RAD and VU seems to be very focussed on research.

The two analyses methods give very different results. When also taking Table 7 and 8 into account, we can see that the LDA topic order of the posterior classification differs a lot from the general model outcomes. This is in line with the expectations, since the LDA assumes independent topics and therefore topics in the general outcome will not be allowed to cluster. The CTM also brings forth a lot more clusters, as was to be expected, since the CTM does not assume independent topics and thus allows topics to correlate.

In general, we can conclude the universities differ a lot in what domains their data science and artificial intelligence programs are more oriented towards. If one would be interested in a specific area within data science or artificial intelligence, it might be beneficial to look into what domain a university would be leaning towards.

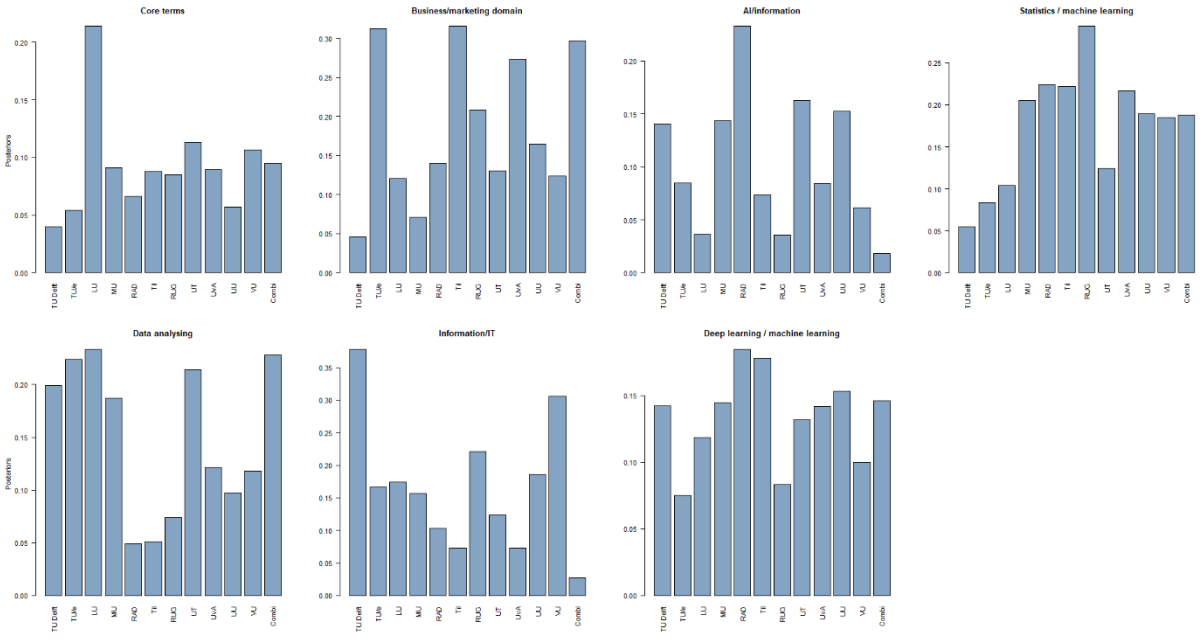


Figure 8. Seven barplots of the posteriors per university for every manually named topic of the LDA outcomes

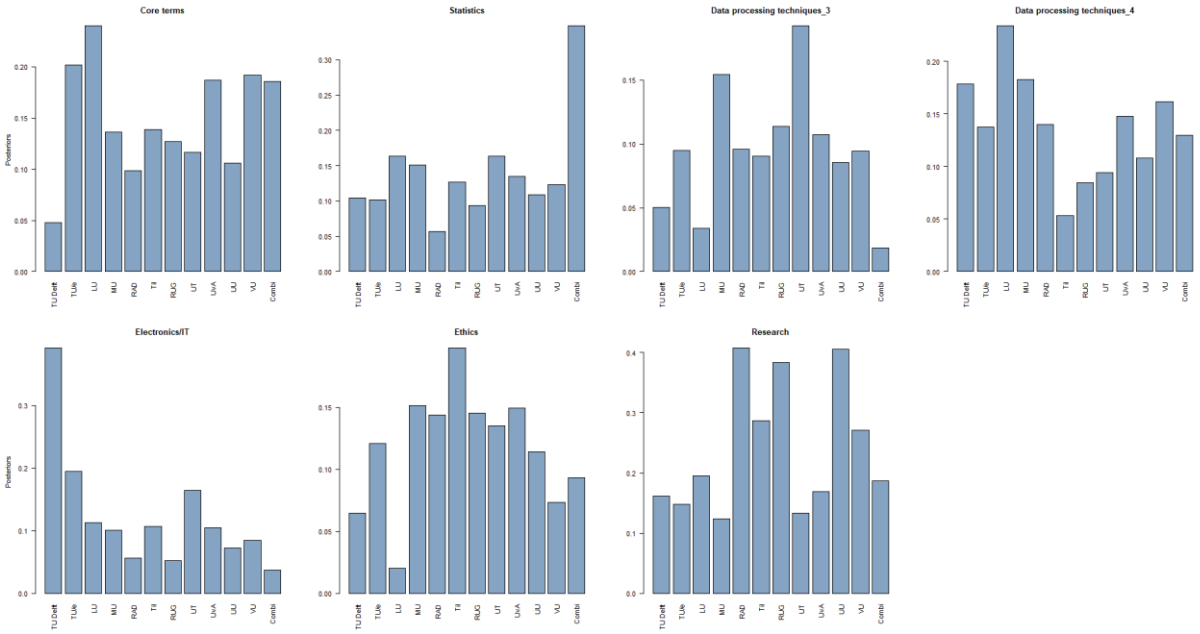


Figure 9. Seven barplots of the posteriors per university for every manually named topic of the CTM outcomes

4.7 Limitations

A limitation of this study is that specific skill sets cannot be derived from the texts alone. Descriptions mostly state key terms such as “deep learning” or “statistics”. As a result, it is hard to derive specific information from the text analyses without further knowledge of this field of science. In other words, expertise will be needed to understand what data science and artificial intelligence graduates are capable of doing. Additional attention and personal effort will have to be put in by the employer to get a good grasp on what the domains of data science and artificial intelligence exactly consist of.

Another shortcoming is that only two metrics have been used to select the optimal number of K topics per model. It is important to rest important decisions of optimization based on a variety of indicators. However, investigated metrics have been examined and have been found to be not informative enough to be operationalized. In future research, more optimizers should be examined to be able to make better informed decisions.

Lastly, both program and course descriptions have been used in these analyses. This leads to very broad descriptions and more narrowed down descriptions being put on the same

pile and being considered as the same. In addition, some programs have a higher number of courses than other programs, which might cause the results to be skewed towards the programs with a higher number of courses (and thus a higher number of descriptions and therefore a higher number of words). These programs might have more influence on the results than programs with a lower number of courses. This is also relevant for the posterior classification, since different universities have different totals of programs included in this study. Therefore, results might be skewed towards the universities with a higher number of programs included.

Despite these shortcomings, the findings do shed a light on the very ill-defined concept of data science and artificial intelligence. The most important domains of data science and artificial intelligence were exposed in a data set of 1.009 descriptions and therefore this might be considered the start of the large task of completely defining these fields.

4.8 Conclusion

In this study, the following research question has been investigated: “What are the skill sets within data science related master programs throughout the Netherlands and which of these skill sets are subject specific (i.e. Biology, computer science, etc.)?”. Three text analyses methods, LSA, LDA and CTM with an optimal number of topics have been performed on 1009 program and course descriptions of data science and artificial intelligence programs. These methods point out the important domains of data science and artificial intelligence for which skill sets are needed, further stated in Table 10. These methods also show if and which universities’ data science and artificial intelligence programs are leaning towards a specific side domain. These findings can form a basis for the start of the complex journey of precisely defining the fields of data science and artificial intelligence.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J. Roy. Statist. Soc. Ser. B 44* 139–177. MR0676206

Aitchison, J. (1985). A general class of distributions on the simplex. J. *Roy. Statist. Soc. Ser. B 47* 136–146. MR0805071

Aitchison, J., & Shen, S. (1980). Logistic normal distributions: Some properties and uses. *Biometrika 67* 261–272. MR0581723

Allen, J., Belfi, B., Fouarge, D., Holtrop, N., & Kozole, S. (2021). Niet-routinematige vaardigheden in hbo-profielen. ROA. *ROA Reports No. 003* https://doi.org/10.26481/umarep.2021003

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blei, D.M., & Lafferty, J.D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D.M., & Lafferty, J.D. (2009). *Topic Models.* Chapman and Hall/CRC

Deerwester, S., Dumais, S.T., Furnas, G.W., Landbauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1: 61–84. http://doi.org/10.3166/dn.17.1.61-84

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1: 5228–5235. http://doi.org/10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software, 40* (13).

Jones, T., Doane, W., & Attborn, M. (2021). textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.5.

Juan, C., Tian, X., Jintao, L., Yongdong, Z., & Sheng, T. (2009). A density-based method for adaptive lda model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008* 72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011.

Markow, W., Braganza, S., Taske, B., Miller, S.M., & Hughes, D. (2017). The Quant Crunch. How the Demand for Data Science Skills is Disrupting the Job Market. *Burning Glass Technologies.* Boston. (Non-Scientific)

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262-272.

Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed Algorithms for Topic Models. *Journal of Machine Learning Research, 10*, 1801–1828.

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015) An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications, 42*(13), 5645-5657.

Rajkumar Arun, V., Suresh, C.E., Veni Madhavan, M.N., & Narasimha Murthy. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Advances in Knowledge Discovery and Data Mining, Mohammed, J., Zaki, Jeffrey, Xu Yu, Balaraman Ravindran, & Vikram Pudi (eds.). *Springer Berlin Heidelberg*, 391–402. http://doi.org/10.1007/978-3-642-13657-3_43

Sigelman, M., Bittle, S., Markow, & Francis, B. (2019). The Hybrid Job Economy. How New Skills Are Rewriting the DNA of the Job Market. *Burning Glass Technologies.* Boston. (Non-Scientific)

Zins, C. (2007). Knowledge map of information science. *Journal of the American Society for Information Science and Technology, 58*(4), 526-535.

Appendix A

Search keys Data Science and Artificial Intelligence

- knowledge map AND data science
- knowledge map AND data science OR AI OR artifical intelligence
- knowledge map AND artificial intelligence
- knowledge map AND computer science
- knowledge map AND data science OR AI OR artifical intelligence OR computer science
- definition of data science

Appendix B

Erasing non-informative words: “ad” and “hoc”

The words “ad” and “hoc got printed in one of topics of the LSA model. The words “Ad”, “hoc” and “ad_hoc” would be stated here, which would contain 30% of the total shown words per topic, whereas the words would not hold any interpretable information. The words ad and hoc were only used in one description, but were used a lot and always together. This would lead to the LSA prioritizing this word to the extent where it would end up in one of the top 10 words of one of the 7 topics. This would lead to too much information being lost and therefore we decided to delete the words from the descriptions. Therefore the optimal number of K metrics and all analyses were rerun. Since the algorithms for finding an optimal number of K would run for two full days on a normal computer with a quad core processor. In light of not spending too much time we chose to ignore other combinations of words that would lead to some information being lost. An example of this case is in the LSA where k = 7 outcomes, topic 4 where “multimedia”, “search” and “multimedia_search” would be stated. These cases erose after removing the words “ad” and “hoc” and therefore this might become an endless cycle of removing non-optimal words from the analyses.

Appendix C

Top ten most frequent words per topic for the LDA

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	data	data	data	data	model	data	model
2	process	datascience	multimedia	model	data	machinelearning	algorithms
3	visualization	research	datascience	analysis	methods	knowledge	theory
4	project	programme	ethical	methods	linear	techniques	problem
5	mining	business	privacy	techniques	problem	programming	techniques
6	topic	knowledge	law	naturallanguageprocessing	statistical	system	reinforcement
7	datascience	methods	big	statistical	algorithms	language	artificialintelligence
8	event	analysis	information	timeseries	theory	algorithms	machinelearning
9	probabilistic	big	system	language	analysis	datamining	methods
10	information	program	legal	modeling	optimization	apply	basic
Core	*Core terms*	*Research domain*	*Law / ethics*	*Statistics / techniques*	*Statistics*	*Machine learning*	*Artificial Intelligence*

	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
1	image	system	network	research	deeplearning	business
2	analysis	design	social	project	neuronetwork	information
3	techniques	financial	system	marketing	information	security
4	processing	human	research	business	retrieval	data
5	computer	data	cognitive	design	model	management
6	vision	smart	human	knowledge	data	thinking
7	model	model	computational	problem	deep	services
8	system	complex	neuroscience	thesis	web	system
9	recognition	support	software	development	system	architecture
10	object	health	information	scientific	applications	processes
Core	*Analysis process*	*Subdomains*	*Human brain*	*Marketing / research*	*Deeplearning / networks*	*Subdomains*

Appendix D

Top ten most frequent words per topic for the CTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	network	model	data	data	data	system	research
2	datascience	data	process	techniques	information	problem	design
3	programme	statistical	naturallanguageprocessing	algorithms	model	algorithms	scientific
4	data	analysis	techniques	analysis	programming	quantum	project
5	social	techniques	model	machinelearning	visualization	techniques	knowledge
6	research	methods	web	datamining	system	optimization	field
7	knowledge	theory	language	information	retrieval	artificialintelligence	researchproject
8	artificialintelligence	linear	datascience	knowledge	language	design	software
9	project	timeseries	information	theory	processing	methods	speech
10	information	regression	mining	applications	big	model	questions
Core	*Core terms*	*Statistical analyses*	*Natural language processing*	*Machine learning*	*Information*	*AI / optimization*	*Research*

	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
1	image	system	deeplearning	business	project	model
2	machinelearning	design	reinforcement	data	research	system
3	model	security	algorithms	marketing	thesis	computational
4	data	data	model	datascience	economic	methods
5	techniques	health	neuronetwork	innovation	system	modeling
6	methods	business	system	management	analysis	human
7	analysis	management	methods	services	problem	processes
8	processing	software	deep	knowledge	data	theoretical
9	computer	digital	problem	concepts	services	neuroscience
10	vision	information	techniques	customer	start	cognitive
Core	*Analyses*	*Information security / business*	*Machine learning*	*Business domain*	*Research*	*Modelling / Human brain*