University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 16/2020 (322)

SO CLOSE AND SO FAR. FINDING SIMILAR TENDENCIES IN ECONOMETRICS AND MACHINE LEARNING PAPERS. TOPIC MODELS COMPARISON.

Marcin Chlebus
Maciej Stefan Świtała

Warsaw 2020

WORKING PAPERS 16/2020 (322)

So close and so far. Finding similar tendencies in econometrics and machine learning papers. Topic models comparison.

Maciej Stefan Świtała, Marcin Chlebus*

Faculty of Economic Sciences, University of Warsaw
** Corresponding author: mchlebus@wne.uw.edu.pl*

Abstract: The paper takes into consideration the broad idea of topic modelling and its application. The aim of the research was to identify mutual tendencies in econometric and machine learning abstracts. Different topic models were compared in terms of their performance and interpretability. The former was measured with a newly introduced approach. Summaries collected from esteemed journals were analysed with LSA, LDA and CTM algorithms. The obtained results enable finding similar trends in both corpora. Probabilistic models – LDA and CTM – outperform the semantic alternative – LSA. It appears that econometrics and machine learning are fields that consider problems being rather homogenous at the level of concept. However, they differ in terms of used tools and dominance in particular areas.

Keywords: abstracts, comparison, interpretability, tendencies, topics

JEL codes: A12, C18, C38, C52, C61

1. Introduction

In the literature, topic models rarely have been being compared with each other. The vast majority of the papers conducting analysis on textual data focuses on implementing LDA in multiple contexts as improving text classification (Ramage et al., 2009), information retrieval (Wei and Croft, 2006) or dynamic patterns capturing (Al Sumait et al., 2008). Sometimes LSA is applied for e.g. translation problems (Tam et al., 2007) or information networks (Wang et al., 2013). There are hardly any positions in the literature that take into consideration analysing text corpora containing papers' abstracts too. Moreover, no research taking into account abstracts of papers associated with econometrics and machine learning can be found.

When analysing the papers which cover the topic of comparing different topic models, different approaches are applied for grouping several pieces of texts. What is mutual for all of them is a strong desire to obtain both informative and interpretable topics (mostly: Chang et al., 2009; Stevens et al., 2012). It makes the researchers to compare various algorithms in terms of their quality as well as suggest several measures of topic models performance (Aletras and Stevenson, 2013; Mimno et al., 2011; Newman et al., 2010).

When conducting a study comparing many topic models, articles of different kinds are usually analysed. They used to be associated with popular science or journalism (Chang et al., 2009; Rubin et al., 2010; Stevens et al., 2012). It seems that the heterogeneity of actual topics in such corpora makes them desirable. Sometimes, personal opinions or complaints are the objects of interest (Niraula et al., 2013). It happens that the corpora are prelabelled (Chiru et al., 2014). It enables using well-established performance measures that are usually applied to different models solving classification problems. Surprisingly, scientific papers are hardly ever taken into consideration. What is more, no topic modelling researches that would analyze the tendencies being present in econometrics and statistics as well as machine learning were found. It creates a niche for studies that would apply any topic modelling algorithms to these particular corpora.

It seems to be a common practice that semantic based models are confronted with the probabilistic ones. The first group is usually represented by Latent Semantic Analysis (LSA) whereas well-established Latent Dirichlet Allocation (LDA) and slightly new Correlated Topic Model (CTM) stand for the second one. Probabilistic Latent Semantic Analysis (PLSA) is also taken into account as a combination of both mentioned ideas. Stevens et al. (2012) took into consideration LSA and LDA being the most popular approaches of the mentioned types respectively. Neither the last not the former was reported of the highest performance in all of

analysed cases. Still, it was LSA that was reported of having highest correlation between used topic coherence measures and topic ranks assigned by experts evaluating the output. The conclusions that were drawn in the studies conducted by Bergamaschi and Po (2014) or Niraula et al. (2013) named LDA a better performing model. Chiru et al. (2014) compared both algorithms using a prelabelled corpora. Comparing models performance with accuracy, Kappa nad F-measure they avoid naming either LSA or LDA a better method. Titov and McDonald (2008) compared extensions of both approaches and stood for the probabilistic one. When comparing PLSA with LDA and CTM, the last two clearly outperformed the first (Chang et al., 2009).

A vast majority of researchers uses a human evaluation of the topics interpretability. Still, some of them suggest automatic measures. Considering only the ones that enable comparing models with substantially different mechanics, it should be concluded that they are based on analysing topics top terms. Newman et al. (2010) recommends taking into account the co-occurrences of such tokens for calculating every single topics coherence. An alternative is suggested by Mimno et al. (2011). Aletras and Stevenson (2013) present another concepted deeply rooted in the idea of n-grams – a context windows around each of most important terms are considered. Nevertheless, the mentioned ideas compare tokens only within a single topic. As a result, the similarity of the words describing only one group is analysed.

2. Aim of the work

The main goal of the research is to conduct two empirical studies – on different text corpora – in aim of comparing three different approaches to topic modelling: LSA, LDA and CTM. The choice of the considered models was made with respect to their typology, popularity and level of concepts complication. As a result, the model based on semantics and linear algebra (LSA) was confronted with the probabilistic ideas – most popular in the literature LDA and rather complex CTM. The expected output should be a recommendation of the best performing algorithm. Conclusions should be made with respect to both an automatic measure of models quality and a human evaluation of topics interpretability. A brief comparison of different algorithms seems to be important because of not so many elaborations of such kind being present in the literature. As a result, many researchers limit themselves to using only one, most popular approach which does not have to be the best performing one (e.g. Al Sumait et al., 2008; Ramage et al., 2009; Tam et al., 2007; Wang et al., 2013; Wei and Croft, 2006). Even if it is, elaborations about its advantages and maybe weaknesses can be useful for improving the existing solutions.

Another aim of the paper is to introduce a new measure of topic models performance. Its idea is based on most prominent tokens co-occurrence in the analysed corpus. Broadening the literature, it takes into consideration both the similarity within every single topic and the desire of significant differences between them. The last aspect is usually neglected in the literature (Aletras and Stevenson, 2013; Mimno et al., 2011; Newman et al., 2010; Roder et al., 2015; Stevens et al., 2012). It creates a bottleneck for improving the whole idea of automatic topic models performance measures. The concept of the suggested solution can be easily explained using a system of nets with nodes being the most important tokens and connections occurring when both terms can be observed in the same document at least once. The shortest distances between the nodes describe the coherence of each topic as well as its dissimilarity in comparison to the remaining ones. The averages of them are being considered for every single topic. They are calculated separately for the pairs of tokens suggesting a particular topic and the pairs consisting of one term describing a certain subject and the other from the remaining tokens. Ratio of the computed values is taken into account for the purposes of optimization and making comparisons.

Another purpose of the study is associated directly with the data used for modelling. The models are about to compare the tendencies observed in the papers collected from different branches of science. In other words, abstracts of articles that analyse econometric and statistical issues were compared with the machine learning researches reviews in terms of identified topics. Such fields were chosen since the ideas rooted in econometrics and statistics are usually applied to the problems that are also commonly analysed with a usage of machine learning tools. It appears that at the level of concept, the problems analysed among both fields are very similar. As far as the tools differ between the scientific branches, it seems interesting if any mutual topics can be found. Moreover, the approaches recommended by the econometricians and data analytics using machine learning algorithms can be perceived as competing on many grounds. The shares of mutual trends identified in both corpora can name the strengths as well as weaknesses of both areas.

3. Material and Methods

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997; Landauer et al., 1998) is a approach that was initially established for the needs of psychology. Still, the clarity of the idea as well as a strong mathematical background made it one of the most common approaches for analysing textual data. Contrary to other models, LSA does not use any probabilistic estimations

and is based on linear algebra and semantic similarity between words. What is more, it is often used as a benchmark when different models are compared (e.g. Bergamaschi and Po; 2014).

LSA can be broadly described as a model analysing a large collection of natural text. Its output is a representation capturing the similarities of words among documents. The model bases on Singular Value Decomposition (SVD). Conducting such analysis starts with presenting the text in a form of matrix (denoted as A) with n rows and m columns. The rows describe unique terms that occur in the considered text collection. Each column represents a text context i.e. a particular document. Cells in the matrix stand for the frequency of words in the analysed passage. Each cell A_{ij} is transformed and weighted by a function expressing word's importance. Word frequency is converted into a general inverse form B_{ij}.

$$B_{ij} = \mathrm{A_{ij}} \frac{m}{\mid \forall k \in \{1, \dots, m\}: A_{ik} > 0 \mid} \qquad [1]$$

The transformed matrix is then decomposed using the SVD approach into $B = U\Sigma V^T$. As a result, a product of three matrices is further analysed. The first component U denotes the original rows. The third matrix V depicts the initial columns. It should be addressed that both matrices consist of the eigenvectors of BB^T (which is called a word similarity matrix) and B^TB respectively. Another important detail that should be taken into account is that Σ is a diagonal matrix. It contains such values that after a multiplication of all three matrices, the initial one B is perfectly reconstructed. Therefore, the values of Σ have to be the singular values of BB^T.

The eigenvectors in U matrix enable distinguishing rows in BB^T. As a result, when looking for the best approximation of the word similarity matrix in k-dimensional space, one should take into consideration the first k rows in U (assuming that singular values in Σ are sorted in a descending order). The U object with only k rows in the literature is usually denoted with Λ_k. The i-th row of this particular matrix, expressed as $\Lambda_k(i)$, is called the LSA feature vector for the i-th word.

Idea of decomposition is applied here since any matrix can be perfectly decomposed in a way that the number of used factors is at most the original matrix's smallest dimension. Moreover, it can be mathematically proved that using fewer factors will result in reconstructing a matrix being the best least-squares fit. The dimensionality is usually reduced by removing coefficients in the diagonal component Σ. A common practice is deleting the smallest ones (Landauer et al., 1998).

LSA distributes the analysed documents among a specified number of topics using inter-document similarity. Assuming that vector f_i represents the word frequencies in an i-th document – d_i and f_{ij} denotes the frequency of j-th term in this document, one can calculate such metric in a following way:

$$M_{ij} = \cos(\lambda_i, \lambda_j) = \frac{\sum_k \lambda_{ik}\lambda_{jk}}{\sqrt{\sum_k \lambda_{ik}^2 \sum_k \lambda_{jk}^2}} \quad [2]$$

where $\lambda_i = \sum_j f_{ij}\Lambda_k(j)$.

Another model taken into consideration is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The main idea is an assumption of all the pieces of text being represented over latent topics as random mixtures. Distribution over words is what characterizes the topics. When considering N words, for each term w_n, a topic z_n from multivariate distribution with a parameter θ is drawn. Every single word is chosen from a conditional probability $p(w_i \mid z_n, \beta)$. Parameters N and θ are drawn from Poisson and Dirichlet distributions respectively. Still violating an assumption of N being a number of words has no meaningful impact on the estimations because it is independent from θ and z. Describing the former, the model assumes known and fixed dimensionality k. As an implication, the same assumption is made when considering the dimensionality of z. A $k \times V$ matrix, denoted as β, is also used for parametrizing the word probabilities, where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.

If the random variable denoted as θ takes only non-negative values and the assumption of $\sum_{i=1}^{k} \theta_i = 1$ is fulfilled, then it takes values in the $(k-1)$-simplex. Its probability density can be therefore expressed as:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \quad [3]$$

where $\Gamma(x)$ denotes the Gamma function and α is a vector of length k with non-negative components.

With parameters α and β defined, a joint distribution considering θ, z and w can be calculated:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \quad [4]$$

It enables obtaining probabilities of words and corpus:

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \quad [5]$$

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad [6]$$

where: θ_d is a document-level variable whereas z_{dn} and w_{dn} are word-level variables.

One more detail that should be addressed is a very important difference between LDA and LSA. In case of the former, the matrix containing word-document co-occurrence C is normalized and decomposed into two matrices instead of three:

$$C = \Phi \times \Theta \quad [7]$$

where the right-hand side's symbols denote respectively: a matrix being a mixture components (with dimensions: words × topics) and a matrix called mixture weights (with dimensions topics × documents). The idea comes from a fact that the diagonal matrix considered in SVD which is used for LDA can be easily absorbed in the remaining two matrices. Moreover, LDA puts initial constraints on the topic and word distributions. On the other hand, decomposition of the main matrix applied in LSA seems to be more convenient in term of computation (because of an orthogonal basis).

The last of the models being considered is the Correlated Topic Model (CTM) (Blei and Lafferty, 2007). It is commonly classified as a hierarchical topic model. The words observed in each document are modelled using a mixture model. The components are shared by all analysed pieces of text.

It is initially assumed that a document consisting of N words arises from a generative process. It can be described as drawing a topic assignment, for n-th word in d-th document – $z_{d,n}|\eta_d$, from a multivariate distribution with a parameter being a function of η_d. This particular parameter is drawn from a normal distribution with parameters (μ, Σ) which are respectively: k-vector and a $k \times k$ covariance matrix (where k denotes a specified number of topics). Each word $w_{d,n} \mid \{z_{d,n}, \beta_{1:k}\}$ is assumed of being drawn from a multivariate distribution with a parameter being equal to $\beta_{z_{d,n}}$. Important notification is that $n \in \{1, \dots, N_d\}$.

Contrary to LDA and its variations, CTM takes into consideration the dependencies between the elements of the simplicial vector. Latent Dirichlet Allocation models draw the topic proportions from Dirichlet distribution. Same draw could be simulated by drawing elements from independent Gamma distributions with taking into account normalization of the outcoming vector. Correlated Topic Models use multivariate Gaussian distribution to produce multivariate parameters. Including correlations between topics has obvious advantages and reflects the empirical findings. Still, such methodology makes the further calculations rather complicated. When $\beta_{1:k}, \mu, \Sigma$ are given, posterior distribution considering the latent variables under words in document's condition is estimated:

$$p(\eta, z \mid w, \beta_{1:k}, \mu, \Sigma) = \frac{p(\eta \mid \mu, \Sigma) \prod_{n=1}^{N} p(z_n \mid \eta) p(w_n \mid z_n, \beta_{1:K})}{\int p(\eta \mid \mu, \Sigma) \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n \mid \eta) p(w_n \mid z_n, \beta_{1:K}) d\eta} \quad [8]$$

It should be addressed here, that calculating the sum over K values of z_n inside the product over terms results in taking into consideration a combinatorial number of words. K^N can be too much in terms of computational tractability. Even if it is not, another problem can be easily spotted – the distributions of topic assignments and proportions cannot be named conjugated. Therefore, the integral cannot be computed analytically. Simplifications are used for estimating its value. The literature suggests deterministic alternatives to Markov Chain Monte Carlo (MCMC) method (Blei and Lafferty, 2007).

Another computational problem occurs when estimating the topics under documents' collection being given $\{w_1, \dots, w_d\}$. The objective is maximizing the likelihood of the documents being a function of $\beta_{1:K}$ as well as μ and Σ. A common problem in models using latent variables is the need for marginalizing the latent structure out when computing the marginal likelihood. Blei and Lafferty (2007) recommend using variational expectation-

maximization (EM) for solving the problem. In contrary to well-known EM, its mentioned variant uses a variational approximation in the first step.

The optimization of models need to be performed considering a certain measure or criteria. There is no denying that the literature is very heterogenous when it comes to choosing the way of accessing topic models' quality. Many suggested solutions cannot be used for comparing the probabilistic models with the semantic alternatives. For example, the well-established perplexity (Bahl et al., 1983) measure cannot be used for Latent Semantic Allocation since it requires log-likelihoods for its computation. Those are not calculated in this particular method. Still, there are some measures that are capable of comparing models of very different mechanics (Mimno et al., 2011; Newman et al., 2010; Roder et al., 2015; Stevens et al., 2012). The authors of the previously described topic measures used to take into account the coherence of topics by analysing the co-occurrences of their top terms. It need to be addressed that the analysed pairs of tokens are compared only within a single topic. As a result, despite the fact that the similarity of the words within each topic is evaluated, the dissimilarity between the topics is not analysed. Therefore, there is still a space for suggesting new measures that would capture both effects at the same time.

One can imagine a net with nodes being topics' top terms identified during modelling. The nodes in the net are connected then and only then if within the considered corpus it can be found at least one document with both words appearing together. Computing shortest paths between the nodes can be named a basis for the measures used for comparing topic models in this paper.

The concept introduced in this paper takes into account the distances between nodes included in the described net. There is no denying that a well-defined topic should characterize itself with having such top terms that are often observed together i.e. in the same document and at the same time, such tokens should be quite rarely observed together with the top terms from the other topics. Therefore, the net is examined for the distances between the nodes that stand for a single topic top terms. An average of the distances will be further called **a within measure**. It can be expressed in a following way:

$$within\ measure = \sum_{k=1}^{K} \frac{1}{K} \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \min \{dist(x_{i,k}, x_{j,k})\}}{N(N-1)} \quad [9]$$

where K and N stand for number of topics and number of top terms found for each topic respectively. An i-th term from the ones being named most prominent in topic k is denoted with $x_{i,k}$. The component $\min\{dist(x_{i,k}, x_{j,k})\}$ is the shortest distance in a net which nodes are all the top terms and connections occur only when both tokens can be observed together in at least one document.

If the within measure is equal to 1, it means a perfectly identified topics (when analysing each topic's top terms alone). Such value means that for each pair of top terms there is at least one document in which both terms were observed. Such solution is similar to the ones already existing in the literature (Aletras and Stevenson, 2013; Mimno et al., 2011; Newman et al., 2010). Still, it is not identical cause it assumes that finding only one paper with both words together is enough for naming them being sufficiently associated with each other. It seems that such assumption can be violated in case of large corpora. Strong homogeneity of analysed texts can be also perceived as a problem.

Were one about to consider the dissimilarity of different topics, the distances between each topic top terms and all the other top terms should be calculated. An average of the distances is named **a between measure** for the further reading convenience. It can be calculated as:

$$between\ measure = \sum_{k=1}^{K} \sum_{l=k+1}^{K} \frac{1}{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \min\{dist(x_{i,k}, x_{j,l})\}}{N^2(K-1)} \quad [10]$$

where again K and N stand for number of topics and number of top terms found for each topic respectively and an i-th term from the ones being named most prominent in topic k is denoted with $x_{i,k}$. The component $\min\{dist(x_{i,k}, x_{j,k})\}$.

When the between measure is equal to 1, it can be interpreted as for each pair of terms, where one element is drawn from the considered topic's top terms and the other from the ones that are most prominent for the rest of the topics, there is at least one document where they both happen to be observed together. Contrary to the within measure, the between metric means a better defined topic, when is higher.

Having two measures, one need to be maximized and the other minimized, a ratio of both can be taken into consideration. As a result the optimization criterion when looking for best parameters among different models was **the between measure divided by the within measure**. Such metric implicates the better topics, when it is higher. A huge advantage is that

it can be used for comparing any topic models since for its computation only the topics' top terms are required. It is calculated as a division of within and between measures:

$$measures\ ratio = \frac{\sum_{k=1}^{K} \frac{1}{K} \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \min \{dist(x_{i,k}, x_{j,k})\}}{N(N-1)}}{\sum_{k=1}^{K} \sum_{l=k+1}^{K} \frac{1}{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \min \{dist(x_{i,k}, x_{j,l})\}}{N^2(K-1)}} \quad [11]$$

which can be simplified to:

$$measures\ ratio = \frac{N(K-1)}{N-1} \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \min \{dist(x_{i,k}, x_{j,k})\}}{\sum_{l=k+1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} \min \{dist(x_{i,k}, x_{j,l})\}} \quad [12]$$

4. RESULTS AND DISCUSSION

Both analysed corpora were collected using web-scraping techniques. The number of abstracts was decided with a deep consideration of both models' ability of accessing any interpretable topics as well as the computation time of the compared algorithms. 1 888 documents were included in the econometrics and statistics corpus. The majority of the abstracts were collected from the Journal Of Applied Econometrics and the Econometrica. Another dataset that was used for the purposes of the research can be briefly described as a corpus made of 3 216 abstracts with a biggest share of European Journal Of Operational Research's summaries. A set of operations was required to be performed when preparing the data for modelling. All of the texts were tokenized at first. After that, the tokens that were identified as numbers or dots were removed from the further analysis. So were the stop words e.g. about, because did, example, for, these, will etc. Any words describing the publisher of the paper were not analysed too. The remaining ones were stemmed i.e. they were reduced to the semantic core. It was decided to remove top ten most frequent terms from both corpora. In case of the econometrics and statistics abstracts they were: model, use, estim, paper, data, result, show, effect, find, can. When taking into account machine learning summaries, tokens: paper, problem, use, model, result, can, propose, base, optim, show were not considered. For the purpose of justifying such action, it should be once again addressed that these particular words would not have any explanatory power when it comes to interpretability of the topics.

Both datasets were divided into training and test subsets in a proportion of 60:40 before conducting any further transformations. It should be addressed here that none of them were performed in a way that it would make the models to fit the test set.

TF, IDF and TF-IDF metrics were calculated for each word's appearance. It was confirmed with the conclusions drawn from the literature (Chang et al., 2009; Ramos, 2003) that TF-IDF can be misleading in case of topic modelling. In the analysed corpora, the IDF varied a lot, TF was usually very low. It means that the tokens were not repetitive within a single document. Still they happened to appear in different numbers of documents. In fact, it was IDF that enabled differentiating the tokens in terms of frequency. Therefore, it was used instead of TF-IDF. The preprocessing of the data was performed as it usually is recommended among the other researchers. Often is it observed that the word frequency within the documents is neglected in favour of analysing only the shares of documents with a certain word occurring (e.g. Chang et al., 2009).

Subject to the IDF statistics, both least and most frequent tokens were removed from the considered corpora. The key issue was to specify the number of tokens needed for further analyse. It was decided to use different numbers of terms when considering uni-, bi- and trigrams. 3 000, 60 000 and 120 000 tokens were tried respectively in the case of econometrics and statistics abstracts. The dataset including machine learning papers' summaries was limited to 5 000, 100 000, 250 000 tokens. Again different numbers for using different kinds of n-grams was assumed to be reasonable. The idea was to use a certain number of words and n-grams from the middle of sorted by IDF list of tokens. Using the same amount in case of unigrams as when considering higher n-grams would undoubtedly affect the results. The numbers of tokens being considered was chosen arbitrary. Still, it should be perceived as a rather educated guess. Some test models were tried before and the outputs were analysed carefully. As far as any rule can be drawn, it seemed like using a half of available tokens made the results at least interpretable to a certain extent. Moreover, on no account should it be forgotten that the computation time of the algorithms is commonly known of being positively correlated with both number of tokens used and number of topics. Additional benefit was gained from limiting the corpora though.

Analysing the optimization results, the variability of both the between and the within measure as well as their ratio was taken into consideration. All of them were studied among different numbers of topics. It was conducted with respect to all models variants – with n-grams of different kind applied. As an implication, various numbers of tokens were tried too. In case of the econometrics and statistics abstracts, LSA results seemed to be similar when different n-grams used. The within measure varied slightly but was always below the between measure.

Therefore, the models could be assumed reasonable and fulfilled the assumptions of similarity within topics as well as the one considering dissimilarity between them. Rather a decreasement could be observed when higher numbers of topics were included. Still, the ratio seemed to stabilize at a certain level. Only when taking into account the case of 120 000 tokens and trigrams it looked like tending to decrease. One can assume that the results obtained on the training set were rather stable when different parameters used. Taking into account the LDA results, it could be easily observed that when applying unigrams, the ratio of the measures increased with the higher number of topics. In contrary, when using bigrams and trigrams, both measures were very close to each other. Even some cases when the topics were less similar inside than to each other could be observed. Low frequencies of higher n-grams should be perceived as a reason of violating the model performance. Using rare collocations in a huge amount can create an unwanted noise in the analysed corpus. Still, such problem did not occur when analysing CTM which is a probabilistic model too. What happens to be similar is that both LDA and CTM recommend a rather high number of topics.

The results obtained when taking into account the machine learning abstracts seem to confirm most of the conclusions drawn before. LSA was rather stable in terms of the between to within measures ratio. Any sudden moves could be spotted only when the low numbers of topics were analysed. Both LDA and CTM were in favour of the relatively high numbers of topics. What was very different in comparison to the previously analysed corpus was that both probabilistic solutions had problems when bi- or trigrams were used. Some cases should be perceived as not very reasonable. Still, the optimization of the ratio of the measures enables avoiding such coincidences.

Best performing models' variants for each dataset and method were found. When considering the econometrics and statistics corpus, LSA performed best when trigrams and 120 000 tokens included. It recommended using relatively low number of topics. An optimal number was named 4. It was reported with a highest measures' ratio – 1.1413. Next best performing variants suggested using trigrams and numbers of topics lower than 15. Measures ratios were: 1.1185 for the model with 13 topics, 1.1160 when using 10 ones. Further alternatives ended with measure of 1.1142, 1.1122 or 1.1093. LDA performed noticeably worse on the training set than the top LSA alternatives. Moreover, the results were rather similar to each other when taking into account the optimization criterion. The most prominent results were obtained for unigrams and 3 000 unique tokens. The best performing model used 32 as a number of topics. The ratio was 1.0835 then. The best alternatives were found with high values of this particular parameter too – 28, 40, 34, 37 and 24. As far as CTM was optimized using a relatively

short vector of parameters, the interpretation in this case is most obvious. Using unigrams seems to be most important for obtaining the best results. Next, the higher number of topics, the better. As an implication, the highest ratio of 1.1205 was reported for 40 topics and unigrams. Bet alternatives were found with measure being equal to 1.1062, 1.0825, 1.0812, 1.0435 and 1.0412. Were one about to name the best method basing only on the training set results, it would be LSA (1.1413) before CTM (1.1205) and worst performing LDA (1.0835).

When taking into account the corpora consisting of machine learning abstracts, the results were reported similar to the case of the first dataset. LSA performed best with recommending a very low number of topics. Here it was the lowest possible value of 2. Alternatives also recommended relatively low values of this particular parameter – 9, 7, 3, 11 and 8. Bigrams seem to be the best solution in all of these coincidences. It should be addressed that the best solution was noticeably better (1.1773) that top alternatives (best reported with ratio of 1.1254). When analysing LDA alternative, this model was reported of performing worse on the training set than LSA. The best variant was found with ratio of 1.0706. The following options scored 1.0701, 1.0698, 1.0688 etc. The suggested numbers of topics were relatively high, with the optimal value of 39 and 36, 40, 37, 35, 38 as the following ones. Most sophisticated model – CTM, was named worst performing on the training data with ratio of 1.0702. It recommended 30 topics and unigrams as the best set of parameters. Still, the difference between this solution and LDA can be assumed neglectable. Optimization with a rather low number of variants could violate the quality of this particular model a bit.

Still, the most important part of comparing the models should be considering their performance on the test set. LSA happened to be very unstable in terms of between to within measures ratio on both sets of abstracts. The results obtained on the training ones were much higher that the measures calculated for the defined topics with a consideration of the test sets. The scores deteriorated from 1.1413 to 1.0109 on the first dataset and from 1.1773 to 1.0160 in case of the second corpus. LDA and CTM happened to be more stable and finally outperformed LSA on both corpora. Taking LDA into consideration, the ratio dropped from 1.0835 to 1.0230 when analysing the first corpus. Analysing the machine learning abstracts, it decreased from 1.0706 to 1.0176. CTM approach ended with test scores of 1.0420 and 1.0175 when the train performances were respectively 1.1205 and 1.0702.

Comparing both probabilistic solutions on both datasets ended with contradictory conclusions. When taking into account the first corpus, CTM was reported with a better performance (1.0420 versus 1.0230). In the second case, LDA was found better (1.0176 to 1.0175). Still, it should be addressed that the differences between models performance on test

data in the machine learning coincidence were relatively small. Were one to optimize CTM with a deeper consideration of different number of topics, this model could outperform LDA. It somehow confirms the conclusions drawn by Chang et al. (2009). The researchers demonstrated that in general CTM brings about better results than LDA. However in some cases they obtained results that recommended using less complex idea instead of the one taking into account the correlations between topics.

Summarizing the LSA results, this particular models performance can be named the worst. The probabilistic alternatives should be assumed better in both conducted analysis. Such conclusion resembles the majority of studies conducted in the empirical literature (Bergamaschi and Po, 2014; Niraula et al. 2013).

What is interesting, LSA is the only solution that was improved when applying n-grams. It was also reported performing much better on the training sets than when measuring the quality on the test ones. Both issues can be perceived as correlated. The semantic solutions bases on word co-occurrence. Collocations are observed together with the words creating them. Therefore, it seems obvious that they should improve the performance in terms of the introduced measures. However, when such collocations are so rare, that they cannot be observed among the abstracts assigned to the test set, the performance can be very unstable. What is more, low numbers of topics can be the reason of so variating quality of LSA. Using few sets of top terms makes the impact of one single difference between training and test sets much more significant than when having many topics.

Drawing conclusions from the probabilistic models performance, LDA and CTM should be named similar. Both are in favour of recommending relatively high numbers of topics. Moreover, using bi- and trigrams does not improve their quality. It seems impossible to recommend the best solution from the analysed ones without a deep consideration of the estimated topics interpretability. Therefore, a human evaluation was applied as a final criterion.

Table 1. Best topic models considering results on econometrics and statistics abstracts training set.

Econometrics and statistics abstracts						
	Number of tokens	N-grams type	Number of topics	Within metric (TRAIN SET)	Between metric (TRAIN SET)	**Between to within ratio (TRAIN SET)**
Latent Semantic Analysis (LSA)						
1.	120 000	trigrams	4	1.3361	1.5250	**1.1413**
2.	120 000	trigrams	13	1.1096	1.2411	**1.1185**
3.	120 000	trigrams	10	1.1483	1.2816	**1.1160**
4.	120 000	trigrams	6	1.1769	1.3114	**1.1142**
5.	120 000	trigrams	3	1.4617	1.6257	**1.1122**
6.	120 000	trigrams	11	1.1416	1.2664	**1.1093**
Latent Dirichlet Allocation (LDA)						
1.	3 000	unigrams	32	1.1940	1.2937	**1.0835**
2.	3 000	unigrams	28	1.1785	1.2768	**1.0834**
3.	3 000	unigrams	40	1.2112	1.3111	**1.0825**
4.	3 000	unigrams	34	1.1997	1.2976	**1.0816**
5.	3 000	unigrams	37	1.1990	1.2965	**1.0813**
6.	3 000	unigrams	24	1.1723	1.2664	**1.0803**
Correlated Topic Model (CTM)						
1.	3 000	unigrams	40	1.1450	1.2829	**1.1205**
2.	3 000	unigrams	30	1.6862	1.2927	**1.1062**
3.	3 000	unigrams	20	1.1668	1.2631	**1.0825**
4.	3 000	unigrams	10	1.0831	1.1711	**1.0812**
5.	60 000	bigrams	40	1.0390	1.0842	**1.0435**
6.	120 000	trigrams	40	1.0458	1.0888	**1.0412**

Source: own elaboration.

Table 2. Best topic models considering results on econometrics and statistics abstracts test set.

Econometrics and statistics abstracts									
Model	Number of tokens	N-grams type	Number of topics	Within metric (TRAIN SET)	Between metric (TRAIN SET)	Between to within ratio (TRAIN SET)	Within metric (TEST SET)	Between metric (TEST SET)	**Between to within ratio (TEST SET)**
LSA	120 000	trigrams	4	1.3361	1.5250	1.1413	1.0056	1.0165	**1.0109**
LDA	3 000	unigrams	32	1.1940	1.2937	1.0835	1.0271	1.0578	**1.0230**
CTM	3 000	unigrams	40	1.1450	1.2829	1.1205	1.0388	1.0824	**1.0420**

Source: Own elaboration.

Table 3. Best topic models considering results on machine learning abstracts training set.

Machine learning abstracts						
	Number of tokens	N-grams type	Number of topics	Within metric (TRAIN SET)	Between metric (TRAIN SET)	**Between to within ratio (TRAIN SET)**
Latent Semantic Analysis (LSA)						
1.	100 000	bigrams	2	1.2353	1.4544	**1.1773**
2.	100 000	bigrams	9	1.1100	1.2490	**1.1254**
3.	100 000	bigrams	7	1.1496	1.2918	**1.1238**
4.	100 000	bigrams	3	1.3827	1.5425	**1.1156**
5.	100 000	bigrams	11	1.1068	1.2262	**1.1079**
6.	100 000	bigrams	8	1.1035	1.2147	**1.1008**
Latent Dirichlet Allocation (LDA)						
1.	5 000	unigrams	39	1.0483	1.1223	**1.0706**
2.	5 000	unigrams	36	1.0415	1.1146	**1.0701**
3.	5 000	unigrams	40	1.0487	1.1219	**1.0698**
4.	5 000	unigrams	37	1.0490	1.1211	**1.0688**
5.	5 000	unigrams	35	1.0352	1.1063	**1.0686**
6.	5 000	unigrams	38	1.0419	1.1130	**1.0682**
Correlated Topic Model (CTM)						
1.	5 000	unigrams	30	1.1183	1.1968	**1.0702**
2.	5 000	unigrams	20	1.0710	1.1400	**1.0644**
3.	5 000	unigrams	40	1.0597	1.1235	**1.0601**
4.	5 000	unigrams	10	1.0180	1.0535	**1.0349**
5.	100 000	bigrams	10	1.0077	1.0398	**1.0319**
6.	100 000	bigrams	40	1.0129	1.0205	**1.0075**

Source: Own elaboration.

Table 4. Best topic models considering results on machine learning abstracts test set.

Machine learning abstracts									
Model	Number of tokens	N-grams type	Number of topics	Within metric (TRAIN SET)	Between metric (TRAIN SET)	Between to within ratio (TRAIN SET)	Within metric (TEST SET)	Between metric (TEST SET)	**Between to within ratio (TEST SET)**
LSA	100 000	bigrams	2	1.2353	1.4544	1.1773	1.5253	1.5497	**1.0160**
LDA	5 000	unigrams	39	1.0483	1.1223	1.0706	1.0028	1.0205	**1.0176**
CTM	5 000	unigrams	30	1.1183	1.1968	1.0702	1.0281	1.0461	**1.0175**

Source: Own elaboration.

First of all, the best LSA model describing the econometrics and statistics abstracts was taken into consideration. The first topic was reported with time, studi, method, empir, condit, differ, provid, distribut as the most important terms. As far as the enumerated

tokens are very general and not directly connected, it can only be admitted that this particular topic describes some empirical researches. Some words as market and rate suggested connection with macroeconomics. However it was the second topic that suggested macroeconomic issues to a greater extent. Such conclusion can be drawn since words market, price, rate, growth, import, dynam, econom, cycl, labor were found most important. What is more, the remaining top terms clearly demonstrated that this particular topic captured any researches conducted with large datasets. The third topic did not contain any words connected directly with macroeconomics. Moreover, the terms suggested conducting analysis based on some factors distributions or taking into account functions of some parameters. Were one to name the most problematic set of top terms from the analysed models output, the forth one should be chosen. It consisted of the top terms being also observed in the rest of the topics. Some words suggested empirical findings – method, provid, time, set. The others made the macroeconomics being the main object of interest – dynam, state, shock, effic. Still, it was the only topic that contain words seri, stationari as the most prominent tokens. It can be the one that captured the time series analysis.

Figure 1. Econometrics and statistics chosen topics (LSA model).

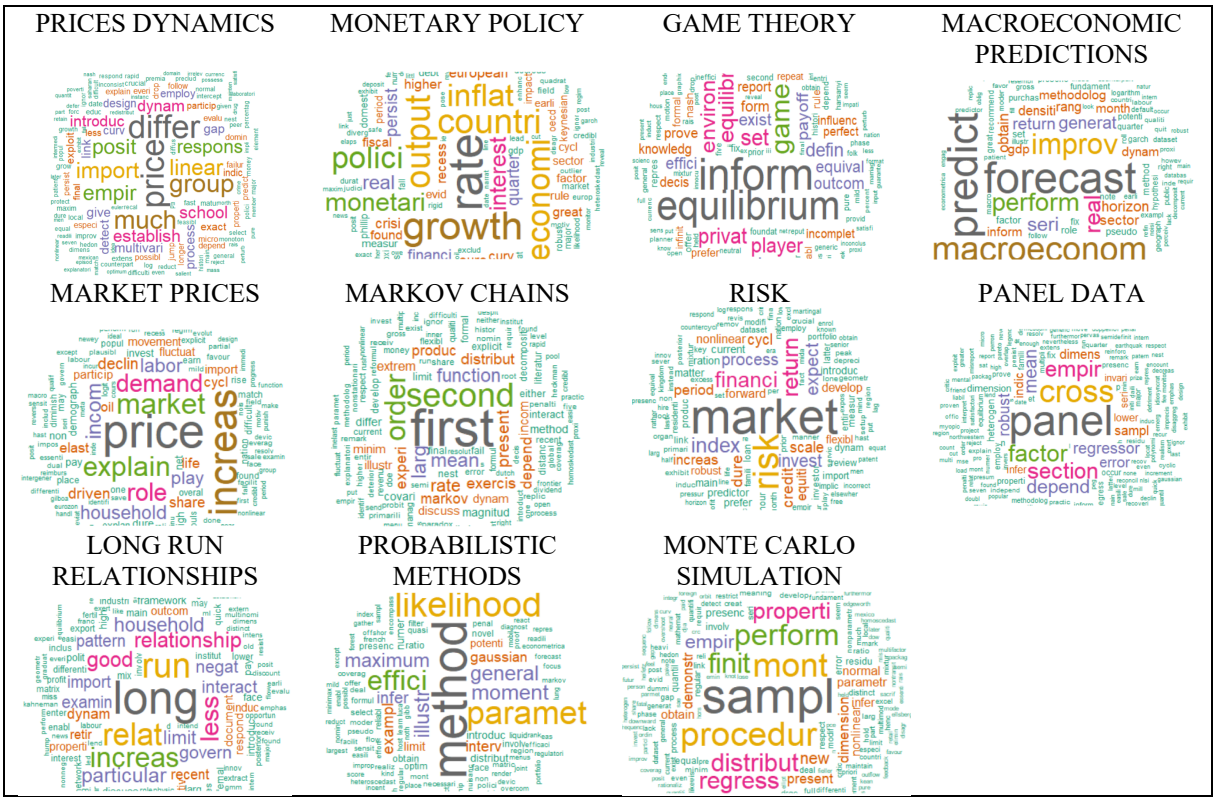
Source: Own elaboration.

When taking into account LDA model, as it was mentioned previously, 32 topics were named. Considering the capacity of the paper, only the main tendencies were briefly described here. It seems that estimating the higher number of topics leads to clearer and more accurate identification of different trends in the literature. Some topics should be named suggesting macroeconomic studies. It looks like the prices variability is quite important for the analysed branch of science. Its dynamics, especially in import is definitely the main object of interest within some topics. Words as rate, polici, growth, monetary, inflat, crisi, interest

seem to suggest that another one is the monetary policy. The problem of its increasement and its impact on households, market and labour was also named in a separate topic. Tokens suggesting critical literature reviews could be observed too. Some topics clearly described optimization problems. Another tendency was identified with words that captured the part of the literature describing the game theory – equilibrium, player, game, payoff, decis. Researches that take into consideration a large amounts of data were also grouped in a stand-alone topic. Microeconomic subjects were classified together too. The problematic of predicting macroeconomic variables was another identified topic. Tokens specific for time series analyses were captured as they were in LSA model. What is very specific, Markov chains applications to different problems was spotted as a whole topic. Another field found among the estimated groups of tokens was credit, financial and market risk. Problems associated directly with the business were also classified together. Next very broad and current topic was panel data analysis. Another one looked like depicting researches that take into consideration any long-run relationships. Probabilistic methods seem to be captured as a stand-alone tendency too. The most important words for analysis using Monte Carlo methods were identified too.

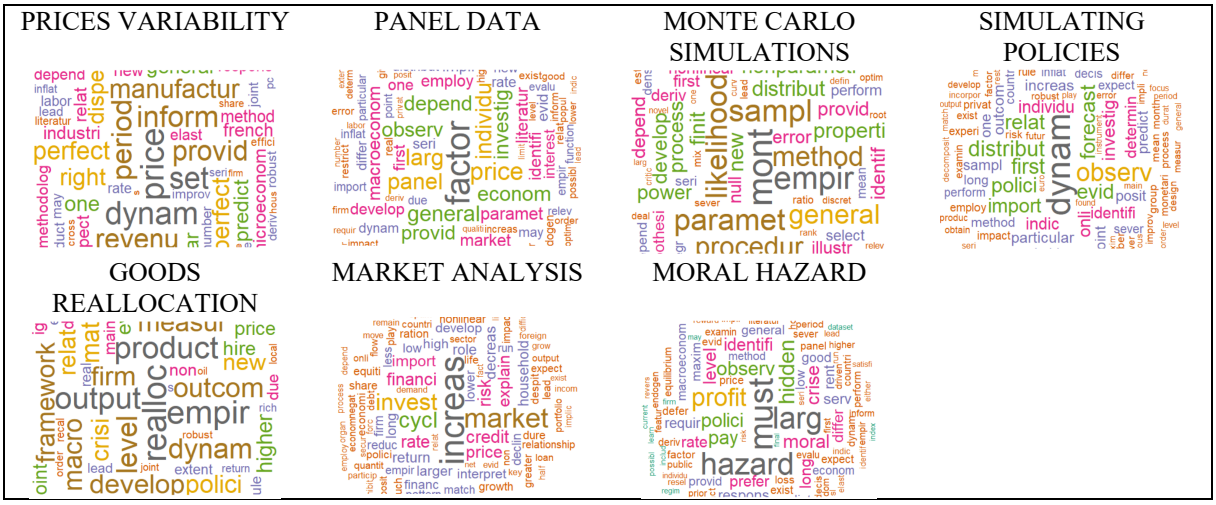
The best performing model – CTM was found performing optimal with 40 topics. It should be admitted that the majority of the word clouds seems to contain similar words in different configurations. Tokens as differ, price, market, larg, set, dynam, can were reported repetitive in many topics. As a result, despite the fact that some differences can be found, it seems very problematic to unambiguously name them. The tendencies that could be identified were similar to the LDA models case. For example, price variability and its impact on economy was captured by several of the estimated topics. Panel data analysis was also observed to be present in certain topics. Monte Carlo simulations was named of having its own group of terms. Optimization problems were found as analysing different policies and simulating different scenarios. Another topic described a problem of goods reallocation. Both can be perceived as related to microeconomic issues identified within LDA topics. Another one captured the words related to different kinds of market analysis. Still, some trends that were observed in LDA topics were not spotted in case of CTM. Monetary policy, game theory, time series analyses and Markov chains applications can be enumerated as examples of the tendencies that were not identified. However, some topics that were not created by LDA algorithm were captured. A stand-alone set of top terms associated with the linear regression method was reported. Moral hazard was found in a separate topic too. Another one represented the papers considering classification problems.

Figure 2. Econometrics and statistics chosen topics (LDA model).



Source: Own elaboration.

Figure 3. Econometrics and statistics chosen topics (CTM model).

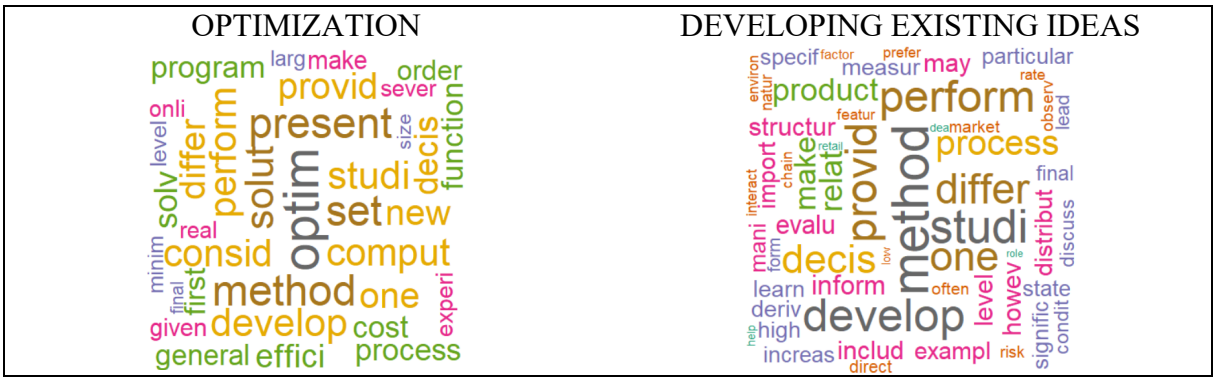


Source: Own elaboration.

The optimized LSA model that was trained on the machine learning abstracts used only two topics. The first one could be generally named considering optimization of different processes. It looked like the papers being assigned to this topic should take into account finding

best solutions for specified problems with a deep consideration of various methods. It was rather the second topic that was focused on developing new ideas. Nonetheless, many of observed tokens repeated in both analysed topics – method, provid, decis, develop, studi etc. However, the importance of them seemed to be noticeably different.

Figure 4. Machine learning chosen topics (LSA model).

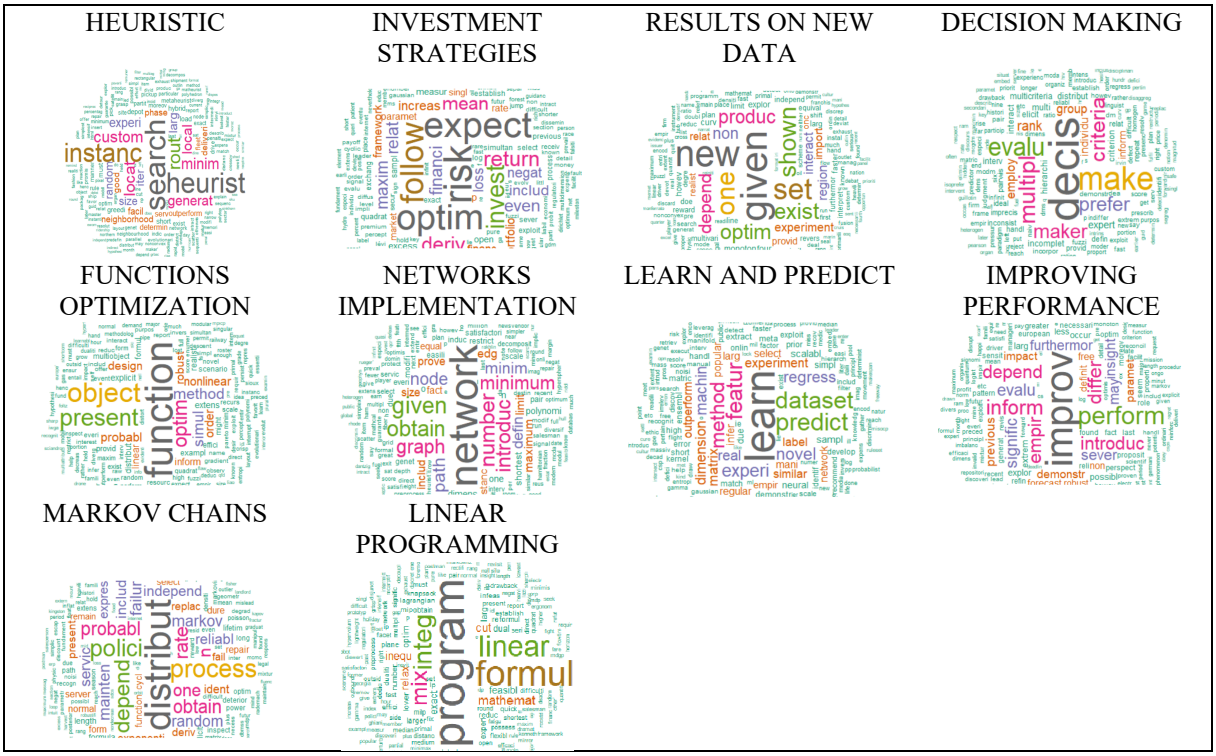


Source: Own elaboration.

Best performing LDA model used thirty nine topics for classifying the machine learning papers reviews. The first one looked like being devoted to heuristic. It seemed to consider the theoretical problems with a strong mathematical background. Another well-defined topic undoubtedly stated for different investment strategies along with their optimization. The next one, described optimization of results when a new (test) data set is given. Topic associated with analysing market share, demand and profits was also identified. Another one described decision making and preferences. Still, it seemed that this particular topic did not include game theory as a similar one identified in econometrics and statistics abstracts. It was rather focused on different algorithms. Next interpretable topic took into account the problem of finding an optimum when analysing a specified function. It could be named a bit similar to the topics identified within the first corpora e.g. the optimal goods allocation problem. Networks implementation was found as a stand-alone topic too. Regression analysis using different machine learning models was found. Again, despite different algorithms being captured by the top terms, the problem can be perceived as a similarity in comparison to the econometrics corpora. So should be the topic associated with making predictions on test set in aim of simulating models performance on the real data. The next one that stood for improving different methods performance should be named another similarity in the conducted comparison. Another group of tokens seemed to describe analysing big data solutions in terms of their

efficiency. Markov chain methods were also identified. Analysing the products demand can be named a main idea of another topic. Linear programming issues were captured by the topic model too. Still, some of the estimated topics were very general and not interpretable enough.

Figure 5. Machine learning chosen topics (LDA model).

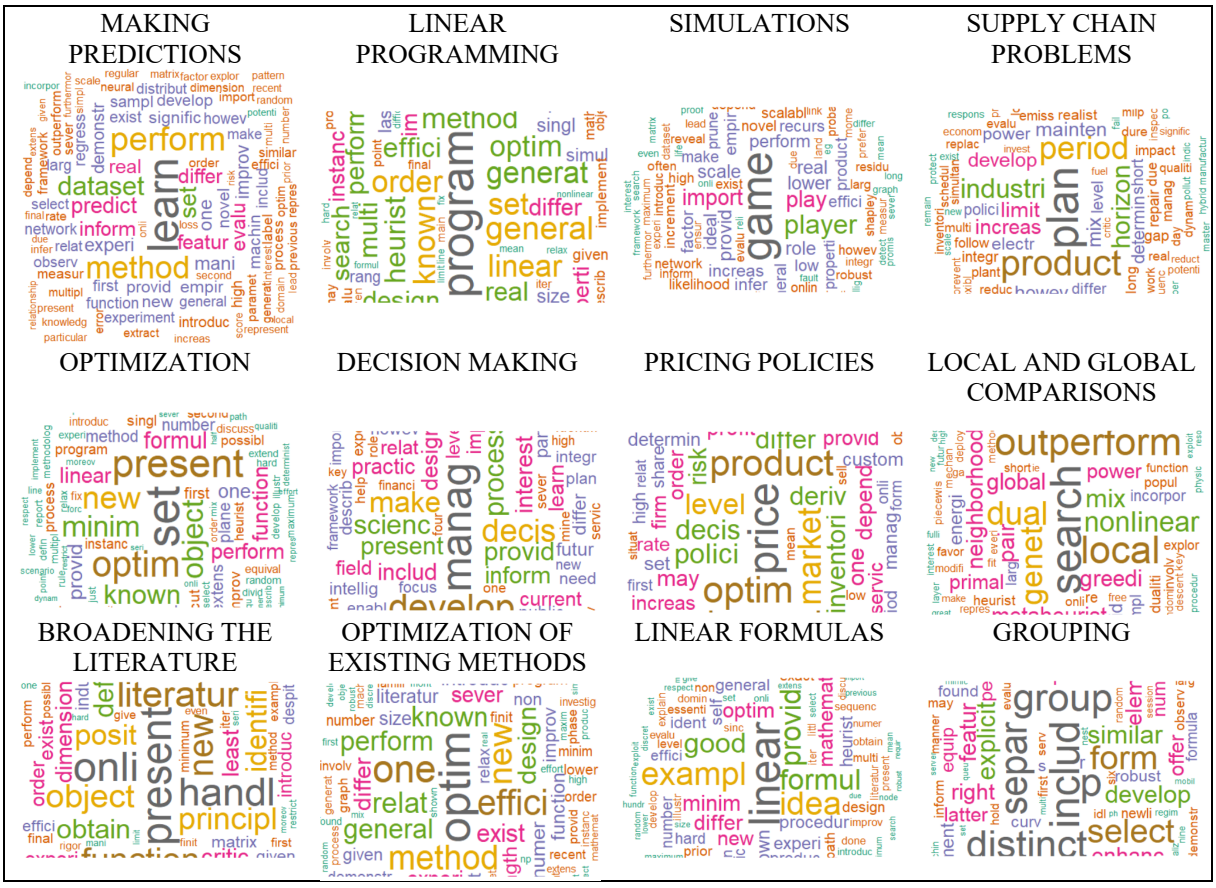


Source: Own elaboration.

Taking into account the CTM model, thirty topics describing machine learning abstracts were prepared. As far as the author is concerned, they are in general less interpretative are the ones computed with LDA. The first one seemed to describe predicting different features with previously learned model. Linear programming along with searching optimal solutions should be named obvious objects of interest in case of some topics. Another suggested that some simulations are described within machine learning papers too. Next of the interpretable topics could describe supply chain problems. Another one stood for unspecified optimization issues. Decision making algorithms were identified too. Next one seemed to be a representation of any analysis connected with pricing policies. Analysing both local and global neighbourhood for making a comparison was found in a separate group of terms. What is more, topics describing some new methods based on the literature review were spotted. So were the ones developing and optimizing the algorithms that were already suggested by some researchers. Another topic

stood for optimizing linear formulas. The next one aimed grouping and then, making a justified selection.

Figure 6. Machine learning chosen topics (CTM model).



Source: Own elaboration.

The introduced measure of topic models performance named LSA worst performing in comparison to the probabilistic approaches. However, the interpretability of the topics seem to be satisfactory. According to this algorithm, basic fields in econometrics and statistics are empirical researches, macroeconomic issues, time series analysis. In case of machine learning it is either optimizing the already established solutions or introducing completely new ones. Nonetheless, the recommended number of topics does not enable identifying any mutual tendencies. The probabilistic alternatives do so. Both LDA and CTM make it possible to find interpretable topics. They are more accurate, probably because of the higher number of topics. In case of the first corpus, monetary policy, game theory, panel data analysis, Markov chains and Monte Carlo methods should be enumerated as main trends in the literature. The fields that can be named most obvious when considering the second dataset are investment strategies,

decision making algorithms, regression problems, classification models, linear programming and big data solutions. Taking this into account, it should be assumed that LSA performs worse than LDA and CTM when it comes to being as informative and interpretable as possible. It should be reminded that such conclusion is similar to the previously conducted researches. Bergamaschi and Po (2014) as well as Niraula et al. (2013) also do not recommend LSA over the probabilistic models. Neither was Chang et al. (2009). Still, it is a contradiction in comparison to Chiru et al. (2014). It can be the data what makes LSA performing better than other models in certain cases.

On no account should be forget that the probabilistic approaches identified many topics that are very general and cannot be assumed fully interpretable. Moreover, there are many tokens that repeat among the topics. Still, it should be addressed that when removing more most occurring words from the corpora, another terms were repeating. As a result it was hard to establish how many words should be removed to make the output clearer. Moreover, removing some often observed tokens as `optim` would have an obvious impact on the topics meaning. What should be mentioned here is that on the literature, it is common that a large number of topics is estimated and not every single one can be perceived as informative (Chang et al., 2009). As an implication, it was assumed that the results obtained in this paper are satisfactory.

It seems not possible to unambiguously name a better performing algorithm when considering the two analysed probabilistic approaches. Probably, when well-optimized, CTM gives slightly better results than LDA when taking into account the suggested measure of performance. It would be consistent with the literature. However, the computation time of such optimization takes a noticeably more time. Nevertheless, is seems necessary that more empirical researches should be tried for comparing these two models. The topics estimated with LDA seem to be a bit more interpretable than the ones computed with CTM. Still, Chang et al. (2009) happened to obtain a case with LDA outperforming the more sophisticated solution too.

Concluding all the above, CTM cannot be named better than LDA when taking into account all of the models aspects. It probably depends on the context of using the models. Still, it appears that in business problems, when the time of computation matters and small differences between the results are neglectable, LDA should be recommended. Both probabilistic approaches were reported better than the semantic alternative. Nonetheless, LSA can be used as a benchmark when conducting analyses based on probabilistic models.

As far as the topics suggested by different models can be named interpretative, they enable identifying the main objects of interest among different branches of science. Were one to compare them, similarities can be found. In case of the analysed corpora, it should be stated

that some problems are constantly being taken into consideration. Both the econometricians and the data analysists using machine learning tools, conduct researches on similar problems. Obviously, the algorithms used for finding optimal solutions are different. Still, the issues analysed in different papers seem to be quite similar at the level of concept. However, there is no denying that both datasets contain topics that are specific for the analysed branches too. They do not repeat between the corpora. It can be an effect of well-established solutions of particular problems that are assigned to only one scholarship. Moreover, it can be the interpretability as well as the explainability of the econometrics methods that results in applying them more often to certain problems.

Optimization seems to be an issue that is present among both datasets. Still, it seems to be more popular within the machine learning abstracts. Moreover, both branches of science focus on optimization considering various issues. The emphasis on certain problems differs among the analysed fields too. In case of econometrics and statistics, the idea of finding optimal solutions is strongly associated with analysing macroeconomic policies. A deep consideration of the monetary policy should be perceived as looking for the best solution. After all, a certain level of inflation is recommended. Making predictions enables establishing this particular level. The fluctuations observed in prices of different goods can be also analysed in terms of a general optimum where the aggregated loss is neglectable. On the other hand, one can easily image taking it into account in aim of maximizing the profits of a certain company. Estimating the demand for particular product should be also perceived as looking for a most desirable recommendation for both the customers and the producers. The game theory assumes that the players are rational and make decisions being optimal for them. Moreover, the econometricians put a great effort into explaining certain markets. Analysing the labour market targets moving towards an equilibrium where an optimal number of employees is observed.

Taking into account the coincidence of the machine learning branch, optimization is probably the most prominent subject. Still, the tools are very different in comparison to the econometrics and statistics. The problems solved with the usage of this particular algorithms are similar at the level of concept but for different reasons are not applied to the same issues. It can be the explainability of different methods that places them in the certain areas. In case of macroeconomic issues it seems to be extremely important to make the results fully interpretable. Same when considering a vast and current topic of financial risk. Econometric and statistic approaches are preferred there because of their ability to be easily explained. Every single parameter means a certain impact on the dependent variable. In case of machine learning tools, the interpretation is usually much harder. Sophisticated models as random forest or gradient

boosting cannot be simply visualised without sacrificing their accuracy. Of course, the surrogated models can be used for this particular purpose. Still, the idea standing behind creating a single decision tree is less interpretable than e.g. logistic regression. Moreover, many law regulations require using methods of certain explainability for solving particular problems. As a result, topic models do not identify using machine learning algorithms for optimization in the same areas as they do in case of the econometrics methods. The optimization methods from the second corpus are reported to be used in analysing different policies and simulating different scenarios. Still, hardly any terms suggesting macroeconomic issues were observed there. A problem of goods reallocation is also present in this corpus what can be perceived as a similarity to the first one. However, it appears that the contexts of applying different tools is different. Nevertheless, there are some machine learning topics that can be named applying optimization tools to the field being analysed by the econometricians. They are supply chain analyses and investment strategies. However, these particular topic cannot be easily identified in case of the first corpus. What can be a certain mutual area for both is optimizing models results. It was not obviously identified within the first corpora. Still, it seems logical, that conducting any estimations with a usage of econometric tools requires making the models as good as possible. It also can be named an optimization. Making the models performing better is a very common and clear tendency found among the machine learning topics.

When it comes to analysing decision making, it can be assumed of being slightly connected with optimization problems. Choosing an appropriate strategy can be named looking for an optimal solution of the problem. Similar topics considering this issue can be observed when taking into account both corpora. The most obvious tool that should be named in case of econometric and statistic ones is the whole concept of game theory. However, it can be easily justified that many different algorithms identified within the first dataset, even if not aiming making best possible decisions directly, can be perceived as applied in such processes. Even time series analyses, ending with a forecast can be a basis for making a business decision. Similar conclusion should be made when taking into consideration the coincidence of different regression and classification models. Such topics are present among both corpora. Predictions made in different scenarios should be named supportive when making any decisions. On no account should we forget that statistics is also used for assuring business choices e.g. by using the confidence intervals.

Moving fluently to the next similarity, making predictions is probably the most obvious one. As mentioned before, this particular topic is associated with optimization as well as decision making. Still, modelling both regression and classification problems seem to be one

of the most common connotation when thinking of both scientific branches. In case of econometrics and statistics abstracts topic related to the former were identified. Linear regression along with panel data analysis seem to be an inseparable part of the first corpus. Classification problems were not observed as a stand-alone topic. Still, when considering e.g. a topic of credit risk, using econometric methods for predicting a binary dependent variable seems to be obvious. Analysing the second dataset, a topic depicting classification models was directly identified. Of course, different algorithms are used in both areas but still the idea can be perceived as being present in econometrics as well as machine learning.

On no account should we forget that simulations are conducted in both econometrics and machine learning. Monte Carlo methods were captured by the topics computed for econometrics and statistics abstracts. Some topics being strongly associated with statistics suggest bootstrapping methods too. Moreover, it looks like researches using Markov chains were identified in both datasets. Again, despite the fact of using different methods, the general idea as well as the purpose of the conducted analysis can be assumed of being similar to each other.

Credit, financial and market risks are considered within the first corpora. Still, topic associated with analysing risk was also found when evaluating models estimated for the machine learning papers reviews. It can stand for the applications of explainable machine learning tools to the mentioned problems. Analysing prices variability seems to be present in both of the analysed datasets too. Still, it is much more frequent in case of econometrics and statistics. In this particular corpus, macroeconomics is usually associated with such studies. Taking into account machine learning abstracts coincidence, it is not only less popular but can be named not so strongly connected with global topics.

Some topics can be observed only when analysing a chosen corpora. Some of them appear to suggest using different methods than taking into account dissimilar problems. Econometrics and statistics abstracts were reported to consider previously mentioned game theory unlikely to the machine learning summaries. It stands for analysing decision making problem as it was named before. The same was found in the second dataset. Another differences in the determined topics are associated with the methodology. Monte Carlo and linear regression can be found only in the first corpus. In contrary, goods allocation and classification models were found in the second one. Still, it cannot be named a difference at the level of generally formulated problem. However, such examples can be named after a deep investigation of the results. Time series happened to be observed only in case of econometric and statistic abstracts. So were theoretical issues assigned to microeconomics as moral hazard. In the corpus

consisting of machine learning summaries, supply chain, linear programming and natural language processing can be named being specific topics.

It looks like both considered branches of science compete in many fields. They recommend various methods for solving similar problems. Nevertheless, some specifications as interpretability, complexity or computation time make them being applied more frequent in certain areas. As a result, analysing the topics obtained from different models, comparing only the top terms in not enough. One need to have at least a basic knowledge of the background of every single method to identify similar tendencies in the literature. It makes the differences harder to be spotted too. Still, some can be found. They are associated with the frequency of using either econometrics or machine learning for solving particular problems. Moreover, methodological dissimilarities are obviously present. Even some areas which appear only among the topics estimated on one dataset can be reported.

All in all, it should be concluded that econometrics and statistics as well as machine learning have mutual tendencies. The topics prepared using LDA and CTM models are interpretative enough to identify the similarities between both branches of science. Still, it should be addressed that as far as the ideas standing behind many of the topics within both corpora are the same, the methods and backgrounds that they are applied for differ a lot. Optimization is a trend being constantly observed in both analysed areas. However, it is applied to very different problems and in various contexts. Considering decision making, the methods used are very heterogenous too. So are they when taking into account another mutual field of making predictions. Simulations appear to be observed in both areas. Moreover, Markov chains happened to be found in both.

5. Concluding remarks

Comparing topic models should be perceived as an important and current subject in the literature. Multiple approaches have already been suggested by many researchers. In the following paper, three solutions – LSA, LDA and CTM were used for making a performance comparison. Their choice was based on the literature review. Moreover, the conceptual similarities between considered algorithms were taken into account. The semantic idea was confronted with the probabilistic alternatives. The popularity of the approaches was also taken into consideration.

The analyses of models performance was combined with a research considering scientific papers abstracts. Two corpora were analysed. One contained econometrics and statistics papers summaries. The other consisted of abstracts of articles that used machine

learning tools for several purposes. It was checked if the analysed branches of science have any mutual topics.

A concept of a new topic models performance measure was presented. It takes into account similarity between the top terms describing every single topic. It should be named quite similar to the well-established topic coherence measures. However, it contributes with considering dissimilarity between the topics. The idea of it was explained and applied in practise.

Topic models were optimized with respect to the introduced measure. Models that performed best were compared considering the test sets from both analysed corpora. In case of econometrics and statistics abstracts, CTM outperformed LDA. When analysing the second corpus, results were contradictory in comparison to the previously obtained. The last was named the better than the latter. LSA ended with being the worst performing model in both coincidences that were taken into account. Furthermore, the topics estimated using different algorithms were checked for interpretability. The probabilistic methods delivered similar results. It seems hard to recommend one over the other. As a result, it can be concluded that the probabilistic approaches are generally better performing than the semantic alternatives. Such results confirm the previous findings exposed in the literature. Still, it seems not possible to honestly recommend CTM over LDA.

The tendencies in abstracts were identified. Topics considering optimization were found in both corpora. This particular idea is very common in both analysed branches of science. However, both econometrics and statistics apply different tools to similar problems. Moreover, the background varies between the corpora. It seems that some problems are analysed more with econometrics. Interpretability of machine learning algorithms can be a serious issue that causes such distribution. Making predictions was named another tendency present in both datasets. Still, it appeared to be a bit hidden in case of the first corpora. In the second one, its presence was clearly demonstrated by certain top terms. More mutual tendencies were identified as decision making algorithms, simulations, considering prices variability and Markov chains application. Again, it should be addressed that these tendencies can be assumed similar at the level of concept. However, several differences including using different algorithms and context make them having certain specific.

It appears that econometrics and statistics as well as machine learning have mutual tendencies that can be identified when analysing summaries of the articles. There is no denying that both branches compete with each other. Different modelling methods are used for particular

studies. Still, many similarities can be found when taking into consideration the conceptual problems.

In the paper, only three topic modelling concepts were taken into account. In the future, comparing more different topic models would be informative. PLSA is the one that could be confronted with the methods tried within this particular research. Furthermore, variations of LSA and LDA should be analysed. Such a broad comparison could enable suggesting a new approach or slight modifications to the existing ones.

Nevertheless, different corpora should be tried for conducting similar researches. Was the analysis repeated using different data, the results could be assured. What is more, another problem connected with natural language processing should be taken into consideration. There is no denying that bank transfer titles can be analysed using topic modelling methods. Looking for a more sophisticated case, legal sources seem to create a broad possibility for conducting similar studies.

The concept of measuring both the similarity within a topic and dissimilarity between it and the others can be obviously expanded. The idea presented in this particular paper assumes that observing two tokens together once has the same meaning as spotting them next to each other in more pieces of texts. It is a simplification that could be omitted. The measures of topic coherence that are suggested in the literature consider the frequencies of word. It could be incorporated to the idea of nets connecting different words. The connections could be weighted with the shares of documents where the words co-occur.

Some empirical study considering how many most occurring words should be removed from the corpus should be conducted. When preparing the analysis for this paper purposes, removing words using IDF concept was applied. Still, it was not enough to get at least satisfactory interpretation of all the topics and more modifications were performed. There is no denying that some words should be treated as another stopwords. Here, it was assumed that top ten tokens with highest TF should be named so. However, as far as some topics were reported being not very interpretable, another algorithm could be suggested in further researches.

References

Al Sumait L., Barbara D., Domeniconi C. (2008), On-line lda: Adaptive topic models for mining text streams with application to topic detection and tracking. Proceedings of the 12th IEEE International Conference on Data Mining, pp. 3-12.

Aletras N., Stevenson M. (2013), Evaluating topic coherence using distributional semantics. Proceedings of the 10th International Conference on Computational Semantics, pp. 13-22.

Bahl L. R., Jelinek F., Mercer R. L. (1983), A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-5, no. 2, pp. 179-190.

Bergamaschi S., Po L. (2014), Comparing LDA and LSA Topic Models for Content-Based Movie Recommendation Systems. Proceedings of the International Conference on Web Information Systems and Technologies, Springer, pp. 247-263.

Blei D. M., Lafferty J. D. (2007), A Correlated Topic Model of Science. The Annals of Applied Statistics, Vol. 1, No. 1, pp. 17-35.

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

Bouma G. (2009), Normalized (Pointwise) Mutual Information in Collocation Extraction. Processing of the Biennial GSCL Conference, pp. 31-40.

Chang J., Boyd-Graber J., Wang C., Gerrish S., Blei D. M. (2009), Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems 21.

Chiru C., Rebedea T., Ciotec S. (2014), Comparison between LSA-LDA-Lexical Chains. In Proceedings of the 10th International Conference on Web Information Systems and Technologies, pp. 255-262.

Damashek M. (1995), Gauging Similarity with n-Grams: Language-Independent Categorization of Text. Science, New Series, Vol. 267, No. 5199, pp. 843-848.

Deerwester S. C., Dumais S. T., Launder T. K., Furnas G. W., Harshman R. A. (1990), Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391.

Gildea D., Hofmann T. (1999), Topic-Based Language Models Using EM. Proceedings of the 6th European Conference on Speech Communication and Technology.

Griffiths T. L., Steyvers M., Tenenbaum J. B. (2007), Topics in Semantic Representation. Psychological Review, Vol. 114, No. 2, pp. 211-244.

Hofmann T. (1999), Probabilistic Latent Semantic Analysis. Proceedings of the Fifteenth Conference of Uncertainty in Artificial Intelligence.

Jiang J. J., Conrath D. W. (1997), Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of Research in Computational Linguistics.

Kakkonen T., Myller N., Sutinen E., Timonen J. (2008), Comparison of dimension reduction methods for automated essay grading. Educational Technology & Society, 11(3), pp. 275-288.

Landauer T. K., Dumais S. T. (1997), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, Vol. 104, No. 2, pp. 211-240.

Landauer T. K., Foltz P. W., Laham D. (1998), An Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259-284.

Lee D. D., Seung H. S. (2001), Algorithms for Non-negative Matrix Factorization. Advances in Neural Information Processing Systems, vol. 13.

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. (2011), Optimizing Semantic Coherence in Topic Models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262-272.

Newman D., Lau J. H., Grieser K., Baldwin T. (2010), Automatic evaluation of topic coherence. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Niraula N. B., Banjade R., Stefanescu D., Rus V. (2013), Experiments with semantic similarity measures based on lda and lsa. Statistical Language and Speech Processing, Springer, pp. 188-199.

Paik J. H. (2013), A Novel TF-IDF Weighting Scheme for Effective Ranking. Proceedings of th 36th international acm sigir conference on research and development in information retrieval, pp. 343-352.

Ramage D., Hall D., Nallapati R., Manning C.D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi labeled corpora. Procceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248-256.

Ramos J. (2003), Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, Vol. 242, pp. 133-142.

Roder M., Both A., Hinneburg A. (2015), Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399-408.

Rubin T. N., Chambers A., Smyth P., Steyvers M. (2010), Statistical topic models for multi-label document classification. Machine Learning 88, 1, pp. 157-208.

Salton G., Buckley C. (1988), Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5).

Sidorov G., Velasquez F., Stamatanos E., Gelbukh A., Chanona-Hernandez L. (2013), Syntactic Dependency-based N-grams as Classification Features. Lecture Notes in Artificial Intelligence, N 7630, pp. 1-11.

Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D. (2012), Exploring Topic Coherence over many models and many topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952-961.

Tam Y. C., Lane I., Schultz T. (2007), Bilingual LSA-based adaptation for statistical machine translation. Machine Translation, pp. 187-207.

Titov I., McDonald R. (2008), Modelling Online Reviews with Multi-grain Topic Models. Proceedings of International Conference on World Wide Web.

Trstenjak B., Mikac S., Donko D. (2014), KNN with TF-IDF Based Framework for Text Categorization. Procedia Engineering 69, pp. 1356-1364.

Wang X., McCallum A. (2005), A Note on Topical N-grams. Technical Report UM-CS-2005-071, University of Massachusetts.

Wang Q., Peng Z., Jiang F., Li Q. (2013), LSA-PTM: A Propagation-Based Topic Model Using Latent Semantic Analysis on Heterogenous Information Networks. WAIM, vol. 7923, pp. 13-24.

Wei X., Croft W. B. (2006), LDA-Based Document Models for Ad-hoc Retrieval. Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178-185.

Wu H. C., Pong Luk R. W., Wong K. F., Kwok K. L. (2008), Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inform. Syst. 26, 3, Article 13.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl