DOI 10.31489/2022Ec4/15-25
JEL C55, E24
UDC 004:331.5

**A.K. Atabayeva***

*Karaganda University of the name of academician E.A. Buketov, Kazakhstan*
*atabaeva@list.ru*
*https://orcid.org/0000-0002-4644-1843*

Scopus Author ID: 1617942538468
Researcher ID: AAR-3212-2021

# Application of text mining technology for comparative analysis of trends in the labor market

***Abstract***

*Object*: The main purpose of the article is to analyze the employment of the population in the post-Soviet space in the context of world events to identify positive and negative trends in employment, as well as characteristic trajectories of development directions.

*Methods*: Modern methods of word processing, in particular text mining, word cloud were used.

*Findings*: The rapid development of modern technologies, Internet applications is accompanied by the generation of large amounts of data, the timely processing of which is today one of the main problems in various spheres of life - social, economic, political, and others. In solving this global problem, modern methods of processing text information, the so-called text mining technologies, come to the rescue. These tools allow to increase the efficiency of solving problems of different levels. The algorithms embedded in the text mining technology reveal the basic concepts of the text, the content and the relationship between them.

The integration of modern text mining systems and the R-Studio programming language makes it possible to conduct research in the field of text analysis and processing. These systems, using statistical methods, process the rating of news documents, materials of scientific documents, blogs, tweets, emails, advertisements and other information. The main task of text analysis is to get a clear idea about the topics of interest, to extract important information. The analysis of text documents by text mining methods is carried out in several stages: information search, text preprocessing, extraction of the required information, application of text methods, analysis and interpretation of the obtained results. For the analysis of texts, articles in Russian in PDF format were selected, including information on trends in the labor market and employment for 1995–2020.

*Conclusions*: over the past 20 years there have been significant changes in the labor market and employment. Text analysis technologies made it possible to reveal that during the study period, the labor market issues of unemployment, employment, employment transformation, the emergence of new forms of employment, social and gender problems, and others raised.

***Keywords:*** text mining, word cloud, TF-IDF, LDA, employment, labor market.

## *Introduction*

The labor-intensive process of the manual method of text analysis remains far in the past. Huge arrays of text data today can no longer be explored without the use of software. Modern information technologies allow researchers to use computer processing methods and text mining. The rapid development of the Internet makes it possible to extract information resources for data processing from scientific articles, online discussions, websites, chats, user reviews, newspapers, social networks, and other open sources. Therefore, text mining methods are the most relevant and in demand at the present time in various fields of business, politics, education, etc., first of all, for visualizing the content of information taken from the texts of reports, speeches and reviews. For example, in the US Department of Health, word clouds have been used to analyze the content of documents to determine if sufficient attention is being paid to the core activities of the organization (Atenstaedt, 2017).

There are many online tools that generate word clouds. One of the first is Wordle.net. These tools visually display frequently occurring words in the text and serve as a quick way to get a general idea of the information being studied (article text, speaker's speech, blog or database posts, respondents' online responses, comments, and others). In some cases, word clouds can reveal specific features in the data that prompt further, deeper exploration. It should also be noted that text analysis has some disadvantages, which are the reason for the rare use of this method in the analysis of scientific articles.

* Corresponding author. E-mail address: *atabaeva@list.ru*

***Literature review***

Recently, most research has been carried out using new technologies in the field of artificial intelligence, machine learning, etc. Of particular interest are the so-called text analysis technologies, since the analysis of patterns and trends is a huge task. Therefore, text mining is widely researched today. Text Mining extracts relevant knowledge from text documents. Various text mining methods convert unstructured data into structured data. Text classification, one of the basic principles of text analysis, requires a number of text processing techniques, the most important of which is natural language processing (NLP) (Udgave, Kulkarni, 2020).

Numerous research papers are published online. The growth in the development of computer and information technology makes it difficult for users to find and classify interesting scientific articles on a specific subject (Cai, Luo, Wang, Yang, 2018). Therefore, it is desirable that there be a mechanism by which scientific papers are systematically classified according to similar topics. This will allow users to quickly and easily find research papers of interest to them. As a rule, searching for research papers on specific topics or subjects takes a long time. For example, researchers usually spend a lot of time on the Internet to find articles of interest to them. The required information is not retrieved effectively due to the fact that the articles are not grouped by topic or there is no access to the necessary information (Bolshakov et al., 2017).

Today, owing to big data technologies, this problem is completely solved. Modern possibilities of analysis, classification and processing of a huge number of research papers make this work efficient, manageable and accessible. The use of automated processing methods every year an increasing number of scientific papers come to the aid of researchers. They allow to describe the essence of the article, catch the direction of the research and a summary before reading the content in the main body of the article. In this regard, the keywords of scientific papers should be written concisely and informatively (Kalabin, Korneeva, 2020).

To classify a huge number of articles into articles of similar subjects, scientists Kim S. and Gil J. (2019) propose to use an article classification system based on term-frequency - inverse document frequency (TF-IDF) schemes and Latent Dirichlet Allocation (LDA) schemes. The proposed system firstly creates a representative keyword dictionary with the keywords that the user enters and with the topics extracted by the LDA. Second, it uses the TF-IDF schema to extract topic words from article abstracts based on a keyword dictionary.

Experimental results show that the proposed system can well classify entire articles with similar topics by keyword ratio. The classification system based on the TF-IDF and LDA schemes is widely used, as it is quite effective (Nguyen, 2019).

Word cloud technologies are also in high demand. Word clouds are an image made up of words that together resemble a cloud shape. The size of a word shows how important it is, e.g., how often it appears in the text - its frequency. People typically use word clouds to easily create summaries of large documents (reports, speeches), create art on a topic (gifts, exhibitions), or visualize data (tables, surveys) (Turner, 2017).

Modernization of modern methods of data processing requires the search for effective ways to enhance the process of using this tool. Atenstaedt (2012) in his research reveals the features and applications of the clouds, which contribute to a more in-depth study of these technologies.

Word cloud is a resource that allows you to create a visual image of keywords, text in an attractive form. There are special programs that generate a cloud by displaying the most frequently used words in large print, for this it is enough to enter text or URL (website address) in a special field. Techniques for working with the word cloud are unusual and useful for those who perceive most of the information with the help of vision. On the one hand, this is just an opportunity to create a beautiful picture for a report or presentation. On the other hand, it is a useful tool with many interesting applications (Ramsden & Bate, 2008).

***Methods***

In the process of research, modern methods of text processing were used – text mining, word cloud.

***Results***

To determine the main trends in the labor market and employment problems, a literature review in Russian was conducted for the period from 1990 to 2020 in the R-Studio program. This program allows using special built-in commands to analyze texts, which help to identify relevant thematic issues (Verzani, 2017).

R-Studio is a program that is both a programming language and an environment for statistical computing and graphing (Mark, 2012).

Text mining (TM) is an innovative method of structural text analysis, which represents a broad perspective of theoretical approaches for processing input textual information. This method is an interdisciplinary

field of scientific activity at the intersection of data mining, automatic text processing, descriptive statistics and informatics.

To analyze text information in the R program, the "tm" (text mining) package is used, which is installed using the install.packages("tm") command. First of all, the so-called "Corpus" is created. A corpus is an object that includes all analyzed texts. A variety of operations can be performed with a text corpus, such as representing all words in capital letters (tolower), removing punctuation marks (removePunctuation), removing extra spaces, and others (Kabakof, 2015).

The general stages of text data processing are: data cleaning; lemmatization; stemming.

Data cleaning includes removing numeric data, spaces, replacing uppercase letters with lowercase ones. Also, stage 1 removes "stop words" or they are also called "noise words". That is, words that on their own do not carry any semantic load (too frequent, too rare, too short, non-nouns, proper names). These include prepositions, suffixes, participles, interjections, numbers, particles, conjunctions. For example, "not", "also", "these", "either", "among", "always" and others.

Stemming is the process of finding a word stem for a given source word (cutting a word to a stem). In the process of stemming, endings are discarded from words. Stemming is based on the rules of language morphology. Thus, stemming cuts endings and suffixes from the word so that the remaining part is the same for all grammatical forms of the word.

Lemmatization is the process of defining the lemma of a word. Lemma is the original, basic form of the word. For nouns and adjectives, it is the singular form of the nominative case, and for verbs, it is the infinitive.

To analyze the texts, articles in PDF format were selected, including information on trends in the labor market and employment for 1995, 2000, 2005, 2010, 2015 and 2020. The analysis includes the following commands:

1) Creation of a database that includes the analyzed PDF files. To perform the analytical part of our task, the first thing to do is to create a PDF database or corpus. The corpus is a database of words. Six PDF-format documents are loaded into the corpus. Therefore, we create a database consisting of documents in PDF format. We upload all PDF files and perform a document upload pre-check to make sure that all required documents are uploaded to the database.

2) Cleaning the case from the so-called noise (tm command). A number of transformations are performed: we change all text to lower case, remove numbers, stop words, all punctuation marks, and spaces. These operations really tidy up and structure our documents so that text can be parsed; in other words, changing text from unstructured format to structured format.

3) In our case, we do not want the program to abbreviate words, so we set stemming to FALSE.

4) Checking frequently occurring words in all documents and determining their number (Fig. 1).

```
> inspect(opinions.tdm[1:10,])#Examine 10 words at a time in across documents
<<TermDocumentMatrix (terms: 10, documents: 6)>>
Non-/sparse entries: 60/0
Sparsity           : 0%
Maximal term length: 11
Weighting          : term frequency (tf)
Sample             :
            Docs
Terms        1995.pdf 2000.pdf 2005.pdf 2010.pdf 2015.pdf 2020.pdf
безработица        40      103       44       23        5        8
бизнес              1        1        6       15        1        6
доходы              2       58       13        9        3        5
занятость          60      119       63      112      108       45
занятых             1       16       33        4        1        2
изменения           1        9       14        4        2        3
людей               1        4        1       10        1        1
места               5        5        2        4        2        5
новые               2        6        1        1        1        5
новых               2        3        2        6        5        9
```

Figure 1. Determination of the number of repetitions of the first 10 frequently occurring words

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

For text analysis, a command is installed that checks for the presence of the first 10 words that appear in all documents and determines the total amount of repetitions of these words in each document. For example,

the word “income” appears twice in the first document, 58 times in the second, 13 times in the third, 9 in the fourth, 3 in the fifth, and 5 times in the sixth. As can be seen from Figure 1, the main topic of the study is employment issues, the frequency of which is the highest in all the studied files. The second largest is the problem of unemployment. “Unemployment” was the most frequently discussed in 2000, and in subsequent periods this problem was raised less and less. This fact confirms the situation on the labor market during the study period. According to Kazakh official statistics, the highest unemployment rate was 10.4% in 2000. This is followed by a gradual decline to 4.9% in 2020 (BNSASPR, 2020). The data in Figure 1 confirm this fact: the largest mention of the word “Unemployment” is observed in 2000 - 103 times, the smallest in 2015 (5 times) and in 2020 (8 times).

Also common to all documents is the word “change”, which implies transformational processes in the labor market. In the documents of 2000 and 2005, it occurs the most times, 9 and 14, respectively. It was during these periods that mass computerization took place, informatization of all spheres of activity, which significantly affected the employment of the population. The emergence of new industries (IT, services), the digitalization of society led to the emergence of new forms of employment, which is reflected in Figure 1: the word “new” is most common and discussed in 2000 and 2020 (9 and 14, respectively).

If we consider the next 10 most popular words, we note that some of the most common words are: “problems”, “work”, and “market” (Fig. 2).

```
> inspect(opinions.tdm[11:20,])#Examine 10 words at a time in across documents
<<TermDocumentMatrix (terms: 10, documents: 6)>>
Non-/sparse entries: 60/0
Sparsity           : 0%
Maximal term length: 10
Weighting          : term frequency (tf)
Sample             :
           Docs
Terms       1995.pdf 2000.pdf 2005.pdf 2010.pdf 2015.pdf 2020.pdf
  переход          3        2        2        2        2        1
  проблема         2        2        5        4        2        4
  проблемы         5       91       64       23       11       15
  работа          49       29        9       45       18       16
  рабочих          4       41       34       22       13       13
  рост             1        8        6        5        2        4
  рынок           36       49       42       25       30       26
  связи            3        7        5        6        2        1
  социальных       3       26        2        2        1        3
  сфере            1        9        9        4        3       11
```

Figure 2. Second 10 frequently occurring words

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

The word “problems” was mentioned 91 times in 2000, which indicates the most difficult period for the labor market (in terms of employment, advanced training, social protection, etc.). This period is characterized by acute social problems (the word “social” occurs 26 times).

Consider terminology that appears at least 20 times in all 6 documents. That is, if we previously analyzed the frequency of use of terms for each individual document, now we will analyze all 6 documents as a whole (Fig. 3).

```
> findFreqTerms(opinions.tdm,lowfreq = 20,highfreq = Inf)#Frequent terms that appear at least 20 times across all documents
 [1] "безработица" "бизнес"      "доходы"      "занятость"   "занятых"     "изменения"   "места"       "новых"       "проблемы"
[10] "работа"      "рабочих"     "рост"        "рынок"       "связи"       "социальных"  "сфере"       "труда"       "трудовых"
[19] "уровень"     "условиях"    "экономики"   "экономического"
```

Figure 3. Terms appearing in all documents at least 20 times

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

One can also consider the total amount of frequently occurring words in documents (Fig. 4).

```
> ft.tdm<-as.matrix(opinions.tdm[ft,])#Sum the count of all frequently occurring words
> sort(apply(ft.tdm, 1, sum), decreasing = TRUE)
занятость     труда безработица  проблемы     рынок     работа   рабочих  экономики    доходы  трудовых  условиях
      507       303       223       209       208       166       127       115        90        71        62
  занятых   уровень социальных     сфере изменения   бизнес    новых рост экономического     связи    места
       57        55        37        37        33        30        27        26        26        24        23
```

Figure 4. Sum of frequently occurring words in documents

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

As Figure 4 shows, the maximum number of words "employment" – 507, "labor" – 303, "unemployment" – 223, and others, reflects the general focus of the subject under study. Based on the analysis carried out, it is possible to note transformational processes in the labor market, as well as to identify the following pronounced problems: incomes of the population, social protection of the population, new working conditions, introduction of new forms of employment, and others.

Thematic modeling.

Let us apply one of the topic modeling methods based on a specific algorithm called LDA (Latent Dirichlet Allocation). It is a mathematical model of a language that captures topics (lists of similar words) and how they cover various texts. By scanning and understanding the importance of words in the text, this algorithm can evaluate what is contained in the text (review), what the reviewer thinks on various topics that are weighted and interconnected. Also, the LDA function monitors which words appear next to others in texts and reviews. This information is captured using probability statistics, which is a deeply mathematical process.

The application of this text mining method is that several documents can be grouped by topic. That is, documents similar to each other are grouped. The following libraries are loaded for this task: tm, a tool for working with PDF files, ggplot, and dplyr.

The next step is to create a document matrix. This is required for data modeling. Since all analyzed texts must be presented in the form of a matrix of document terms. To make this transformation, the corpus we have already created, which is a document, is taken as the initial unit. We place it in the matrix function of the document term "DocumentTermMatrix(document)" and save it as a DTM variable. Now we create our actual model. The first step for topic modeling is to create a model using the "lda" function, where lda is short for Latent Untargeted Allocation. So we use the LDA function and pass in the name of our document matrix (DTM), k equal to 6 says we have 6 documents and we set the initial value so that every time we run the function we get the same results.

When creating a model, there are specific things that we are interested in. First of all, these are beta values, which are part of our model (Fig. 5). Therefore, we create a new variable, which we call "beta_topics" and create a "tidy" function, to which we pass our model "LDA".

```
> #Shows the probability of a word being associated to a topic
> beta_topics<-tidy(Model_lda, matrix = "beta")#create the beta model
> beta_topics#shows all the information in beta_topics
# A tibble: 54,594 × 3
   topic term             beta
   <int> <chr>           <dbl>
 1     1 суволенные   1.14e- 3
 2     2 суволенные   2.02e-67
 3     3 суволенные   4.90e-64
 4     4 суволенные   1.92e-58
 5     5 суволенные   6.77e-66
 6     6 суволенные   4.57e-66
 7     1 пбезработица 1.14e- 3
 8     2 пбезработица 3.26e- 4
 9     3 пбезработица 1.26e-62
10     4 пбезработица 2.10e-57
```

Figure 5. Beta values

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

Figure 5 shows that the terms "laid off" and "unemployment" are most related to the topic under study. In other words, they are part of all 6 documents, and the beta value shows a quantitative relationship with the text. The highest beta in 2015 is 6.77 e-66. The lowest score was in 1990. This means that the higher the beta

value, the stronger the connection of the term with the analyzed topic. Therefore, it can be argued that the problems of layoffs were especially acute in 2015 and 2000 (beta coefficient = 4.9 e-64), and unemployment problems in 1995 (beta coefficient = 3.26 e-4).

By examining the beta values to see which terms or words make up each topic, one can display this visually as a series of graphs (Fig. 6). To do this, one needs to make sure that all terms that are frequently repeated in documents are grouped based on beta values. The chart displays the groups of terms most frequently used in each document. It should be noted that practically in all documents the dominant words are employment, labor market, work, and unemployment.

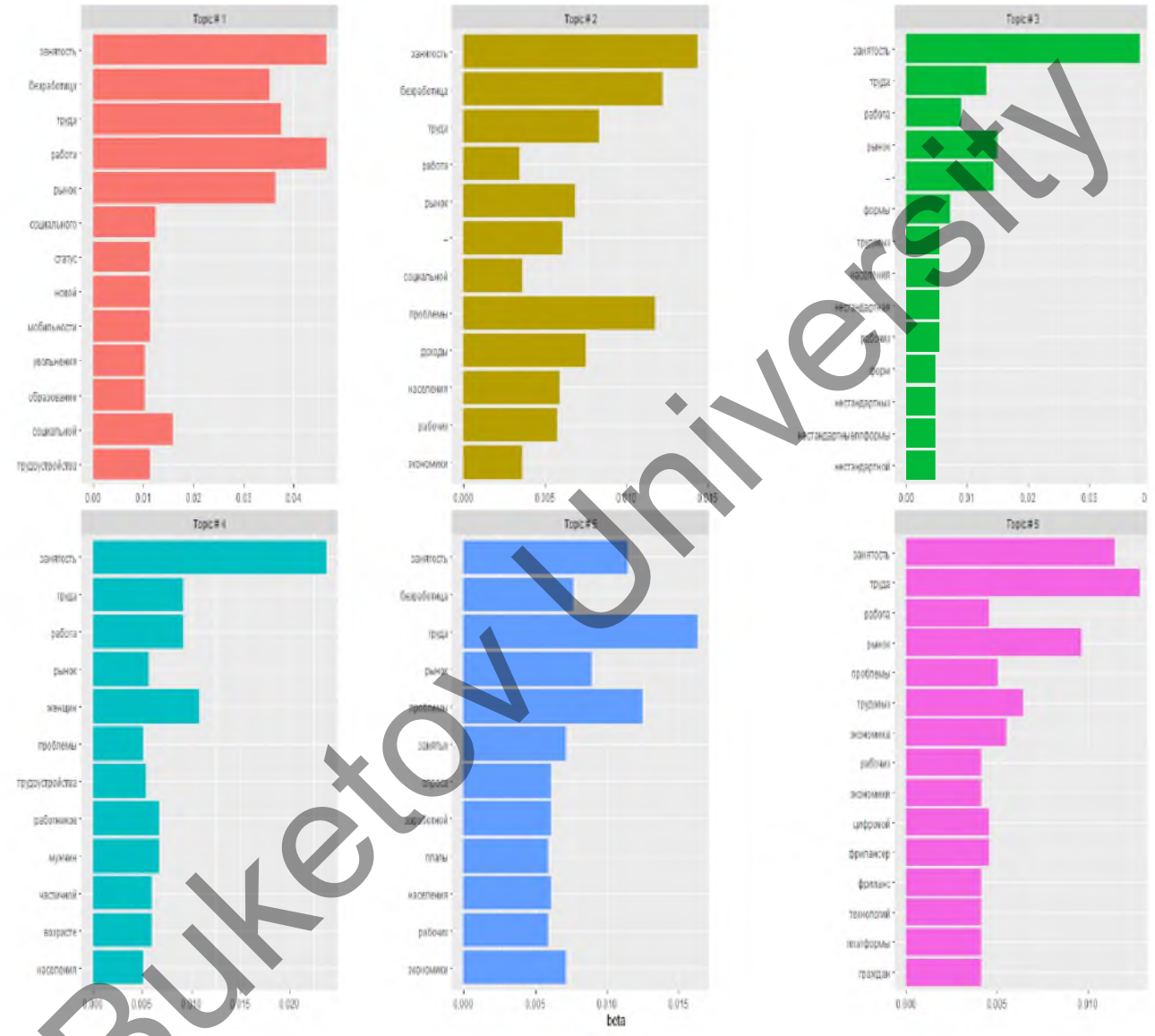


Figure 6. Grouping terms based on beta values

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

In 1995, social problems were raised in the labor and employment market, the problems of layoffs of workers, mobility, education, and employment. The year 2000 also focuses on the problems of employment, to which the problems of income and the economy as a whole are added. In 2005, transformational processes are observed in the employment market, as there is a prevailing use of such terms as non-standard forms of employment. The next period (2010) is focused on gender issues of the labor market, the problems of women's employment and age restrictions are raised. As a solution, the introduction of part-time employment is proposed. The year 2015 again raises the problems of unemployment, which intersect with the problems of workers' wages. In the documents of 2020, much attention is paid to the impact of digital technologies, labor platforms on employment. The consequence of this influence is the emergence of new forms of employment, the words freelance and freelancer are especially often mentioned.

Word cloud.

The word cloud tool (function) shows a random display of all words in the text source, where the size of each word is proportional to the number of repetitions in the text.

A word cloud is a set of frequently occurring words depicted in the same picture in different sizes. The more often a word occurs in the text, the larger it is in the picture. A word cloud is one of the powerful ways to visualize text, which determines the direction of the analyzed information and indicates the main, real trends on the topic under study.

In total, 160 scientific articles by Kazakh authors and authors from the CIS countries were loaded into the program, the key content of which was the topic of employment. To trace the evolution of forms of employment, the list of references was divided into three periods: 1990-2000, 2001-2010, 2011-2020 (Fig. 7).

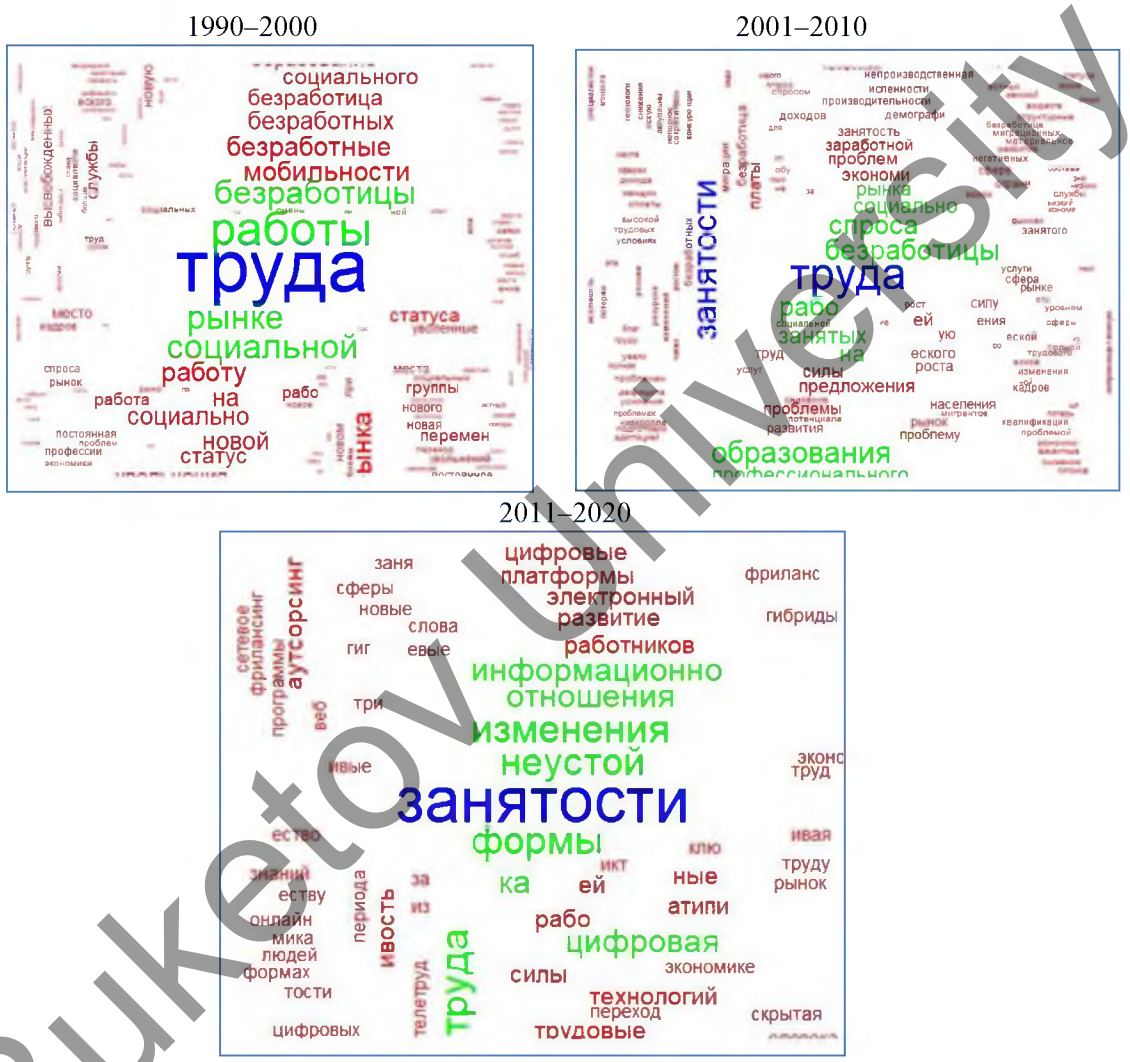


Figure 7. Word cloud for the periods 1990–2000, 2001–2010, 2011–2020

*Note – compiled by the author based on the R program (articles in Russian were used for the analysis)*

The results of the word cloud analysis showed that the following terminology was most often mentioned in the articles:

- in the period 1990-2000 – unemployment, dismissal of workers, shortage of jobs, layoffs, socially new status of an employee, personnel, changes, forced, self-employment, mobility. The transition from a planned economy to a market economy brought with it great changes in the world of work. The closure of many enterprises and factories was accompanied by mass layoffs. In such conditions, people were forced to agree to any work - partial or temporary. To adapt to new working conditions, workers became mobile;

- in the period 2001-2010 – social problems, labor demand, problems of wages and incomes, unemployment, temporary employment, part-time employment, informal employment, labor migration. Despite the stabilization of the economy, there were still unresolved social and income issues. Crisis of 2008-2009 raised the issue of unemployment. This period is characterized by the spread of temporary, part-time, infor-

mal employment;

- in the period 2011–2020 – digital economy, human capital, Internet, new forms of employment, information technology, remote employment, robotization, remote work, education, capital, outsourcing, freelancing. The development of the Internet and information and communication technologies contributes to the emergence of new forms of employment. Scientists are concerned about the consequences of robotization and its impact on the labor market. The problems of the level of education and human capital are being raised, as the requirements for workers in the digital society are increasing.

In Table, we consider the correlation of frequently occurring words with the word "Employment", in order to determine the main trends and directions in the labor market.

Table. Correlation of frequently occurring words with the word "Employment"

| Period | Changes | Social | Status | Problems | Place (work) |
|---|---|---|---|---|---|
| 1990-2000 гг. | 0,99 | 0.87 | 0.84 | 0.81 | 0.72 |
| 2001-2010 гг. | 0.66 | 0.97 | 0.68 | 0.74 | 0.71 |
| 2011-2020 гг | 0.76 | 0.89 | 0.61 | 0.62 | 0.84 |
| *Note – compiled by the author based on the R program* | | | | | |

Analyzing the data in Table, we can conclude that the employment sector has undergone a significant transformation in the first study period (1990–2000). This fact is confirmed by the high values of the correlation coefficients ("Changes" in the field of employment - 0.99). The social sphere (0.87), the status of workers (0.84), jobs (0.72) also underwent strong changes. The transition to a market economy was accompanied by a difficult process of adaptation to new conditions. It took time to build new market labor relations.

In the second decade (2001–2010) there is a slight easing of problematic issues in the field of employment. But the global crisis in the second half of the 2000s caused unemployment and precarious employment, which again exacerbated social tension in the labor market (Social problems - 0.97).

The third decade under study is characterized by changes caused by digitalization processes. The emergence of new remote forms of employment raises concerns about future employment (0.84), social protection of workers (0.89), and the development of human capital.

***Discussions***

With the rapid development of modern technology, new computer and Internet applications are generating large amounts of data at an unprecedented rate, such as video, photo, text, voice, and social media data. This data often has high dimensional characteristics, which poses a major challenge for data analysis and decision making. The right choice of methods shows its effectiveness in processing multidimensional data and increasing the efficiency of the analytical component (Mezentseva, Kolomiiets, 2020).

The choice of features and methods plays an important role in compressing the scale of data processing when redundant and irrelevant features are removed. The feature selection technique can pre-process analysis algorithms, as well as simplify and improve the accuracy of results using the R program (Mastitskii, Shitikov, 2014).

Over the past decade, many companies have been developing special software for processing text information. We note the following of them: Google, IBM, SAS, Angoss Software Corporation, and others. According to Kovtun D.B., the R program is the most accessible to use, since other programs have a number of shortcomings in their work. For example, Google's programs contain restrictions on the analysis of unstructured data, and Google's software is not freely available (Kovtun, 2021).

Topic modeling refers to a wide class of application of machine learning algorithms to text data transformed into a document-term matrix. Topic models are "statistical algorithms aimed at identifying and measuring latent topics within a corpus of text documents". Thematic models are divided into two groups. The first includes documents supposedly having only one theme (single-membership models). Secondly, documents containing several topics (mixed-membership models). Models that assume that each document can have only one topic are implemented, for example, using cluster analysis (k-means, k-medians, etc.). However, models that assume that each document can have many topics have become popular. Currently, there are many such topic models: classical latent Dirichlet placement (LDA), correlated topic models, dynamic topic models, hierarchical topic models and structural topic models (Shipunov et al., 2014).

***Conclusions***

The intellectual analysis of texts made it possible to identify the main trends in the labor market over the past 20 years. Articles of post-Soviet scientists written in Russian were used as sources of analysis.

The main topic of the study is the employment of the population, the frequency of which is the highest in all the studied files. In this regard, some of the most common words in all documents are "problems", "work", "employment", "labor" and "market". The second largest problem is unemployment. This is confirmed by the high beta coefficient in 1995, equal to 3.26 e-4.

Also common and common to all documents is the word "change", which implies transformational processes in the labor market. In the documents of 2000 and 2005, it occurs the most times. It was during these periods that mass computerization took place, informatization of all spheres of activity, which significantly affected the employment of the population. The emergence of new industries (IT, services), the digitalization of society have led to the emergence of new forms of employment, since the word "new" is most often encountered and discussed in 2000 and 2020.

One of the most difficult periods for the labor market was the year 2000, as the word "problems" appears more than 90 times and issues of employment, advanced training, social protection, etc. are raised. This period is characterized by acute social problems.

Based on the analysis carried out, it is possible to note transformational processes in the labor market, as well as to identify the following pronounced problems: incomes of the population, social protection of the population, new working conditions, introduction of new forms of employment, and others.

***References***


Atenstaedt R. Word cloud analysis of the *BJGP*: 5 years on / R. Atenstaedt // British Journal of General Practice. – 2017. - 67(658). – P. 231-232. DOI: https://doi.org/10.3399/bjgp17X690833

Atenstaedt R. Word cloud analysis of the *BJGP* / R. Atenstaedt // British Journal of General Practice. - 2012. - 62(596). – P. 148. DOI: https://doi.org/10.3399/bjgp12X630142

Cai J. Feature selection in machine learning: A new perspective / J. Cai, J. Luo, S. Wang, S. Yang. // Neurocomputing. - 2018. - Vol. 300. – P. 70–79. DOI: 10.1016/j.neucom.2017.11.077.

Kim S. Research paper classification systems based on TF-IDF and LDA schemes / S. Kim, J. Gil // Human-centric Computing and Information Sciences. - 2019. - 9(1). – P.1-21. DOI: 10.1186/s13673-019-0192-

Mark P. J. Learning RStudio for R Statistical Computing / P. J. Mark // Packt Publishing. - 2012. – 126 p.

Mezentseva O. Optimization of analysis and minimization of information losses in text mining / O. Mezentseva, A. Kolomiiets // Herald of Advanced Information Technology. – 2020. - 3(1). – P. 373–382.

Turner D. Word clouds and word frequency analysis in qualitative data / D. Turner. — Quirkos, 2017. [Электронный ресурс]. Режим доступа: https://www.quirkos.com/blog/post/word-clouds-and-word-frequency-analysis-in-qualitative-data/

Ramsden A. *Using Word Clouds in Teaching and Learning* / A. Ramsden, A. Bate // University of Bath, 2008. Retrieved from http://opus.bath.ac.uk/474/

Udgave A. Text Mining and Text Analytics of Research Articles / A. Udgave, P. Kulkarni // Palarch's Journal Of Archaeology Of Egypt/Egyptology. — 2020. — 17(6). — P. 1–7.

Verzani J. Getting started with RStudio / J. Verzani // O'Reilly Media. — 2017. — 98 p.

Большакова Е. И. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пос. / Е.И. Большакова, К. В. Воронцов, Н. В. Лукашевич, А. С. Сапин. — М.: НИУ ВШЭ, 2017. — 268 с.

Бюро национальной статистики Агентства по стратегическому планированию и реформам РК. [Электронный ресурс]. Режим доступа: https://stat.gov.kz/official/industry/25/statistic/5

Кабакоф Р. R в действии. [Электронный ресурс]. Режим доступа: https://www.manning.com.

Калабин А. Л. Анализ информационных критериев отбора значимых признаков в методах Text Mining / А. Л. Калабин, Е. И. Корнеева // Вестн. ВГУ. Сер. Системный анализ и информационные технологии. — 2020. — 2. — С.150–159.

Ковтун Д. Б. Исследование внутриведомственного взаимодействия органов власти РФ на основе документов стратегического планирования с помощью технологии Text Mining / Д. Б. Ковтун // Моск. экон. журн. — 2021. — 2. — С. 1–10.

Мастицкий С. Э. Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. — 2014. — 401 с.

Нгуен М. Т. Тестирование методов машинного обучения в задаче классификации http запросов с применением технологии TFIDF / М. Т. Нгуен // Вестн. Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. — 2019. — № 4. — С. 119–131

Шипунов А. Б. Наглядная статистика. Используем R! / А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников, С. А. Назарова, С. В. Петров, В. Г. Суфиянов. — 2014. — 296 с. [Электронный ресурс]. Режим доступа: http://ashipunov.info/shipunov/school/books/rbook.pdf

**А.К. Атабаева**

## Еңбек нарығындағы тенденцияларды салыстырмалы талдау үшін мәтінді өңдеу технологиясын қолдану

***Аңдатпа:***

*Мақсаты:* Мақаланың негізгі мақсаты халықты жұмыспен қамтудың оң және теріс тенденцияларын, сондай-ақ даму бағыттарына тән траекторияларды анықтау мақсатында әлемдік оқигалар контексінде посткеңестік кеңістіктегі халықтың жұмыспен қамтылуын талдау.

*Әдісі:* Зерттеу барысында мәтінді өңдеудің заманауи әдістері, атап айтқанда *Text mining, Word cloud* қолданылды.

*Қорытынды:* Заманауи технологиялардың, интернет-қосымшалардың қарқынды дамуы деректердің үлкен көлемін генерациялаумен қатар жүреді, оларды уақтылы өңдеу бүгінгі таңда өмірдің әртүрлі салаларында, ягни әлеуметтік, экономикалық, саяси және т.б. негізгі проблемалардың бірі болып табылады. Осы жаһандық мәселені шешуде мәтіндік ақпаратты өңдеудің заманауи әдістері, мәтіндерді интеллектуалды талдау технологиялары көмекке келеді. Бұл құралдар әртүрлі деңгейдегі есептерді шешудің тиімділігін арттыруға мүмкіндік береді. *Text Mining* технологиясына енгізілген алгоритмдер мәтіннің негізгі ұғымдарын, мазмұнын және олардың арасындағы байланысты анықтайды.

Қазіргі мәтін-майнинг жүйесі мен *R-Studio* бағдарламалау тілін біріктіру мәтінді талдау және өңдеу саласында зерттеулер жүргізуге мүмкіндік береді. Бұл жүйелер статистикалық әдістерді пайдалана отырып, жаңалықтар құжаттарының, гылыми құжаттардың материалдарының, блогтардың, твиттердің, электрондық пошталардың, жарнамалардың және басқа ақпараттардың рейтингін өңдейді. Мәтінді талдаудың негізгі міндеті — қызықтыратын тақырыптар туралы нақты түсінік алу, маңызды ақпаратты шығарып алу. Мәтіндік құжаттарды *Text Mining* әдістерімен талдау бірнеше кезеңдерде орындалады: ақпаратты іздеу, мәтіндерді өңдеу, қажетті ақпаратты алу, *Text Mining* әдістерін қолдану, алынган нәтижелерді талдау және интерпретациялау. Мәтіндерге талдау жүргізу үшін 1995-2020 жылдардағы еңбек және жұмыспен қамту нарығындағы үрдістер туралы ақпаратты қамтитын pdf. форматындағы мақалалар таңдалды.

*Тұжырымдама*: Соңғы 20 жылда еңбек нарығында және халықты жұмыспен қамтуда елеулі өзгерістер болды. Мәтінді талдау технологиялары зерттеу кезеңінде еңбек нарығында жұмыссыздық, жұмысқа орналасу, жұмыспен қамтуды трансформациялау, жұмыспен қамтудың жаңа нысандарының пайда болуы, әлеуметтік және гендерлік проблемалар және т.б. мәселелер көтерілгенін анықтауға мүмкіндік берді.

***Кілт сөздер:*** мәтінді өңдеу, Word cloud, TF-IDF, LDA, жұмыспен қамту, еңбек нарығы.

**А.К. Атабаева**

## Применение технологии *Text Mining* для сравнительного анализа тенденций на рынке труда

***Аннотация***

*Цель:* Основной целью статьи является анализ занятости населения на постсоветском пространстве в контексте мировых событий для идентификации позитивных и негативных тенденций в сфере занятости, а также характерных траекторий направлений развития.

*Методы:* В процессе исследования использовались современные методы обработки текстов, в частности, Text mining, Word cloud.

*Результаты:* Стремительное развитие современных технологий, интернет-приложений сопровождается генерацией больших объемов данных, своевременная обработка которых является на сегодняшний день одной из главных проблем различных сфер жизни — социальных, экономических, политических и др. В решении данной глобальной проблемы на помощь приходят современные методы обработки текстовой информации, так называемые технологии интеллектуального анализа текстов. Данные инструменты позволяют повысить эффективность решения разного уровня задач. Алгоритмы, заложенные в технологии *Text Mining,* выявляют основные понятия текста, содержание и взаимосвязи между ними.

Интеграция современных систем текст-майнинга и языка программирования *R-Studio* дает возможность проводить исследования в области анализа и переработки текста. Данные системы с помощью статистических методов обрабатывают рейтинг новостных документов, материалы научных документов, блогов, твитов, электронных писем, рекламы и другую информацию. Основной задачей анализа текста является получение четкого представления об интересующих темах, извлечение важной информации. Анализ текстовых документов методами *Text Mining* выполняется в несколько этапов: поиск информации, предобработка текстов, извлечение требуемой информации, применение методов *Text Mining*, анализ и интерпретация полученных результатов. Для проведения анализа текстов отобраны статьи в формате *pdf,* включающие информацию о тенденциях на рынке труда и занятости за 1995–2020 годы.

*Выводы:* За последние 20 лет произошли значительные изменения на рынке труда и занятости. Техноло-

гии анализа текстов позволили выявить, что на протяжении исследуемого периода на рынке труда поднимались вопросы безработицы, трудоустройства, трансформации занятости, появление новых форм занятости, социальные, гендерные проблемы и др.

***Ключевые слова:*** анализ текстов, Word cloud, TF-IDF, LDA, занятость, рынок труда, Text Mining, новые формы занятости.

### *References*

Atenstaedt, R. (2017). Word cloud analysis of the *BJGP*: 5 years on. *British Journal of General Practice,* 67 (658), 231-232. https://doi.org/10.3399/bjgp17X690833

Atenstaedt, R. (2012). Word cloud analysis of the *BJGP*. *British Journal of General Practice*, 62 (596), 148. https://doi.org/10.3399/bjgp12X630142

Bolshakova, E. I., Voroncov, K. V., Lukashevich, N. V., & Sapin, A. S. (2017). Avtomaticheskaia obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic text processing in natural language and data analysis]. Moscow: Natsionalnyi issledovatelskii universitet "Vysshaia shkola ekonomiki" [in Russian].

Buro natsionalnoi statistiki Agenstva po strategicheskomu planirovaniiu i reformam RK [Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan]. Retrieved from https://stat.gov.kz/official/industry/25/statistic/5[in Russian].

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing,* 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Kabakof, R. (2015). R v deistvii [R in action]. Retrieved from https://www.manning.com [in Russian].

Kalabin, A. L., & Korneeva, E. I. (2020). Analiz informatsionnykh kriteriev otbora znachimykh priznakov v metodakh Text Mining [Analysis of information criteria for selecting significant features in text mining methods]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriia Sistemnyi analiz i informatsionnye tekhnologii [Proceedings of Voronezh State University. System analysis and information technologies Series], 2, 150–159 [in Russian].

Kim, S., & Gil, J. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences,* 9(1), 1-21. https://doi.org/10.1186/s13673-019-0192-7

Kovtun, D. B. (2021). Issledovanie vnutrivedomstvennogo vzaimodeistviia organov vlasti RF na osnove dokumentov strategicheskogo planirovaniia s pomoshchiu tekhnologii Text Mining [The study of intradepartmental interaction of the authorities of the Russian Federation on the basis of strategic planning documents using Text Mining technology]. Moskovskii ekonomicheskii zhurnal [Moscow Economic Journal], 2, 1–10 [in Russian].

Mark, P. J. (2012). Learning RStudio for R Statistical Computing. Packt Publishing. Mastitskii, S. E., & Shitikov, V. K. (2014). Statisticheskii analiz i vizualizatsiia dannykh s pomoshchiu R [Statistical analysis and data visualization with R] [in Russian].

Mezentseva, O., & Kolomiiets, A. (2020). Optimization of analysis and minimization of information losses in text mining. *Herald of Advanced Information Technology*, 3(1), 373–382.

Nguyen, M. T. (2019). Testirovanie metodov mashinnogo obucheniia v zadache klassifikatsii *http* zaprosov s primeneniem tekhnologii TFIDF [Machine Learning methods testing within http requests classification problem with the use of TFIDF algorithm]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriia Sistemnyi analiz i informatsionnye tekhnologii [Proceedings of Voronezh State University. System analysis and information technologies Series]*, 4, 119–131 [in Russian].

Ramsden, A., & Bate, A. (2008). *Using Word Clouds in Teaching and Learning*. University of Bath. Retrieved from http://opus.bath.ac.uk/474/.

Shipunov, A. B., Baldin, E. M., Volkova, P. A., Korobeinikov, A. I., Nazarova, S. A., Petrov, S. V., & Sufiyanov, V. G. (2014). Nagliadnaia statistika. Ispolzuem R! [Visual statistics. Let's use R!]. Retrieved from http://ashipunov.info/shipunov/school/books/rbook.pdf [in Russian].

Turner, D. (2017). Word clouds and word frequency analysis in qualitative data. *Quirkos.* Retrieved from https://www.quirkos.com/blog/post/word-clouds-and-word-frequency-analysis-in-qualitative-data/.

Udgave, A., & Kulkarni, P. (2020). Text Mining and Text Analytics of Research Articles. *Palarch's Journal Of Archaeology Of Egypt/Egyptology*, 17(6), 1–7.

Verzani, J. (2017). Getting started with RStudio. *O'Reilly Media*, 98.