DOI: 10.17323/2587-814X.2020.2.64.83

Demand for skills on the labor market in the IT sector

Andrei A. Ternikov
E-mail: aternikov@hse.ru

Ekaterina A. Aleksandrova
E-mail: ea.aleksandrova@hse.ru

National Research University Higher School of Economics
Address: 3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia

Abstract

One of the most dynamically changing parts of the labor market relates to information technologies. Skillsets demanded by employers in this sphere vary across different industries, organizations and even certain vacancies. The educational system in the most cases lags behind such changes, so that obsolete skillsets are being taught. This article proposes an algorithm of skillsets identification that allows us to extract skills that are needed by companies from different occupational groups in the information technologies sector. Using the unstructured online-vacancies database for the Russian regional labor market, skills are extracted and unified with the use of TF-IDF and *n*-grams approaches. As a result, key specific skillsets for various occupations are found. The proposed algorithm allows us to identify and standardize key skills which might be applicable to create a system of Russian classification for occupations and skills. In addition, the algorithm allows us to provide lists of the key combinations of skills that are in high demand among companies inside each particular occupation.

Key words: job vacancies in IT sector; online vacancies; unstructured data analysis; labor market; demand on skills of job candidates; combinations of skillsets.

Citation: Ternikov A.A., Aleksandrova E.A. (2020) Demand for skills on the labor market in the IT sector. *Business Informatics*, vol. 14, no 2, pp. 64–83. DOI: 10.17323/2587-814X.2020.2.64.83

Introduction

The process of employment in the labor market involves several parties: employers, employees, the educational system and state authorities. One of the most informative indicators for demand assessment is skills, which provide extensive information about competences and abilities demanded from the potential job candidate. However, sets of such skills are dynamically changing in different industries, organizations and even certain vacancies. These changes are connected to economic system fluctuations and labor market restructuring. Moreover, the professional standards that are formed with the help of the educational system become obsolete and inflexible to such changes. The particular interest of this study relates to the problem of identifying key skillsets on the labor market for occupational groups in information technologies (IT).

Several authors highlight issues of skills determination in the IT sphere. Firstly, this branch of the labor market has high volatility of technical and soft skills required, and must be analyzed in a time perspective [1–8]. Secondly, skills, especially technical skills, have an outstanding structure due to the precise formulation of programming languages, technological stack, interface instruments, etc., so that it is easier to classify them in attribution to several job positions [9–11]. Thirdly, the adoption of new technologies requires changing combinations of skills of workers in order to perform newly created tasks [12–16].

IT has penetrated a large part of the labor market. Technical specialists with certain sets of competences and knowledge are hired in spheres of economics and finance, public management, retail industry, etc. Thus, such specialists are also required to be competent in the professional activities of a particular company. Unique combinations of skills in certain areas can be formed in the education system not only in IT specialties. The rethinking of educational policy regarding the formation of skills cannot be separated from the demand from the labor market, which needs effective tools for identifying combinations of professional skills that are required by employers.

The present work concentrates on the creation of the algorithm of key skillsets, determining the particular occupational group, extraction in the IT sphere. The main question is: which skills are needed by companies from different occupational groups in the IT sector.

The paper is structured as follows. The first section contains the overview of related works and methods that are used for classification and clusterization purposes of online labor market data. The second section relates to the main algorithm representation which allows us to extract and classify information from an unstructured job advertisements database and its implementation to real data obtained for local labor market. The third section provides results of the work in terms of proposed key skills and combinations relating to different occupations of the IT sector of the labor market. Finally, in the concluding remarks we discuss the theoretical and practical implementation of the proposed algorithm and extended results are presented in the last section.

1. Related works and methods of skill demand analysis

Many researchers create various algorithms in order to extract information about occupations and particular skills from online job advertisement databases [17–25]. Such sources provide an extensive amount of information about the labor market. However, this data is, in general, unstructured. The main methods of processing this information are based on Natural Language Processing (NLP) techniques such as TF-IDF (Term Frequency – Inverse Document Frequency) matrices, *n*-grams (contiguous sequence of *n* items), classification techniques based on manual mark-up of data

sample (LDA, KNN, SVM, etc.), clusterization of data [17–25].

Online vacancies databases, in general, have unstructured text fields that contain information about an occupation and competences required. However, such fields are manually filled by the companies' representatives, and that demands that data preparation procedures and algorithmic techniques be implemented in order to extract the appropriate information in standardized form. Some research papers try to resolve the classification task of how to match job titles and job descriptions from online advertisements with widely used classifications of occupations and skills such as ISCO[1], ESCO[2] and O*NET[3] [17–20]. Others implement classification models on the basis of expert markups [2, 4, 11, 13, 21]. In other words, the sample portion of data is analyzed and labelled by the domain experts and, then, this information is used to transfer this knowledge to the whole dataset. In addition, researchers use clusterization approaches for the occupations and skills determination in the data preparation process, thereby formulating the separate groups of jobs and competences after machine-based separation [19, 21–25]. Thus, the combination of different approaches and algorithms of data preparation and standardization allows us to provide the basis for an analytical research of labor market issues. A brief description of data, approaches and pipelines which are used in related works is presented in *Table 1.*

The information presented in the table allows us to summarize and systematize approaches for data organization, its processing and selection of criteria for identifying combinations of skills.

All authors present their algorithms of information extraction and systematization on the basis of online job advertisements. However, the manner of their implementation differs from the stated research task. For example, if the main research objective relates to the process of matching the unstructured text fields from job advertisements with the official classification for occupations and skills [17, 18, 20, 21] classification algorithms are implemented on the basis of finding similar patterns in job title description with the extended text information from official classifications and a significant amount of expert manual mark-up data.

The other approach relates to the data-driven approach where obtained data is manually corrected by domain experts in order to provide the appropriate systematization [19, 22–25]. These works focus on the data preparation part and clusterization algorithms. Despite the difference in research objectives, the common techniques of data preparation and extraction of standardized information are, in general, applied. All authors use the TF-IDF approach and tokenization (including stopword removal and stemming procedure) in order to process a wide amount of unstructured textual information. In addition, *n*-grams are used for more than one-word extraction. As a result, a set of unified patterns of information (e.g. occupations and skills) is obtained. However, the authors do not provide a generalized algorithm for matching different variants of the same pattern notation within the noisy data management process.

The choice of groups of occupations and skills is highly dependent on the official classifications and the volume of data. Thus, the level of such groups' aggregation demands an expert view based on the data characteristics. In general, the data is available for a one-year period and the search for appropriate patterns for the unstructured fields are simplified only for job advertisements published in one language.

[1] International Standard Classification of Occupations, https://www.ilo.org/public/english/bureau/stat/isco/

[2] European Skills/Competences, Qualifications and Occupations, https://ec.europa.eu/esco/portal/home

[3] The Occupational Information Network, https://www.onetonline.org/

Table 1.

Related works on online job advertisements analysis

Main direction	Data				Number of extracted groups		Methods of data processing				Number of manually processed data entries	Similarity indexes for terms' matching	Authors
	Volume	Period	Source(s)	Language	Occupations	Skills (terms)	TF-IDF	*n*-grams	Clusterization	Classification			
Clusterization of occupations	1.460	4 months (April – July 2018)	LinkedIn	English	8	96	+	+	Unweighted Pair Group Mean Average method	–	>900	Jaccard	[24]
	12.849	7 months (July – November 2015 and October – November 2016)	5 sources	English	69	NS*	+	–	Latent Semantic Indexing. Singular Value Decomposition	–	750	Cosine	[19]
Classification of occupations	75.546	27 months (February 2013 – April 2015)	WollyBI	Italian	9	NS	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	57.740	Levenshtein. Jaccard. Sørensen–Dice	[17]
	40.000	1 month (year is not specified)	12 sources	Italian	62	542	+	–	Weighted Word Pairs (WWP) extraction	LinearSVC & Perceptron classifier	412	Levenshtein	[21]
Clusterization of occupations and key skills extraction	2.786	3 months (Fall 2015)	10 sources	English	4	180	+	+	Latent Dirichlet Analysis	–	180	Centrality degree	[22]
	2.638	3 months (May – July 2018)	Indeed. com	English	48	480	+	–	Latent Dirichlet Analysis	–	480	% of occurrences	[23]
	1.050	5 months (July – November 2017)	6 sources	English	2	2.335	+	–	Latent Class Analysis. Singular Value Decomposition	–	NS	VARIMAX Rotation	[25]
Classification of occupations and key skills extraction	6.222	4 months (June – September 2015)	3 sources	Italian	6	NS	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	1.007	Random Forest importance	[20]
	~2 mln	24 months (2016–2017)	WollyBI	Italian	22	8	+	+	–	SVM	NS	Levenshtein. Jaccard. Sørensen–Dice	[18]

*NS stands for "Not specified"

Different research pipelines address different metrics of classification and clusterization model assessment. The point of interest here is how several patterns and terms could be matched with the stated domain. The authors use tokenization for raw texts and *n*-grams for construction of the set of terms. Similarity is found by implementing the Similarity indexes. In the case of preserving the word order the Levenshtein distance is the appropriate measure but if only intersection of common terms is valuable for detecting similarity – the Jaccard index is preferable.

2. Algorithm of skills demand analysis and related data

The proposed algorithm which allows us to conduct skills demand analysis is organized for online job advertisements data. These data are obtained from the open-source Application Programming Interface of HeadHunter[4], the largest Russian online recruitment platform[5]. The typical structure of an online vacancy is presented in *Table 2*.

Along with the presented data structure and methods used in the related works, the main interest of the work is to organize the process of knowledge extraction (groups of occupations and unified skills) from unstructured data. Then the opportunity to determine highly demanded skills and skillsets within occupational groups can be performed in an accurate manner. Despite the use of classification algorithms for aggregation of job occupations in related works, the current dataset has already been codified by the data producer. Thus, we assume at the preliminary stage of analysis that occupational groups already exist. In order to organize the algorithm description, several concepts need to be formalized and simplified for research purposes.

Definition 1. **Online vacancy.** Let I be the set of vacancy identifiers; H is the set of specialization codes; S is the set of key skills in text format. Suppose $V = \{v_1, \ldots, v_n\} : n \in \mathbb{N}$ is the set of vacancies; an online vacancy v is a 6-tuple $v = (i, C, d, p, g, K)$, where $i \in \mathrm{I}$, $C \subseteq H : |C| \in \overline{\{1,6\}}$ is the subset of specialization codes, d is the vacancy published date, p is the text name of vacancy's position, g is the

Table 2.

Structure of a typical HeadHunter online job advertisement

Field	Field type		Description
	Structured	Unstructured	
Vacancy ID	+		Numeric code
Specialization ID	+		Set (from 1 to 6 including) of numeric codes[6]
Published date	+		Long date format
Position Name		+	Text
Job description		+	Text
Key skills		+	Set of texts (30 at maximum: each up to 100 symbols)

[4] HeadHunter API, https://dev.hh.ru

[5] SimilarWeb: websites ranking, https://www.similarweb.com/top-websites/russian-federation/category/jobs-and-career/jobs-and-employment

[6] HeadHunter API: Specializations, https://api.hh.ru/specializations

text description of the vacancy, $K \subseteq S : |K| \leq 30$ is the subset of skills (in text format) for a particular vacancy.

Current research is concentrated on the IT sector and the methodology is tested on the local job market (the city of Saint Petersburg). Online-vacancies from the IT sphere in Saint Petersburg were extracted from 2015 till 2019 (the official HeadHunter classification is used to obtain IT-related vacancies by Specialization ID). Each data entry of a particular vacancy (v) contains of the ID-code of the vacancy (i), HeadHunter specialization codes (C) and the list of skills required for a particular employer (K). The research objectives are concentrated on the process of skills' determination, thus, the portion of data where skills are given in the separate data entry are used. Such a sample consists of 63.869 vacancies from May 2015 till September 2019. Each vacancy includes from 1 to 6 professional area codes (HeadHunter professional area classifier). The distribution along 36 areas (inside the group of IT sphere) is presented in *Table 3*.

Despite the HH classifiers' distribution, some spheres could be deleted and merged in onw bigger subgroup. In accordance with the classification introduced in [20] we define 6 + 1 groups of IT specialists. After deleting vacancies with not a purely IT profile and regrouping, 56.000 observations (vacancies) are obtained. The share of deleted professional areas is 10.2%. The new distribution among the rest of aggregated occupational groups of vacancies and their names are presented in the *Table 4*.

***Definition 2*. Job occupation (occupational group).** Let H be the set of specialization codes (identifiers). Suppose O is the ordered set of aggregated job occupations, a job occupation $o \in O$, is a 2-tuple $o = (L, a)$, where $L \subseteq H$ is the subset of specialization codes attributing the particular aggregated group with text name a.

Table 3.

Distribution of vacancies by HeadHunter specializations in data sample

HeadHunter Specialization ID	Share, %	Name
1.221	20.37	Software Development
1.82	7.99	Engineer
1.9	4.90	Web Engineer
1.89	4.52	Internet
1.10	4.18	Web Master
1.327	3.83	Project Management
1.225	3.42	Sales
1.137	3.21	Marketing
1.272	3.11	System Integration
1.295	3.04	Telecommunication
1.211	2.99	Support, Helpdesk
1.117	2.90	Testing
1.270	2.86	Networks
1.273	2.73	System Administrator
1.25	2.69	Analyst
1.172	2.62	Entry Level, Little Experience
1.50	2.25	ERP
1.400	2.11	SEO
1.536	2.10	CRM Systems
1.474	2.04	Startups
1.359	1.75	E–Commerce
1.116	1.58	Content
1.475	1.51	Video Games Development
1.113	1.50	Consulting, Outsourcing
1.246	1.42	Business Development
1.420	1.39	Database Administrator
1.395	1.04	Banking Software
1.203	1.01	Data Communication and Internet Access
1.110	0.98	IT Security
1.161	0.86	Multimedia
1.296	0.67	Technical Writer
1.274	0.66	Computer Aided Design Systems
1.3	0.64	CTO, CIO, IT Director
1.277	0.61	Mobile, Wireless Technology
1.30	0.34	Art Director
1.232	0.18	Producer

Table 4.

Distribution of vacancies among IT occupational groups

Name	Short name	Share, %	HeadHunter Specialization ID
High–level IT specialists	high	13.10	1.327, 1.272, 1.25, 1.113, 1.3
Low–level IT specialists	low	3.66	1.172, 1.296
Engineering professionals	engineers	16.18	1.82, 1.295, 1.117, 1.277
Software developers	soft	22.67	1.221
Web and multimedia developers	web	20.13	1.9, 1.89, 1.10, 1.400, 1.475, 1.161
Administrators and database designers	admin	19.30	1.211, 1.270, 1.273, 1.50, 1.536, 1.420, 1.395, 1.203, 1.110
Others	others	4.96	1.474, 1.359, 1.274

To simplify the further analysis of key skills extraction for particular occupational groups (O, where $|O| = 7$ for proposed dataset), all data processing is organized on the whole sample during the 5 years of data presented. As an assumption for such aggregation, the relative distribution of vacancies in occupational groups is used (*Figure 1*). Thus, the proportion of data in the sample is relatively the same for each occupational group in a time perspective.

Following the research objectives of the current work, key skills and their combinations should be unified and extracted along the set of vacancies (V). However, before introducing the skills' extraction algorithm, each online vacancy that may relate to several job occupations should be mapped in the new data structure (IT online vacancy).

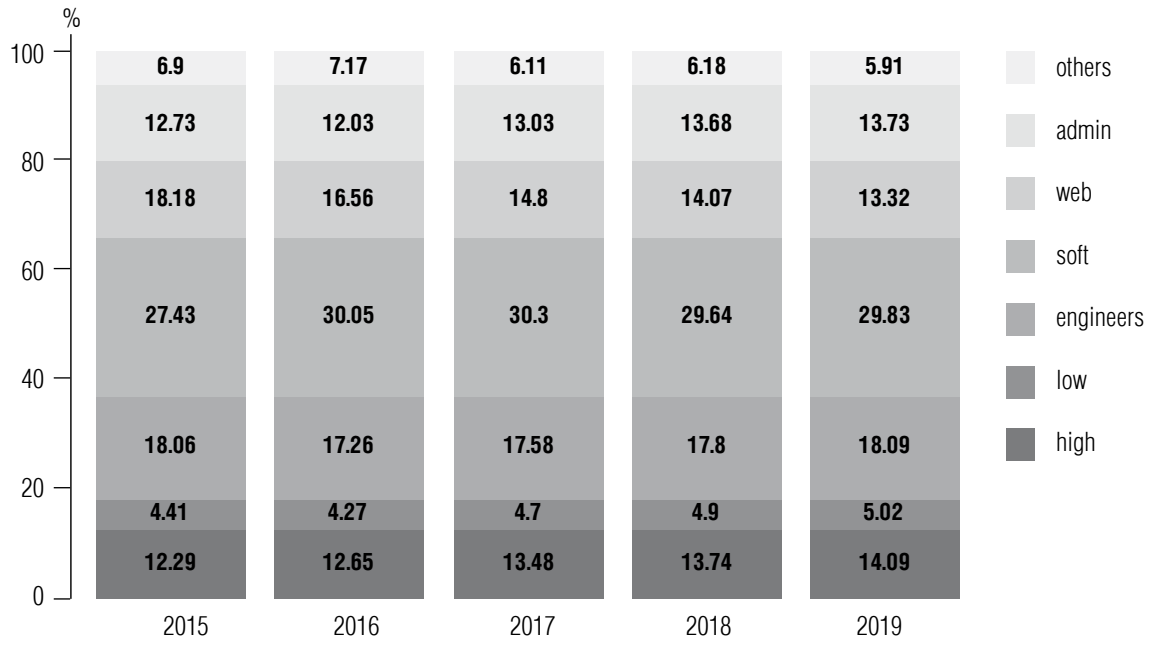

Fig. 1. Distribution of vacancies by occupational groups in time perspective

In order to provide the results of finding key skills and skillset by job occupations, the skill extraction algorithm is organized. Hereinafter, the algorithm of skills' extraction is implemented to IT online vacancies.

Input: IT online vacancies J.
Output: the set of standardized terms (skills) $\tilde{K}$, matched to database of online-vacancies.

1. let $\tilde{S} \subseteq S$ denote the set of unique text descriptions (skills) obtained from J
2. let B denote the set 2-tuples: text description (skill) and its frequency (number of occurrences) in J
3. **foreach** $\tilde{s}_i \in \tilde{S}$ **do**
4. $b_i \leftarrow$ occurrences of $\tilde{s}_i$ in J
5. $B[i] = (\tilde{s}_i, b_i)$
6. **end foreach**
7. sort B in descending order by b
8. **procedure** FrequentTerms(h, t)
9. $\tilde{h} \leftarrow$ subset of h if $h_i > t$, $\forall\ h_i \in h$
10. **return** $\tilde{h}$
11. **end procedure**
12. introduce threshold t
13. $\tilde{B} \leftarrow$ FrequentTerms($B(b)$, t)
14. $T \leftarrow$ 3-tuple set of manually standardized terms $T = (u, x, xs)$, where u denotes identifier of standardized term (skill), x – the name in text format, xs – the set of synonyms in text format for particular pair (u, x)
15. **function** Tokenizer(j)
16. white space normalizer
17. punctuation removal
18. lowercase
19. stemming (English & Russian)
20. stopwords removal (English & Russian)
21. **end function**
22. **procedure** NGrams(J, n)
23. **for** j in J **do**
24. $G \leftarrow n$-grams of size n for Tokenizer(j)
25. add G to ngramterms
26. **end for**
27. **return** ngramterms
28. **end procedure**
29. introduce thresholds t_1, t_2, t_3
30. ngram1:= FrequentTerms(NGrams($B(s)$, 1), t_1)
31. ngram2:= FrequentTerms(NGrams($B(s)$, 2), t_2)
32. ngram3:= FrequentTerms(NGrams($B(s)$, 3), t_3)
33. for obtained databases with n-grams provide manual processing (clear uninformative terms)
34. each entry in these n-grams databases has the set of identifiers
35. **procedure** MatchTerms(X, Y)

let L is the set of unique combinations from X and Y, where $L = \{l_1 | l_1 \in X, l_2 | l_2 \in Y\}$; l_1, l_2 are sets of identifiers $(i, \tilde{s})$
for l_1, l_2 in L **do**
$M \leftarrow$ Jaccard Similarity: $\frac{|l_1(i) \cap l_2(i)|}{|l_1(i) \cup l_2(i)|}$
if >0.5 **do**
add (l_1, l_2) to termsmatched
end if
end for
return termsmatched
end procedure
for each pair of datasets: ngram1, ngram2, ngram3 provide MatchTerms(X, Y) $\rightarrow$ M1, M2, M3
for each pair of datasets T, M1, M2, M3 provide MatchTerms($\tilde{X}, \tilde{Y}$) procedure, where $\tilde{X} := T$, $\tilde{Y} := T$ {M1,M2,M3}
$\tilde{K} \leftarrow X$ left-join Y
$\tilde{K}$ – is output database, with standardized terms, their synonyms, unigrams, bigrams, trigrams

Definition 3. **IT online vacancy.** Let $J \subseteq V$ be the set of IT online vacancies, where $J = \{j_1, ..., j_m\}$: $m \leq n$, $m \in \mathbb{N}$. Let $\mathbf{c}$ denote the classification codes from online vacancy as follows: $\mathbf{c} = (c_1, ..., c_z) \in C$, where $z \leq 6$. Let $\mathcal{L}$ denote the ordered set of labels obtained from job occupations with relation $(L, a) \in O \mapsto \mathcal{L}$ in the following form $\mathcal{L} = (\lambda_1, ..., \lambda_q)$, where $q = |O|$. Introduce the function

$$f(\mathbf{c}, \lambda_q) = \begin{cases} 1, \mathbf{c} \subseteq L_q, \\ 0 \end{cases}$$

that associates the occupational classification codes from job occupation o with codes from aggregated job occupations $\mathcal{L}$. Introduce mapping relation $H: C \mapsto \mathcal{L}$ that provides the multilabel classification and maps the set of aggregated job occupations $\mathcal{L}$ on the basis of the occupational classification codes as follows $\tilde{O} = H(\mathbf{c}) = \{\lambda \in \mathcal{L} \; f(\mathbf{c}, \lambda) = 1\}$, where $\tilde{O}$ is the set of aggregated group names. Thus, an IT online vacancy o is a 3-tuple $j = (i, \tilde{O}, K)$.

In *Table 5*, the distribution of obtained vacancies by aggregated groups (job occupations) is presented. The distribution of job occupations assigned by companies in the database is not homogeneous. In other words, a portion from 6 to 30% in each job occupation is strongly related to the occupation itself. The other major part of vacancies relates to more than one aggregated group. Thus, in the following analysis the diversification of skills that are related to a particular job occupation is needed.

Specifically, assigning the algorithm to the dataset of IT online vacancies the procedure of extracting skills (skillsets) is as follows. In the data sample 13.347 raw unique skills are presented. The descriptions of such skills are not unified in general. In other words, each company can enter its own text string from 1 to 100 characters. For example, one skill's entry may contain one word/phrase or a sentence containing such words separated with punctuation symbols or spaces (no generic format). In order to automate the extraction of certain skills and unify different forms of notation of one term, text mining techniques are used.

According to [20], on the first stage of data preprocessing n-grams (contiguous sequence of items) of words can be constructed. From the vector corpus (TF–IDF) of skills pre-

Table 5.

Distribution of aggregated occupations in data sample

Short name	Number of vacancies	% of non-mixed vacancies by occupational groups
high	19.266	16.28
low	5.383	17.28
engineers	23.787	10.15
soft	33.333	29.78
web	19.312	9.29
admin	20.825	6.18
others	7.293	15.80

sented in vacancies' descriptions, uni-, bi- and tri-grams were constructed with the use of the following tokenizer: removing all punctuation and extra spaces, lowercase to all letters, words' stemming both in English and Russian language, stop words removal. Within these extracted terms, the initial structure and formulation of skills were saved. The first step is the extraction of meaningless words (non-informative itself) and messy data separation from informative patterns. For the unigram database, 348 entries were extracted from 5.234 non-unique terms; for bigram – 577 out of 1.090; for trigram – 110 out of 303. The second step is the creation the database of synonyms for already obtained patterns. HeadHunter API: Suggestions (Key skills suggestions) allows us to obtain a portion of synonyms for manual processing of obtained terms. After such synonym extraction, the term matrix for 707 terms was obtained (1.296 entries for manual checking). The third step is the addition of terms from the Stack Overflow Developer Survey[7] (108 terms for the most popular IT technologies names) and final correction of appropriate terms (database with reference terms and their synonyms).

As a result, 435 standardized terms were obtained within 420 synonyms for them. Such dataset contains both technical (hard) and non-technical (soft) skills for the given sector of the labor market. The last stage is composed through intersection of raw skills (codified with unique identification codes), matched exactly with specific terms obtained from a manually corrected list of HeadHunter synonyms and the results from TF-IDF matrices (for uni-, bi-, tri-grams), resulted within the pairs: skill ID and term. In order to automatically define the closeness between several terms (on the basis of the unique set of IDs for each term) and match the rest of the data with the given standardized terms, the Jaccard distance measure is used. For example, the similarity between two sets of words (terms) $A$and B could be found with the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

[7] Stack Overflow Annual Developer Survey, https://insights.stackoverflow.com/survey/

This measure is appropriate for categorical data closeness comparison and its value is in the range from 0 to 1 inclusive. However, the choice of the cutoff-point highly depends on the data and research objectives. As the threshold for identifying close terms, the level above 0.5 is used (after manual processing of obtained datasets). Thus, 53.672 vacancies (95.8% of the initial sample) contain at least one of standardized skill from the previously obtained dataset of terms and their synonyms. The percentage representation of the Top-20 standardized skills (by the number of occurrences in the sample) among the whole dataset is presented in *Table 6*.

However, following the objectives of the current research, the particular analysis of relevant and highly-demanded (from employer's side) skills (and their combinations) lies behind the determination of relevant skills that are at the same time strongly related to a particular occupational group (specific skills) and supported by a relatively large number of employers.

After the data preparation of the dataset of vacancies, the following data structure is obtained: the dataset of 305.217 observations (particular skill/term from the vacancy), where each observation has the ID of a standardized skill, the ID of the vacancy and the occupational group code. In order to provide the classification of skills in accordance with the stated groups of vacancies, the process of finding pairs and triplets of skills was conducted for each vacancy group. After obtaining the pairs and triplets of skills (non-zero by Jaccard Similarity), the highly matched (threshold by Jaccard Similarity) skills were extracted. The general scheme of the proposed algorithm implemented to the dataset is presented in *Figure 2*.

On the first step, all 435 skills, pairs of them (C^2_{435} = 94.435) and triplets (C^3_{435} = 1.362.345) were used for finding the Jaccard Similarity for each of 7 groups of vacancies (occupations).

Table 6.

Top-20 skills by their occurrence in the sample dataset for the IT sector

Skill Name	% of occurrences in database of skills
HTML/CSS	6.73
JavaScript	4.69
1C	3.48
SQL	3.25
PHP	2.63
Git	2.53
Linux	2.32
Java	2.28
MySQL	1.86
Negotiation skills	1.61
Sales Skills	1.52
Business communication	1.51
English	1.45
Testing Framework	1.43
Python	1.40
jQuery	1.36
C/C++	1.30
OOP	1.29
C#	1.28
.net	1.27

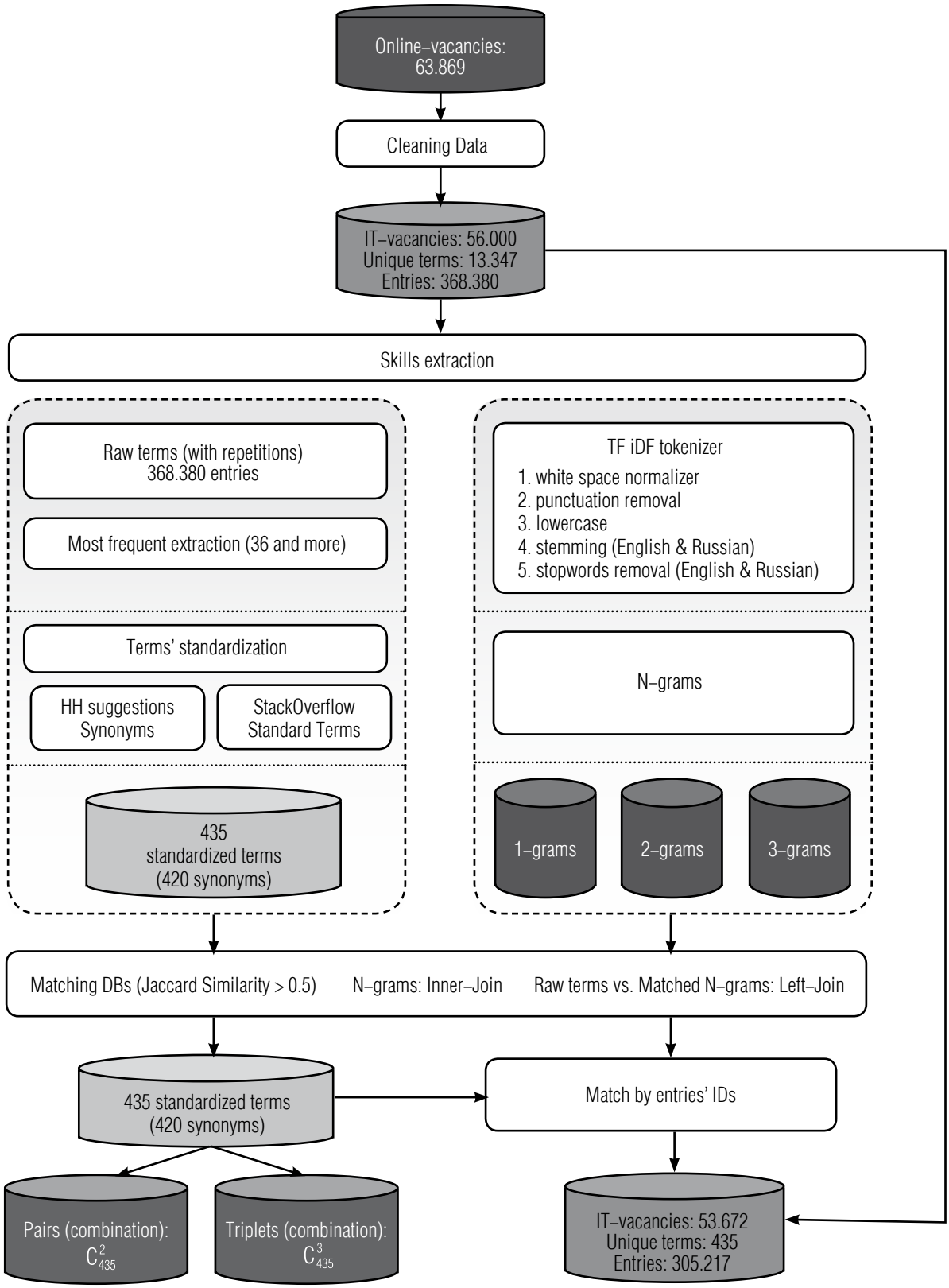

Fig. 2. Baseline algorithm of skills extraction from unstructured database

Secondly, using pairs and triplets of skills (unique combinations without repetitions), each dataset with terms was ranked by Jaccard Similarity (after removing such observations, where Jaccard Similarity equals zero) within their quantiles (permilles: 0.1% step for pairs and triplets in order to produce the variability).

For each pair and triplet of skills, such a measure was calculated on the basis of the number of vacancies that include such combinations. Thirdly, the differences in ranks for each pair of vacancies' groups were found. Fourthly, in order to extract the specific features (set of skills), the outliers in such distribution were found (as a provision for highly diverse skills and skillsets that can describe and separate groups of vacancies. The statistical logic behind this shows that the distribution of ranks' differences is quite close to normal and the detection of outliers (too vast difference in ranks of skills and skillsets) allows us to provide the appropriate selection of skills which can separate different groups of vacancies. For example, several pairs of such groups are represented in *Figure 3*

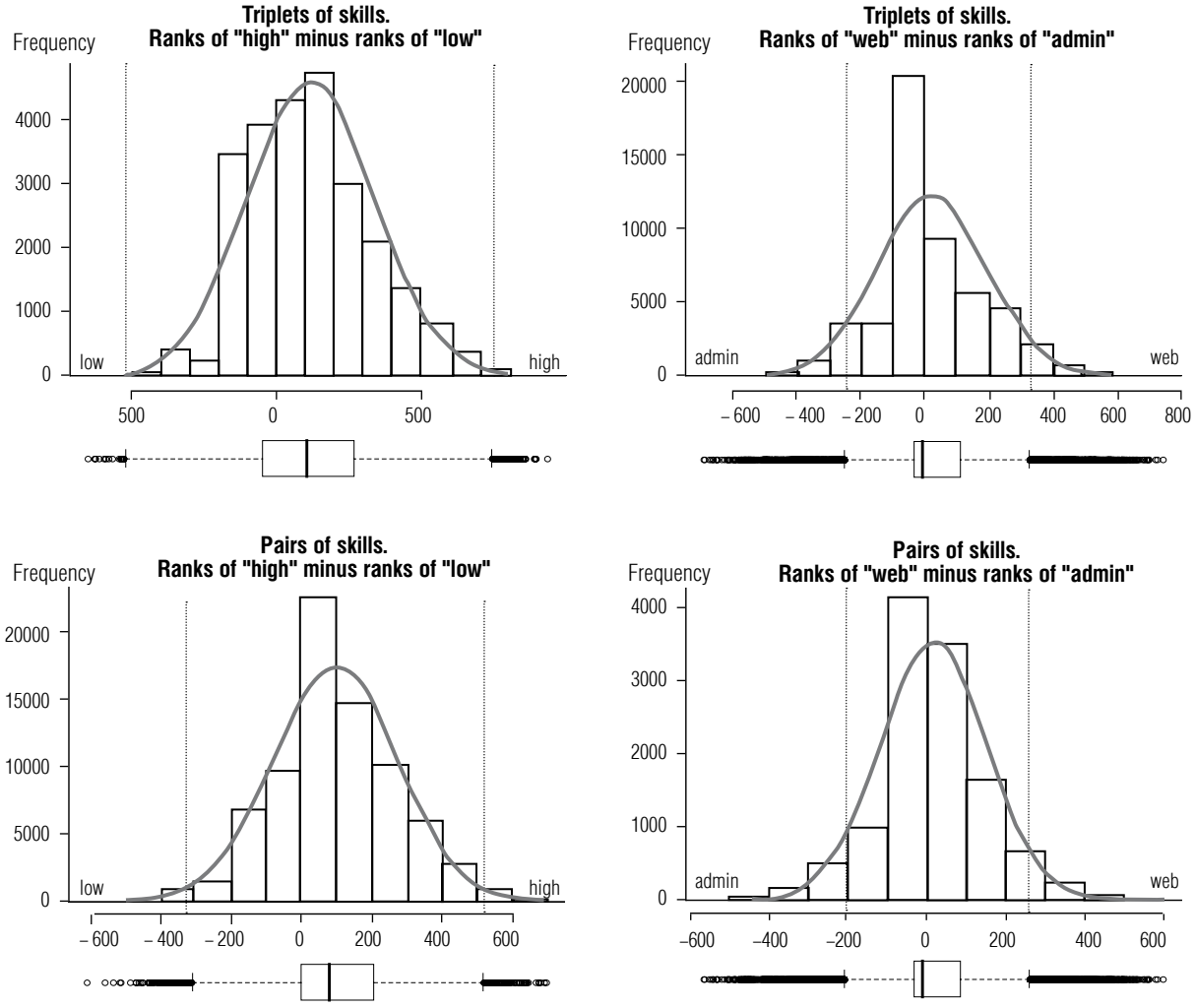

Fig. 3. Distribution of rank differences by Jaccard Similarity for occupational groups

[8] IQR – interquartile range

(cutoffs for skills detection are boundaries of whiskers in boxplot: 1.5 IQR[8] below and upper for relatively appropriate quantile rank difference).

Fifthly, for extracted pairs and triples, the procedure of addition of unique skillsets (different sets of skills) for each pair of an occupation's groups is provided. Thus, in the cross-intersection of skills (technically, with zero value of Jaccard Similarity) only those in above 95% (by quantile difference) are added in order to detect initially key (and different) skillsets. Sixthly, for each pair of vacancies' groups ($C_7^2 = 21$), core and determinant skills (and their combinations) were determined (and skills, which are unique in certain class in the last decile, were added for them unique skills by cross-intersection). Thus, three matrices 7×7 were obtained, where on the cross-section of i-th row and i-th column ($i \neq j$) the unique sets of skills (codified separately for unique skills, their pairs and triplets) are presented (distinguished and unique above 95% threshold skills of i-th group, comparing with j-th group of occupations). Eighthly, such skills were extracted in the following manner: the presence of core skills that determines each occupational group was already set, thus, with the use of by-row intersection (for each of given matrices), determinant ones (core and different for the given occupational group) are extracted. The following thresholds are used: for pairs the threshold of at least 2/3 different from the other groups (4 and more out of 6 the rest groups repeated); for triplets 100% different skills (6 out of 6). Using the logic given above, the lists of such skills were obtained for each particular occupational group of vacancies that represents at the same time core (highly-demanded) skills from companies and skills inherent in the particular occupation.

3. Results

On the stage of key skills and skillsets determination for different occupational groups in the IT sector, the most popular skills are extracted. In accordance with the different occupations, such skills are presented in the form of the Word Clouds by Top-50 skills for each occupation (by the number of occurrences in vacancies' description) in *Figure 4*.

However, within the presence of vacancies that are related to several occupations, several skills are duplicated among different groups of occupations. Thus, at the stage of extraction of pairs and triplets of skills such duplication is reduced using the cross-intersection of determinant skills. The most in demand and at the same time occupational specific pairs of skills are presented in *Table 7*, triplets – in *Table 8*[9].

As a result, from the qualitative point of view, the sets of skills in high demand are extracted for different occupational groups. Moreover, using pairs and triplets of skills, the specific combinations of skills are obtained. Thus, the proposed methods of skills preparation and extraction could be useful for a broader understanding of the demand side of the labor market and provide more evidence for the educational system in order to maintain and renew educational standards to follow the trends (in skills) on the labor market.

Conclusion

Along with the results obtained in this work, it is worth mentioning that the market is slightly diverse in terms of certain occupational groups segregation. In other words, this work provides an opportunity to run the set of classification and clusterization algorithms in order to provide the other occupational separation. In addition, the results are limited by the presence of posted vacancies in the specific online source of data but

[9] Full lists of pairs and triplets may be presented by the authors upon request

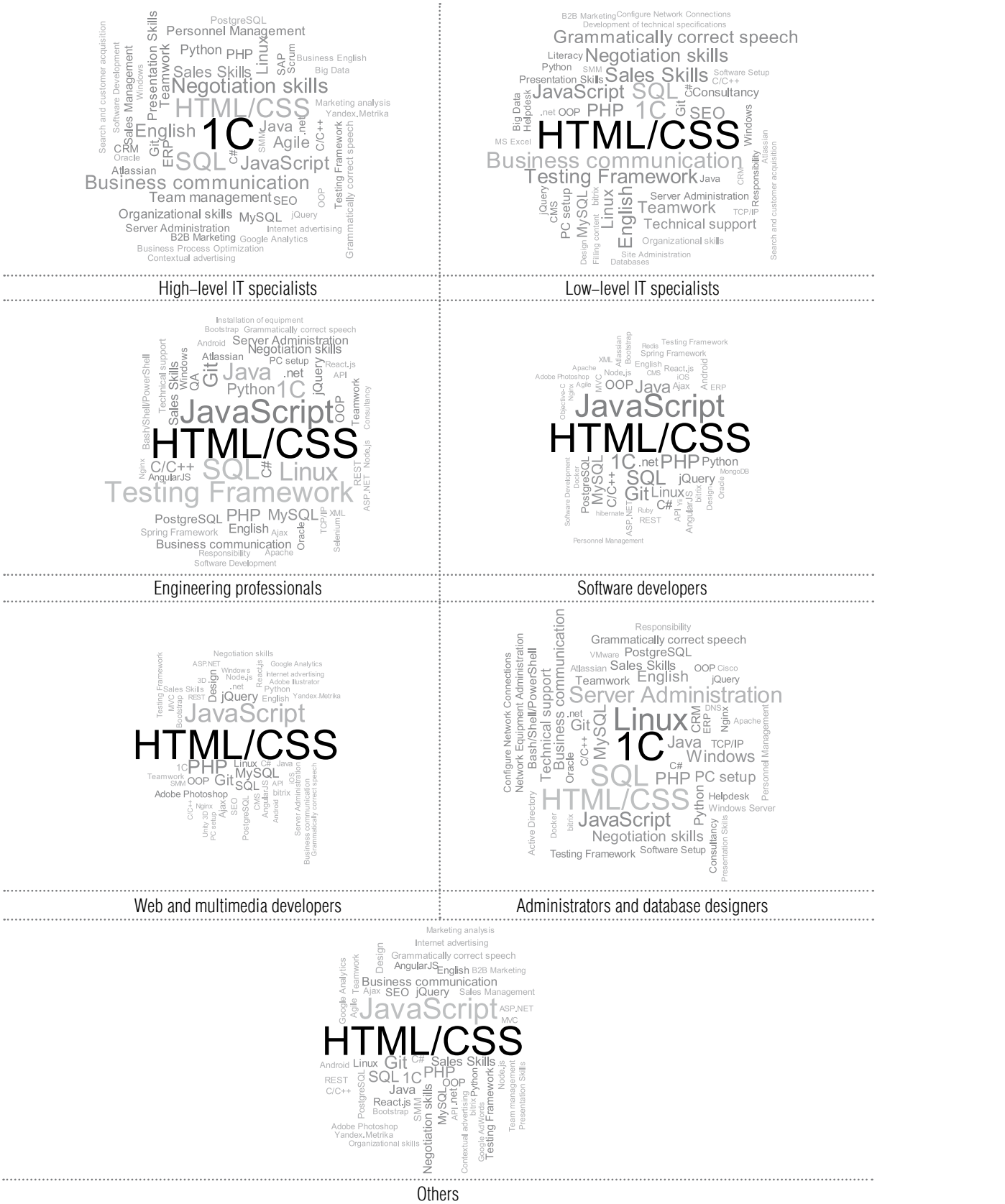

Fig. 4. Top–50 skills by occupational groups

Table 7.

Key pairs of skills by occupational groups

Skill 1	Skill 2	Jaccard Similarity	Group
Dart	Flutter	0.167	high
Billing	Solaris	0.136	high
Arduino	Raspberry Pi	0.125	high
Technical means of information protection	Assembly	0.073	high
Means of cryptographic information protection	Assembly	0.065	high
Production automation	CAD	0.125	low
CCNA	OSPF	0.125	low
A/B tests	Mobile Marketing	0.100	low
Arduino	ARM	0.083	low
Business Process Optimization	Citrix	0.038	low
Network monitoring systems	Google Cloud Platform	0.071	engineers
Cordova	Xamarin	0.065	engineers
Personnel Management	Yandex.Metrika	0.014	engineers
Elasticsearch	Node.js	0.011	engineers
MS SharePoint	Windows	0.010	engineers
Firebase	Google Cloud Platform	0.083	soft
Grammatically correct speech	Drawing up contracts	0.015	soft
Elasticsearch	Yii	0.011	soft
Scrum	TFS	0.010	soft
Contextual advertising	Search and customer acquisition	0.006	soft
Business Intelligence Systems	Olap	0.063	web
3D	Altium Designer	0.022	web
SPA	Unit Testing	0.018	web
Writing Articles	Google AdWords	0.017	web
API	Mercurial	0.016	web
Website technical audit	Technical translation	0.074	admin
Analytical research	System analysis	0.033	admin
Apache	Windows Server	0.029	admin
REST	Xsd	0.022	admin
API	Xsd	0.016	admin
Proofreading Texts	Adobe Lightroom	0.111	others
Mobility	Billing	0.111	others
Pandas	Wifi networks	0.100	others
Website technical audit	SMO	0.091	others
A/B tests	Business Analysis	0.080	others

Table 8.

Key triplets of skills by occupational groups

Skill 1	Skill 2	Skill 3	Jaccard Similarity	Group
ARM	GCC	Raspberry Pi	0.019	high
Media planning	Marketing campaign planning	facebook	0.018	high
CentOS	EJB	NetBeans	0.017	high
Video processing	Adobe Premier Pro	SketchUp	0.013	high
Business planning	Mobile Marketing	Product Marketing	0.012	high
Production automation	Instrumentation	CAD	0.050	low
Process Automation	Instrumentation	CAD	0.048	low
Production automation	Process Automation	CAD	0.043	low
Debian	OSPF	VLAN	0.043	low
Analytical research	Business Analysis	Product Marketing	0.038	low
Video processing	Image processing	Adobe Lightroom	0.020	engineers
Conducting correspondence in a foreign language	Writing Press Releases	Technical translation	0.018	engineers
Writing Press Releases	Written translation	Technical translation	0.015	engineers
Image processing	Adobe After Effects	Adobe Lightroom	0.014	engineers
FreeBSD	OSPF	VLAN	0.014	engineers
Search engine optimization sites	Work with exchanges	Website technical audit	0.021	soft
Mathematical analysis	MATLAB	R	0.017	soft
Mathematical statistics	MATLAB	R	0.016	soft
Proofreading Texts	Writing Press Releases	Presentation Preparation	0.013	soft
Work with exchanges	Website technical audit	SMO	0.013	soft
Microsoft Azure	TensorFlow	Torch/PyTorch	0.020	web
Banner advertising	Video processing	Adobe Premier Pro	0.016	web
Reporting	Tax reporting	Billing	0.016	web
Proofreading Texts	Writing Press Releases	Rewriting	0.015	web
Microsoft Azure	Spark	TensorFlow	0.014	web
Mathematical analysis	Olap	VBA	0.019	admin
Mathematical analysis	A/B tests	R	0.018	admin
Chef	LDAP	Wifi networks	0.015	admin
BGP	Chef	LDAP	0.013	admin
Chef	LDAP	OSPF	0.013	admin
Internal website optimization	Website technical audit	SMO	0.063	others
Internal website optimization	Russian search engines	SMO	0.048	others
Flask	Pandas	Wifi networks	0.037	others
Mobility	Electronic document management	Billing	0.031	others
Internal website optimization	Lidogeneration	SMO	0.027	others

could be aggregated on the level of the population, using the official statistics (if the objectives of the work will be directed to economic issues: salary, changes in time perspective, etc.).

Points for discussion are as follows. Firstly, the proposed database for the analysis has a highly diverse set of already defined occupations. In other words, introducing classification or clusterization for detecting occupational groups could improve the overall results. Nevertheless, the provided list of skills' combinations is constructed in terms of occupation-specific skillsets extraction maintenance.

Secondly, following the logic of mixed occupations that could be declared by the employer in one specific vacancy, the skills' grouping (e.g. "soft" and "hard" skills) may be used by the feature for classification purposes. Moreover, there are skills that are related to the technology itself and the framework for its implementation that cannot be separated in one-way or both directions.

Thirdly, provision of the larger sequence of words in -grams (4 and more) may provide more evidence for extraction skills from unstructured databases. However, the computing power for calculating such algorithms could be extremely high and may demand the simplification of similarity metrics calculation (e.g. using hash-functions and approximate formulas).

Fourthly, the extended implementation of the algorithm could be aimed at detection of key skillsets in the other sectors of the labor market, capturing changes in a time perspective and the organization of cross-regional comparison.

Finally, several contributions of the current work could be highlighted. Firstly, the proposed algorithm allows us to identify and standardize key skills which might be applicable for creation of the system of Russian classification for occupations and skills. Secondly, the algorithm allows us to provide lists of the most popular (key) combinations of skills that are in high demand by companies and employers inside each particular vacancy. Finally, the flexibility of the algorithm allows us to combine it with classification and clusterization techniques of data analysis that could be useful for research into the labor market.■

References

1. Autor D.H., Levy F., Murnane R.J. (2003) The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, vol. 118, no 4, pp. 1279–1333. DOI: 10.1162/003355303322552801.
2. Bensberg F., Buscher G., Czarnecki C. (2019) Digital transformation and IT topics in the consulting industry: A labor market perspective. *Advances in consulting research: Recent findings and practical cases* (ed. V. Nissen). Cham, Switzerland: Springer, pp. 341–357.
3. Christoforaki M., Ipeirotis P.G. (2015) A system for scalable and reliable technical-skill testing in online labor markets. *Computer Networks*, vol. 90, pp. 110–120. DOI: 10.1016/j.comnet.2015.05.020.
4. Florea R., Stray V. (2018) Software tester, we want to hire you! An analysis of the demand for soft skills. Proceedings of the *19th International Conference on Agile Processes in Software Engineering and Extreme Programming (XP 2018), Porto, Portugal, 21–25 May 2018* (eds. J. Garbajosa, X. Wang, A. Aguiar), pp. 54–67.
5. Goles T., Hawk S., Kaiser K.M. (2008) Information technology workforce skills: The software and IT services provider perspective. *Information Systems Frontiers*, vol. 10, no 2, pp. 179–194.
6. Johnson K.M. (2016) Non-technical skills for IT professionals in the landscape of social media. *American Journal of Business and Management*, vol. 4, no 3, pp. 102–122. DOI: 10.11634/216796061504668.
7. Kappelman L., Jones M.C., Johnson V., McLean E.R., Boonme K. (2016) Skills for success at different stages of an IT professional's career. *Communications of the ACM*, vol. 59, no 8, pp. 64–70. DOI: 10.1145/2888391.

8. Litecky C.R., Arnett K.P., Prabhakar B. (2004) The paradox of soft skills versus technical skills in is hiring. *Journal of Computer Information Systems*, vol. 45, no 1, pp. 69–76.

9. Havelka D., Merhout J.W. (2009) Toward a theory of information technology professional competence. *Journal of Computer Information Systems*, vol. 50, no 2, pp. 106–116.

10. Hussain W., Clear T., MacDonell S. (2017) Emerging trends for global DevOps: A New Zealand perspective. Proceedings of the *IEEE 12th International Conference on Global Software Engineering, Buenos Aires, Argentina, 22–23 May 2017* (ed. R. Bilof), vol. 1, pp. 21–30. . DOI: 10.1109/ICGSE.2017.16.

11. Wowczko I. (2015) Skills and vacancy analysis with data mining techniques. *Informatics*, vol. 2, no 4, pp. 31–49. DOI: 10.3390/informatics2040031.

12. Bailey J., Mitchell R.B. (2006) Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills. *Journal of Computer Information Systems*, vol. 47, no 2, pp. 28–33.

13. Brooks N.G., Greer T.H., Morris S.A. (2018) Information systems security job advertisement analysis: Skills review and implications for information systems curriculum. *Journal of Education for Business*, vol. 93, no 5, pp. 213–221.

14. Casado-Lumbreras C., Colomo-Palacios R., Soto-Acosta P. (2015) A vision on the evolution of perceptions of professional practice. *International Journal of Human Capital and Information Technology Professionals*, vol. 6, no 2, pp. 65–78. DOI: 10.4018/IJHCITP.2015040105.

15. Föll P., Thiesse F. (2017) Aligning is curriculum with industry skill expectations: A text mining approach. Proceedings of the *25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, 5–10 June 2017* (eds. I. Ramos, V. Tuunainen, H. Krcmar), pp. 2949–2959.

16. Stal J., Paliwoda-Pękosz G. (2019) Fostering development of soft skills in ICT curricula: A case of a transition economy. *Information Technology for Development*, vol. 25, no 2, pp. 250–274. DOI: 10.1080/02681102.2018.1454879.

17. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, vol. 86, pp. 319–328.

18. Colombo E., Mercorio F., Mezzanzanica M. (2019) AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, no 47, pp. 27–37. DOI: 10.1016/j.infoecopol.2019.05.003.

19. Karakatsanis I., AlKhader W., MacCrory F., Alibasic A., Omar M.A., Aung Z., Woon W.L. (2017) Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, no 65, pp. 1–6. DOI: 10.1016/j.is.2016.10.009.

20. Lovaglio P.G., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining*, vol. 11, no 2, pp. 78–91. DOI: doi.org/10.1002/sam.11372

21. Amato F., Boselli R., Cesarini M., Mercorio F., Mezzanzanica M., Moscato V., Picariello A. (2015) Challenge: Processing web texts for classifying job offers. Proceedings of the *2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, California, USA, 7–9 February 2015* (eds. M.S. Kankanhalli, T. Li, W. Wang), pp. 460–463.

22. De Mauro A., Greco M., Grimaldi M., Ritala P. (2018) Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, vol. 54, no 5, pp. 807–817. DOI: 10.1016/j.ipm.2017.05.004.

23. Gurcan F., Cagiltay N.E. (2019) Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access*, no 7, pp. 82541–82552.

24. Pejic-Bach M., Bertoncel T., Meško M., Krstić Ž. (2020) Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, no 50, pp. 416–431.

25. Radovilsky Z., Hegde V., Acharya A., Uma U. (2018) Skills requirements of business data analytics and data science jobs: A comparative analysis. *Journal of Supply Chain and Operations Management*, vol. 16, no 1, pp. 82–101.

About the authors

Andrei A. Ternikov

Doctoral Student, Doctoral School on Economics;

Lecturer, Department of Management, St. Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: aternikov@hse.ru

ORCID: 0000-0003-2354-0109

Ekaterina A. Aleksandrova

Cand. Sci. (Econ.);

Director, International Centre for Health Economics, Management, and Policy;
Associate Professor, Department of Economics, St.Petersburg School of Economics and Management;
Associate Professor, Department of Finance, St.Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: ea.aleksandrova@hse.ru

ORCID: 0000-0001-7067-5087