Processing and Analysis of Russian Strategic Planning Programs

Nikita Alekseychuk[1], Veronika Sarkisyan[1], Anton Emelyanov[2], and Ekaterina Artemova[1]

[1] National Research University Higher School of Economics, Moscow, Russia
echernyak@hse.ru,
[2] Moscow Institute of Physics and Technology, Moscow, Russia

Abstract. In this paper, we present a project on the analysis of an extensive corpus of strategic planning documents, devoted to various aspects of the development of Russian regions. The main purposes of the project are: 1) to extract different aspects of goal setting and planning, 2) to form an ontology of goals and criteria of achieving these goals, 3) to measure the similarity between goals declared by federal and municipal subjects.

Such unsupervised Natural Language Processing (NLP) methods as phrase chunking, word embeddings, and latent topic modeling are used for information extraction and ontology construction as well as similarity computation. The resulting ontology should serve in short-term as a helper tool for writing strategic planning documents and in long-term resolve the need to compose strategic planning documents completely by navigating through the ontology and selecting relevant goals and criteria. The resulting similarity measure between federal and municipal goals will serve as a navigation tool for further analysis.

Keywords: text mining · topic modeling · distributional semantics · governmental strategic planning.

1 Introduction

Recently the Russian government has decided to follow the Digital Economy approach. In the coming years, it is expected to achieve synergy between existing government routines (such as issuing a decree) and newly introduced solutions (e.g., Natural Language Processing (NLP)-based classifier for decrees). IT and AI technologies, such as Data Mining (DM), Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and Text Mining (TM) are becoming widely used in various governmental tasks, including strategic planning, both to analyze collected and stored data and to optimize and automate decision making process in the future.

There are 85 federal subjects in Russia, divided further in municipal formations. The strategic planning of development is organized as follows: the priorities, such as medicine, education, etc., are defined on the state level, further detailed on the federal level. Each municipal formation is responsible for its planning, which should follow both state and federal priorities and goals. From a technological standpoint, the planning process results in composing text documents and publishing them online. Each state unit, responsible for strategic planning, composes and publishes each year a new document, devoted to various aspects of strategic planning. Although these documents have a similar structure, to our knowledge they have never been processed or analyzed automatically. We approach the following problems:

1. to extract development goals and criteria, to which extent the goals are reached
2. to create a unified ontology that connects goals and criteria and identifies higher-level goals and higher-level criteria
3. to build a framework and a visualization tool to compare federal goals and criteria with municipal goals and criteria.

For the reader's convenience, we denote short phrases that relate to goals and criteria, to which extent the goals are reached by goals and criteria, concisely.

The remainder is organized as follows. Section 2 reviews related works. Section 3 presents our pipeline of processing strategic planning documents. Next Subsections are devoted to the steps of our pipeline, including short descriptions of distributional semantics and topic models. The last section describes possible future work directions.

2 Related works

2.1 Processing e-government documents

NLP allows to extract and structure information of governmental activity. Baturo and Dasandi [4] used topic modeling to analyze the agenda-setting process of the United Nations based on the UN General Debate corpus [16] consisting of over 7300 country statements from 1970 to 2014. Authors examined which topics were actively discussed during the investigated period, how topics interacted among themselves and related to structural factors such as population size, GDP, level of conflict, etc.

In [20] Shen et al. explored Web data and government websites in Beijing, Shanghai, Wuhan, Guangzhou and Chengdu to do comparative analysis on the development of the five metropolia e-governments. They extracted the first 30 high-frequency e-government words within the topics under study (standardization, communication, informationization, service, and security) with the help of self-made ROST Content Mining System.

Albarghothi et al. [1] introduced an Automatic Extraction Dataset System (AEDS) tool that constructs an ontology-based Semantic Web from Arabic web pages related to Dubai's e-government services. The system automatically extracts data from the website, detects keywords, and finally maps the page to ontology via Protégé tool.

2.2 Processing petitions

Natural language processing techniques are widely used for the analysis of public opinion and matching it with governmental policies. E-petitions have emerged as a popular tool for expressing public opinion. Therefore there is a need in a tool for aggregating and summating of petition texts.

The concept of e-democracy implicates open communication between government and citizens, which in most cases involves the processing of a large amount of unstructured textual information [18]. Rao and Dey describe the scheme of citizens' and stakeholders' participation in Indian e-governance which allows the government to collect feedback from citizens and correct policies and acts according to it.

In 2009, the second conference on "Working Together to Strengthen Our Nation's Democracy" in the U.S. resulted in decision "to pursue discussions with the White House Office of Public Engagement, supply relevant information, including case material". Under this program, Evangelopoulos and Visinescu [9] had access to the corpus of appeals to the U.S. government, in particular, SMS messages from Africans, sent during Barack Obama's visit to Ghana in July 2009 and data from SAVE Award - initiative, aiming to make the U.S. government more effective and efficient at spending taxpayers' money. For each of the corpus, authors extracted key topics with Latent Semantic Analysis (LSA) to explore trends in public opinion.

Similar online petition portal named e-People exists in the Republic of Korea. In [21] Suh et al. applied keyword extraction algorithms based on $tf - idf$ and K-means clustering to detect and track petitions groups. Additionally, they used radial basis function neural networks to forecast the future trend of petitions.

2.3 Processing documents in Russian

Several Russian research groups have reported on social studies that are based on using natural language processing or text mining techniques. [14] explore the attitude of news agencies towards the Russian-Ukranian conflict using topic modeling, [13] use classification algorithms to detect interethnic hostility on social media, following [2], who use topic modeling for mining ethnic-related content from Russian-language blogosphere. Opinion mining and sentiment analysis are in focus, too: for example, the authors of the paper [8] try to predict humor, while in [7] a general-aimed sentiment classification system is developed. Finally, in [10,12] new language resourse such as thesauri, aimed at socio-political topics, are built.

3 Our pipeline

3.1 Raw data processing

Raw strategic planning programs can be easily downloaded from the web page of the Strategic Planning Centre (https://strategy.csr.ru) or any other related

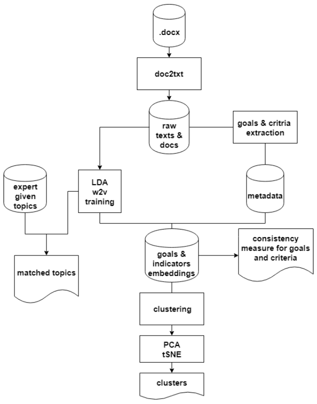


Fig. 1. Our pipeline for processing Strategic Planning Programs

governmental source. The strategic planning programs are distributed mainly in .doc, .rtf or .docx files. The easiest file formats to process are .docx files since they can be easily parsed as XML trees. One of the tools to use for processing .docx files is python-docx package [3].

In total, we downloaded around 70K files from the web. Not all of these files should be used for further analysis, since they contain other types of documents, such as financial plans, reports, etc.

The processing of raw data consists of the following steps:

1. convert .rtf and .doc files to .docx files;
2. extract all tables separately from .docx files;
3. extract raw texts separately from .docx files.

We distinguish between tables and texts while processing raw data, since some (but not all) strategic planning programs have special tables, addressed as “program passport” where there are special fields for desired development goals and criteria. The raw texts would be used further not only to extract development goals and criteria but also to train both topic and embedding models and would be addressed as Strategic planning programs corpus.

3.2 Strategic planning program classification

The downloaded corpus consists not only of strategic planning programs. There are other types of documents that fall out of the scope of the current project. This way it is necessary to distinguish between financial plans, reports and desired federal and municipal strategic planning documents. To do this, we developed a simple rule-based classifier that divides a document into three classes: the municipal program, the federal program, and other. The rules included regular expressions that search for relevant keywords and their stems, such as “муниципал”

[3] https://python-docx.readthedocs.io

("municipal"), "федеральн" ("federal"), "програм" ("program") in the header of the text and in the actual file name. The headers can be easily found by parsing XML trees, extracted from .docx files. In total, we collected and estimated about 5K federal programs from various regions devoted to various aspects of planning and around 25K municipal programs. A similar rule-based approach was applied to table classification: we need to distinguish between program passport and any other table. In this case, the rules are based on the following features: occurrence of the word "паспорт" ("passport") and the shape of the table (the number of rows and columns should exceed certain thresholds). From the raw texts, we extracted those sections, that had such keywords as "цель"("goal") or "индикатор" ("criterion") as a potential source for goal and criteria phrases, and stored these sections separately.

3.3 Goal and criteria extraction

We extracted goals and criteria from the text in an unsupervised fashion. We noted that goals and criteria are usually formulated according to certain patterns. Let us provide a few examples:

- Goals: "развитие дорожной сети в Мурманской области" ("development of the road network in the Murmansk Region"), "строительство новых больниц" ("building new hospitals")
- Criteria: "количество пострадавших в дорожно-транспортных происшествиях" ("the number of road traffic casualties"), "доля смертей от болезней системы кровообращения" ("the ratio of deaths from circulatory system diseases").

These examples can be generalized using the following formula: trigger word + noun phrase (NP) + location phrase (LP). We manually listed around a hundred trigger words for goals (such as "увелечение", "развитие", "модернизация", "оптимизация" ("increasing", "improving", "modernization", "optimization")) and around fifty trigger words ("доля", "объем", "количество", "степень" ("ratio", "volume", "number", "degree").

We used context-free parser Yargy[4] for the Russian language to define part of speech patterns for noun phrases, and location phrases and mystem[5] to identify toponyms. Next, we matched program passports and potential goal and criteria sections to the defined formula and extracted the desired goal and criteria phrases. We stripped location phrases. The last step resulted in more than 100K phrases, of which the majority occurred only once because of specific spelling or extraction mistakes. Finally, we lemmatized both the Strategic planning programs corpus and extracted goals and criteria using mystem again.

In the following sections, we describe the further analysis of the extracted goals and criteria.

[4] https://yargy.readthedocs.io/ru/latest/

[5] https://tech.yandex.ru/mystem/

3.4 Goal and criteria vectorization

Following common NLP approach [19,17], we needed to vectorize each goal and criteria, i.e. represent them in a common space of features to make further use of data mining algorithms possible. Currently, there are two major paradigms of text vectorization: namely, topic modeling [5] and distributional semantics [3], otherwise related to as word embedding models [17]. These approaches are based on a different understanding of what word similarity is: when a topic model is used, two words are similar if they share a common topic, or belong to the same domain. Word embeddings, on the other hand, use the notion of functional similarity: two words are considered similar if they are used in common contexts. We use both approaches and provide below more details on the methods and results.

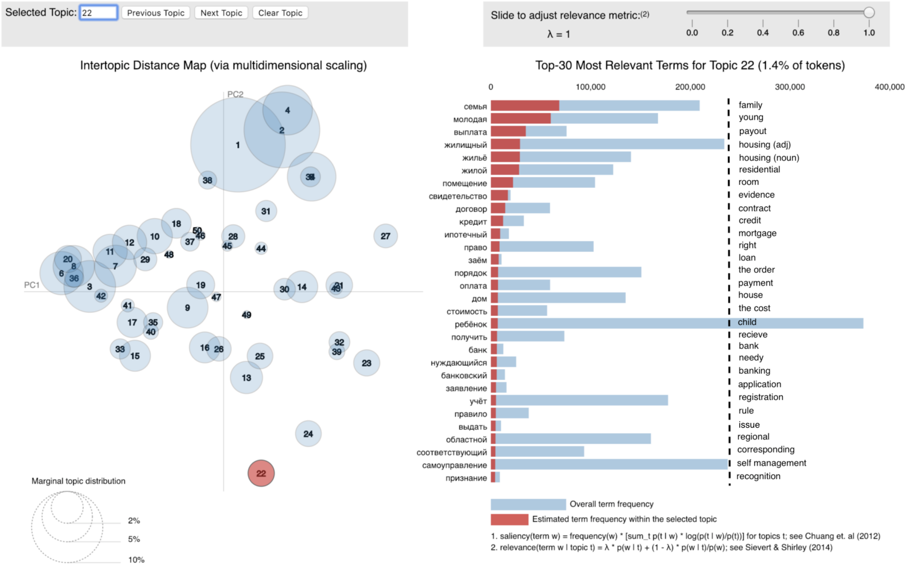
Topic modeling Formally speaking, according to [22], a topic model of the corpus of documents D is the set topics T, the distribution of words $w_i \in V$ in topics $p(w|t)$ for all topics $t \in T$, the distribution of topics $t \in T$ in document $p(t|d)$ for all documents $d \in D$. The resulting topic model can be interpreted as soft clustering of both documents and words, where each document can belong to several topics (clusters) as well as each word can belong to several topics (clusters). Topic models are widely used in various applications such as information retrieval or text classification or as a standalone text mining tool.

We trained a topic model on the Strategic planning programs corpus. We used a popular topic modeling method, Latent Dirichlet Allocation [6], implemented in gensim[6] library, to build the topic model. We chose the number of topics to be equal to 200. The visualization of the final topic model is presented on the Fig 2.

Although the topic model was interesting on its own and provided us with some useful insights of what are the topics and aspects of strategic planning in Russia, we used it mainly for vectorization of goals and criteria. The topic model allowed us to represent each goal and criteria as a vector with 200 dimensions, where each dimension stands for a single topic. After the topic model was trained, the parameters of distributions $p(w|t)$ and $p(t|d)$ were estimated, allowing us to determine the probability of a topic t to generate a document or a phrase, such as a goal or criteria. These probabilities formed the final topic vector for each goal and criteria.

Distributional semantics and word embeddings Word embedding models are another popular unsupervised technology in NLP. They are based on the simple observation, called "distributional hypothesis": words that occur in the same contexts tend to have similar meanings [11]. As a computational model of the distributional hypothesis so-called word-context matrices are constructed. While a word is a regular word or its lemma, a context can be defined differently. However the easiest definition of a context is based on using neighboring words

[6] https://radimrehurek.com/gensim/index.html



Selected Topic: 22

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. Visualization of LDA topic model. Circles on the left represent topics and their relative importance, words on the right form topic # 22. The words are sorted according to their relative importance, too.

of a given the word in a window of size k, i.e., $k/2$ words to the left and the right. More complicated techniques use syntactic contexts [15]. The word embeddings are formed in the result of some form of word matrix decomposition. For example, singular value decomposition of co-occurrence word-context matrix results in a simpler kind of word embeddings, while more sophisticated trainable algorithms such as word2vec [17] (SGNS or CBOW) result in distributed word representation.

We again used gensim[7] library to train SGNS model [17] on the Strategic planning programs corpus. It is a common technique to test the quality of the word embedding model on word pairs, that are semantically related. If two words are related, they should be close in the word embeddings space; if they are not related, they should be far from each other. We manually created a set of such word pairs, that are relevant for the strategic planning domain and tested the trained word embedding models. A few examples, where the word embedding models capture the word relations correctly are presented in Fig. 3. Note that the word embeddings are trained for a single word.

To create vectors for goals and criteria using the word embedding model we exploited the following heuristic. Given a goal, we averaged embeddings of all words, which form the goal, with $tf-idf$ weights.

To compute the similarity between two vectors we, first, normed all vectors so that the length of each vector is equal to unity. Second, we used the Euclidean

[7] https://radimrehurek.com/gensim/index.html

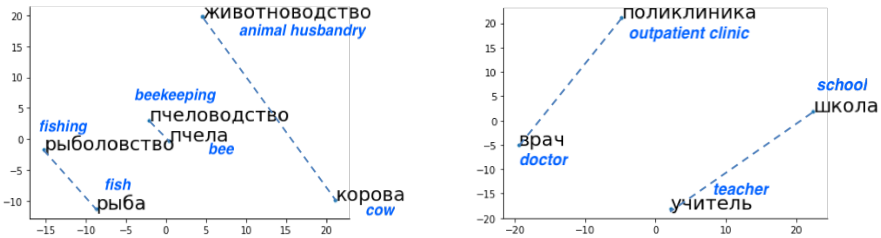


Fig. 3. An illustrative example word pair relations: (on the left) word paired "fish" : "fishing", "bee" : "beekeeping" , "cow" : "animal husbandry" lie in the same plane, however, the distance between "bee" and "beekeeping" is the shortest, most likely, because bee is the only subject to beekeeping, while there are many kinds of fishes and animals. On the right the same phenomena are illustrated.

similarity function, which in this case is equal to the widely used cosine similarity function.

Finally, we got two vectors for each goal and criteria: one was based on the topic model, second was based on the word embedding model. Both models were trained on Strategic planning programs corpus, which is large enough to get the models of high quality.

To sum up, there are two approaches for goal and criteria vectorization: the topic modeling based approach and the word embeddings based approach. These approaches can be compared by understanding what kind of word similarity they reveal. The first one reveals more of topical similarity (i. e. two words belong to the same domain, such as "doctor" and "hospital" belong to the same domain), while the latter reveals the functional similarity (i.e., two goals are used in similar contexts, such as "doctor" and "teacher"). To apply these vectorization approaches for goals and criteria, we need:

1. Topic modeling based approach: to use inference mode for the topic model and infer vectors for goals and criteria;
2. Word embeddings approach average embeddings of all words, which form the goal, with $tf - idf$ weights.

3.5 Goal and criteria clustering

Our preliminary analysis showed, that there are a lot of similar goals, such as "сбор и утилизация бытовых отходов" ("collection and disposal of household waste") and "сбор и вывоз мусора" ("garbage collection and disposal") . We used a clustering algorithm to merge such similar goals. Formally speaking, we selected goals with co-occurrence higher than some threshold and clustered corresponding word embedding vectors using K-means algorithm with K equal to 2K. The 2-dimensional visualization of goals and selected clusters is presented in Figure 4;

Finally, we considered the centroid of each cluster as its representative. This way we reduced the number of goals to 2K and the final list of unique and most important goals extracted from the strategic planning documents.

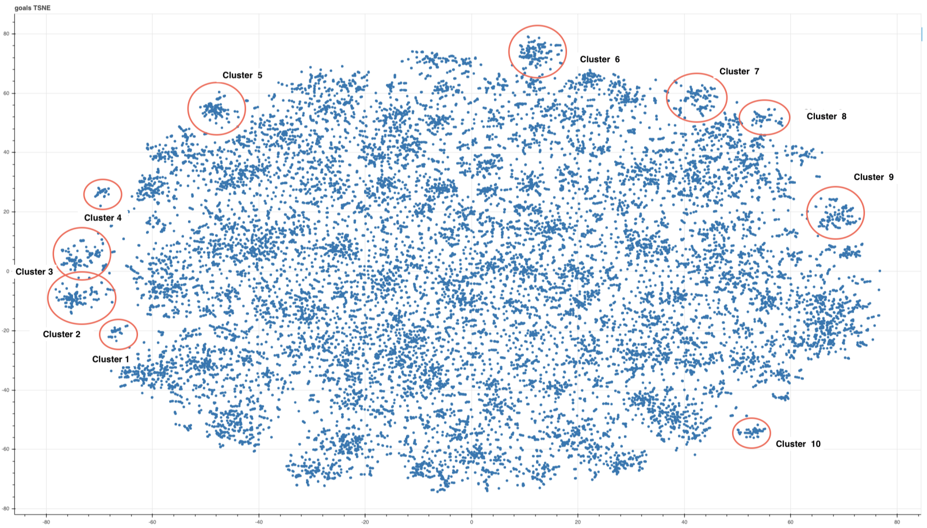


Fig. 4. t-SNE visualization of goals. **Cluster 1:** development of socially oriented non-profit organization sector; **cluster 2:** archive documents storage; **cluster 3:** preservation of library collections; **cluster 4:** creating a uniform cultural space

3.6 Goal and criteria matching

During the last step, we needed to match goals with criteria and form the desired ontology. The overview of the ontology is presented in Table 3.6. The first column stands for the ID of the LDA topic. The second column stands for the goal (the centroid of a cluster, see Subsection 3.5), which is related to this topic. Next column presents a short list of matched criteria, while the long list is presented in the last column. The shorter version is achieved by clustering all matched criteria in $N/5$ clusters, where N is the total number of matched criteria. The fourth column presents the cluster of goals.

Topic ID	Goal (centroid)	Criteria (centroids)	Cluster of goals	Cluster of criteria

Table 1. The desired ontology design

We need to admit that exact matching between goals and criteria cannot be extracted from the original unprocessed data due to the difference in formats and technical issues of parsing the raw data. To form the desired ontology, the following steps are conducted:

1. Find up to three relevant topics for each centroid of goal clusters using the inference of the topic model;
2. Find criteria that a) co-occur with the centroid or any other goal from the goal cluster more than three times and b) that have high similarity using

either topic model or word embeddings model with the goal centroid; at this point, we achieve a long list of criteria, say, its length is N;

3. Cluster all criteria in $N/5$ clusters using word embeddings representation of criteria. If $N < 5$ this step is omitted.

Two illustrative examples of the entries in the resulting ontology are provided in Tables 3.6 and 2. For the sake of space, we omit the whole clusters of goals and centroids and provide manually chosen topic names instead of their computational IDs.

Topic ID	Труд, занятость (employment)
Goal (centroid)	Снижение уровня общей безработицы (Unemployment decrease)
Criteria (centroids)	Уровень безработицы (Unemployment rate), количество ярмарок вакансий (number of job fairs)

Table 2. An example of ontology entry

Topic ID	Безопасность (safety)
Goal (centroid)	Снижение гибели и травматизма людей на водных объектах (reduction of death and injuries of people on water bodies)
Criteria (centroids)	Количество обученных спасателей (number of trained rescuers), Количество оборудованных мест массового отдыха на водных объектах (the number of equipped places for mass recreation on water bodies)

Table 3. An example of ontology entry

The resulting ontology enumerates more than 7K entries, following the design, presented in Table 3.6. Thoroughly expert evaluation of the whole body of the ontology needs to be conducted as an obligatory direction for future work.

3.7 Municipal and federal matching

There arises another analytical task: since we got strategic goals on different levels of country division, namely, federal and municipal divisions, we were able to evaluate, whether the strategic goals of a municipal formation coincide and are consistent with the strategic goals of a federal subject. We may expect municipal formation to follow the strategy of a federal subject to a certain degree, although this is not necessary. However, even if the municipal formation follows the strategy of the regional subject, the wording of the goals still might be different. Hence we needed to exploit more sophisticated NLP techniques rather than straightforward word matching. Another reason to apply NLP techniques here is that the regional subject might set quite a broad goal, such as “развитие сельского хозяйства” (“agricultural development”), the municipal formation may choose to concentrate on the particular part of this goal, such as “развитие овцеводства” (“sheep breeding development”). Although the word2vec embeddings we calculated beforehand are not meant to capture such hierarchical relations, we still tried to use it for this task, since little end-to-end algorithms for discovery of hierarchical relations are available.

In the previous steps, we calculated goal embeddings for each municipal and federal subject. To calculate the matching between federal and municipal formation we constructed a similarity matrix in such a way that its rows represented goals of municipal formation and columns represented goals of regional subjects. The weights of the matrix represented cosine similarity (cossim) between corresponding embeddings. Further on, the following metric was used:

$$M = \frac{\sum \text{cossim of the closest goals for the current regional subject}}{\text{number of goals of the regional subject}} \times \frac{\sum \text{cossim of the closest goals for the current municipal formation}}{\text{number of goals of the municipal formation}} \tag{1}$$

By term **cossim of the closest goals for the current regional subject** we imply column-wise maximum of the matrix (row-wise respectively for municipal formations). The interpretation of this operation is to find the closest goal of municipal (regional) formation for every municipal (regional) goal. Note that the metric, defined in Eq. 1 varies between 0 and 1, which makes different regional subject-municipal formation pairs consistently comparable. The closer to unity, the more similar are the goals of the municipal formations to the goals of the regional subject, the closer to 0 – the more diverse they are.

After all, metrics are calculated, we got the list of municipal formations and corresponding regional subjects along with the similarity scores. We averaged these scores for the visualization of the whole Russian map or to plot one region as it can be seen below in Fig. 5.

4 Conclusions and future work

In this paper, we present an ongoing project on the analysis of the strategic planning documents. We explore a part of the corpus of strategic planning documents that consist of strategic planning programs only. We limit our scope to the extraction and analysis of strategic goals and criteria for achieving these goals. In this direction, the following tasks are considered:

- development of the pipeline for downloading, parsing and processing of strategic planning documents;
- extraction of keywords that stand for strategic goals and criteria (addressed throughout as goals and criteria) using rule-based NLP techniques;
- training and visualization of a topic model along with a model of word embeddings, that allow two types of goals and criteria vectorization, used further to compute the similarity measure between goals and criteria;
- construction of the ontology of the goals and criteria that presents the whole scope of strategic goal settings;
- visualization of the suggested consistency measure that shows how consistent are the goals of a municipal formation are by the goals of a federal subject.

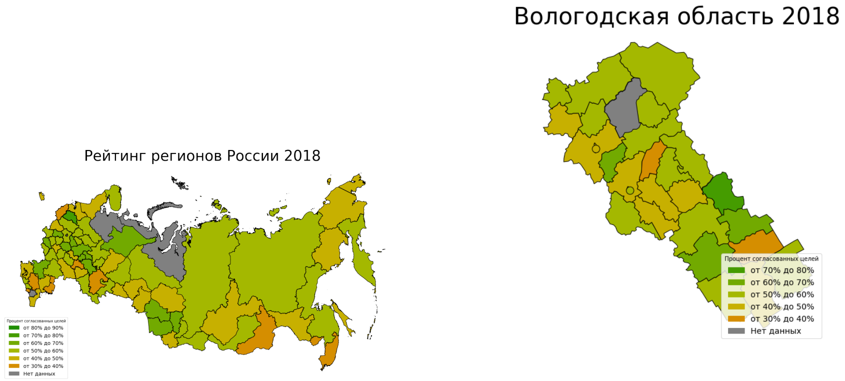


Fig. 5. (On the left) Rating of Russian regions: goals consistency measure; (On the right) goals consistency measure for Vologda region. The green color indicates the values of the cosine similarity measure: the brighter the color is, the closer are the values to unity. The grey color means no data is provided.

The main contributions of the paper are the following. Firstly, we use well-known and purely computational techniques that require little manual tuning, to create an ontology of strategic goals and criteria to achieve these goals. This ontology might be of future use for those who compose strategic planning documents, as it unifies different wording of goals and criteria. Secondly, we develop a measure that is capable of capturing hierarchical semantic relations between phrases, of evaluating the consistency of planning of federal subjects and their municipal formations. This measure is visualized on a map, which provides an analytical tool to manage and evaluate the strategic planning process in municipal formations. Thirdly, our approach and pipeline can be easily adapted to any other domain, where semi-structured documents are available. In this project, we were able to avoid manual or crowd-sourcing annotation of the documents for further training of extraction algorithms due to a specific style of writing and usage of certain noun phrase patterns. In any other domain which obtains the same linguistics specifics, our approach is applicable.

The future work directions include, but are not limited to manual evaluation of the constructed ontology and the consistency measure. Each step of our pipeline can be evaluated by computing descriptive statistics, such as the number of correctly extracted / clustered / defined as similar items and the number of missed or incorrectly processed items. This evaluation is, however, rather time-consuming and requires expert knowledge, so the procedure of the evaluation is still under development.

References

1. Albarghothi, A., Saber, W., Shaalan, K.: Automatic construction of e-government services ontology from arabic webpages. Procedia computer science **142**, 104–113 (2018)
2. Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., Vorontsov, K.: Mining ethnic content online with additively regularized topic models. Computacion y Sistemas **20**(3), 387–403 (2016)
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 238–247 (2014)
4. Baturo, A., Dasandi, N.: What drives the international development agenda? an nlp analysis of the united nations general debate 1970–2016. In: the Frontiers and Advances in Data Science (FADS), 2017 International Conference on. pp. 171–176. IEEE (2017)
5. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
7. Chetviorkin, I., Loukachevitch, N.: Evaluating sentiment analysis systems in russian. In: Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing. pp. 12–17 (2013)
8. Ermilov, A., Murashkina, N., Goryacheva, V., Braslavski, P.: Stierlitz meets svm: Humor detection in russian. In: Conference on Artificial Intelligence and Natural Language. pp. 178–184. Springer (2018)
9. Evangelopoulos, N., Visinescu, L.: Text-mining the voice of the people. Communications of the ACM **55**(2), 62–69 (2012)
10. Galieva, A., Kirillivich, A., Loukachevitch, N., Nevzorova, O., Suleymanov, D., Yakubova, D.: Russian-tatar socio-political thesaurus: Publishing in the linguistic linked open data cloud. International Journal of Open Information Technologies **5**(11), 64–73 (2017)
11. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
12. Kirillovich, A., Nevzorova, O., Gimadiev, E., Loukachevitch, N.: Ruthes cloud: Towards a multilevel linguistic linked open data resource for russian. In: International Conference on Knowledge Engineering and the Semantic Web. pp. 38–52. Springer (2017)
13. Koltsova, O., Alexeeva, S., Nikolenko, S., Koltsov, M.: Measuring prejudice and ethnic tensions in user-generated content. ANNUAL REVIEW OF CYBERTHERAPY AND TELEMEDICINE 2017 p. 76 (2017)
14. Koltsova, O., Pashakhin, S.: Agenda divergence in a developing conflict: Quantitative evidence from ukrainian and russian tv newsfeeds. Media, War & Conflict p. 1750635219829876 (2017)
15. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 302–308 (2014)
16. Mikhaylov, S., Baturo, A., Dasandi, N.: United nations general debate corpus (2017). https://doi.org/10.7910/DVN/0TJX8Y, https://doi.org/10.7910/DVN/0TJX8Y

17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
18. Rao, G.K., Dey, S.: Decision support for e-governance: a text mining approach. arXiv preprint arXiv:1108.6198 (2011)
19. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
20. Shen, Y., Liu, Z., Luo, S., Fu, H., Li, Y.: Empirical research on e-government based on content mining. In: Management of e-Commerce and e-Government, 2009. ICMECG'09. International Conference on. pp. 91–94. IEEE (2009)
21. Suh, J.H., Park, C.H., Jeon, S.H.: Applying text and data mining techniques to forecasting the trend of petitions filed to e-people. Expert Systems with Applications **37**(10), 7255–7268 (2010)
22. Vorontsov, K., Potapenko, A.: Additive regularization of topic models. Machine Learning **101**(1-3), 303–323 (2015)