

Prospectus

1 Topic

Autocomplete Russian addresses using semantic network.

2 Research questions

1. How to extract part of address from loose structured data?
2. Is it possible to process extracted data to structured parts of addresses?
3. Is it possible to autocomplete structured parts of addresses to full addresses?
4. How to select the correct address from different collisions?

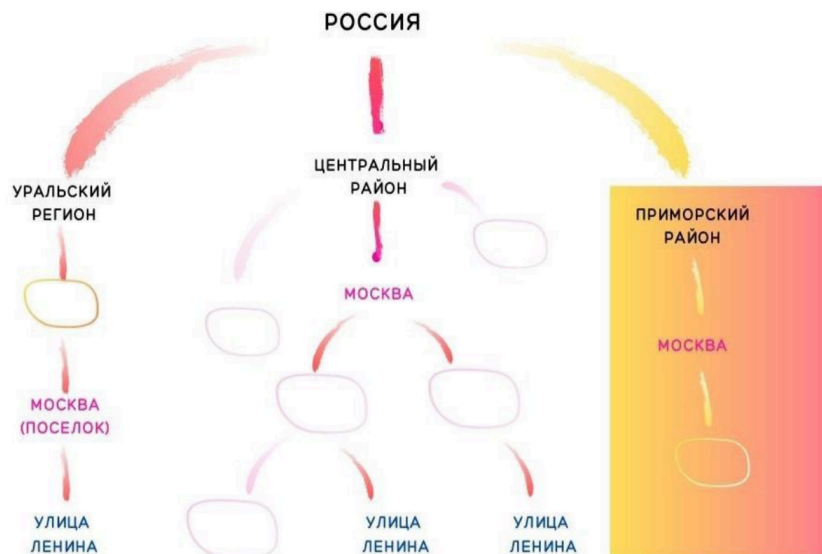
3 Working hypotheses

1. Native language's struct helps to recognize parts of address.
2. Data from FIAS (federal information address system) structured.
3. Structured data may be a template for autocomplete.
4. Frequency dictionaries help solving collisions.

4 Roadmap

In the modern world we have a problem with autocomplete. As usual it bases on frequency dictionaries and works only with full data. But if we want to fill spaces in low structured data modern autocomplete systems create very imprecise results. So we create a roadmap to solve this problem.

First we need to download data from FIAS. FIAS stores data in XML files, which we may translate into a semantic network. A Semantic network will store data in a dependency graph.



Then We need to convert all unique words from FIAS into a hash table. Next we need to split text to words and filter them using a hash table. After that we will have unstructured data which communicate with addresses. For structuring this data we need to group words to tokens using Native language rules. Grouping will be done by taking words from left to right till loose association in the hash table.

For example:

“Город СПб улица ленина” - base text.

“Город” - valid token.

“Город СПб” - valid token.

“Город СПб улица” - invalid token.

That means that “Город СПб” will be separated from text and used in analysis. Tokens should be placed in the dependency graph. After we choose all branches, which includes all tokens and sort them using frequency dictionary. Top ten results will be suggested to the user. If there are no correct results, the user may give more information about the address.

After we prepare data and develop algorithms we will collect data from websites. Websites often contain information which explain how to get to some destination point.

5 Works Cited

1. Slapoguzov A., Malyuga K., Tsopa E. Word Sense Induction for Russian Texts Using BERT//Proceeding of the 28th conference of FRUCT association, 2021, Vol. 2, pp. 621
2. Tsopa E., Mileschin A.A., Slapoguzov A., Klimenkov S. Sentiment Analysis of Russian Text Using the Semantic Network//Proceedings of the XII Majorov International Conference on Software Engineering and Computer Systems, 2021, Vol. 2, pp. 100
3. Uskov I., Yarkeev A., Tsopa E. Named numeric characteristics extraction from text data in Russian//CEUR Workshop Proceedings, 2020, Vol. 2590, pp. 1-8
4. Pismak A., Gavrilov A., Klimenkov S., Yarkeev A., Tsopa E., Nikolaev V. Method of Semantic Refinement for Enterprise Search//In Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2020, Vol. 2, pp. 307-312

6 Keywords

Semantic networks, addresses, autocomplete, Russian language.