Graph-Based Approach to Constructing Word Definitions

Danilova Veronika
23.12.2017

How to write a comprehensible word definition?

Can we create such a word definition automatically?

Some manual...

1) Find examples of the word in use **- corpora**
2) Examine how the word functions in the examples that you find
3) Determine the part of speech of the word, for the sense/senses you wish to define **- morphological parser**
4) Make further distinctions within the parts of speech **- morph. parser**
5) **Choose a specific sense of the word and think about the word's meaning in that sense**
6) Think of synonyms and antonyms for the word **- word2vec**
7) **Describe the word**

How graphs can be helpful?

If we have enough contexts of our word, we can

1. create a graph of its usage with the words in its contexts (n-grams)
2. calculate graph characteristics of these words
3. choose the most significant words on the basis of calculated characteristics
4. use them to define our word

My plan

1) Choose a number of words (nouns) that have no more than 2-3 senses or have one main sense, while the others are far on the periphery of the usage
2) Collect their contexts from Ruscorpora
3) Extract 2-5-grams for each word
4) Create a graph for each word, where an edge between the words (nodes) means that they appeared in some text at the distance of 0-4 words
5) Calculate centrality for each node in each graph
6) For each graph extract top-50 most central words
7) Compare them with dictionary definitions
8) Draw conclusions

Fish

Number of nodes: 32125

Number of edges: 833573

Average degree: 51.8956

Max degree of a core: 127

Number of nodes in max k-core: 623

Result:

рука, дело, дом, ловить, река, мясо, дать, хотя, оказаться, пока, новый, огромный

Tree

Number of nodes: 49019

Number of edges: 1795039

Average degree: 73.2385

Max degree of a core: 233

Number of nodes in max k-core: 799

Result:

это, большой, сад, работа, леса, ствол, растение, какой-то, ряд, лист, жить, часть, метр

Stone

Number of nodes: 43143

Number of edges: 1369113

Average degree: 63.4686

Max degree of a core: 180

Number of nodes in max k-core: 767

Result:

сторона, сделать, просто, дерево, берег, россия, ребёнок

Salt

Number of nodes: 21524

Number of edges: 399313

Average degree: 37.1040

Max degree of a core: 71

Number of nodes in max k-core: 509

Result:

мочь, масло, каждый, продукт, делать, использовать, вод, также, количество, перец, тело, хлеб, вещество, например, организм, процесс, затем

Sun

Number of nodes: 44941

Number of edges: 1561000

Average degree: 69.4689

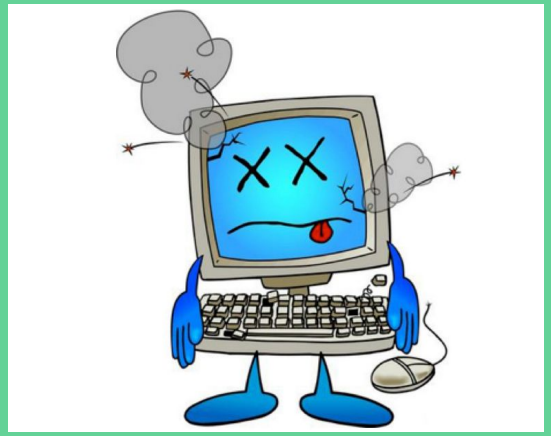
Max degree of a core: 215

Number of nodes in max k-core: 655

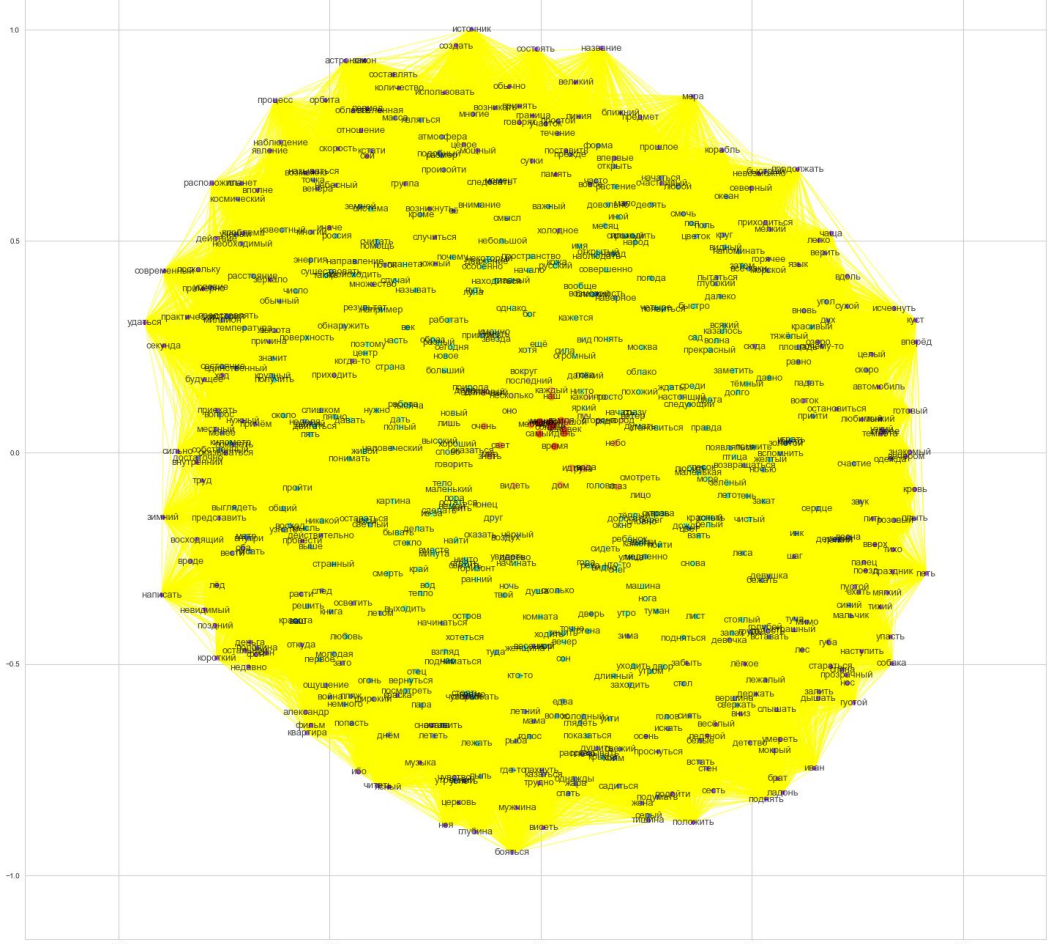
Result:

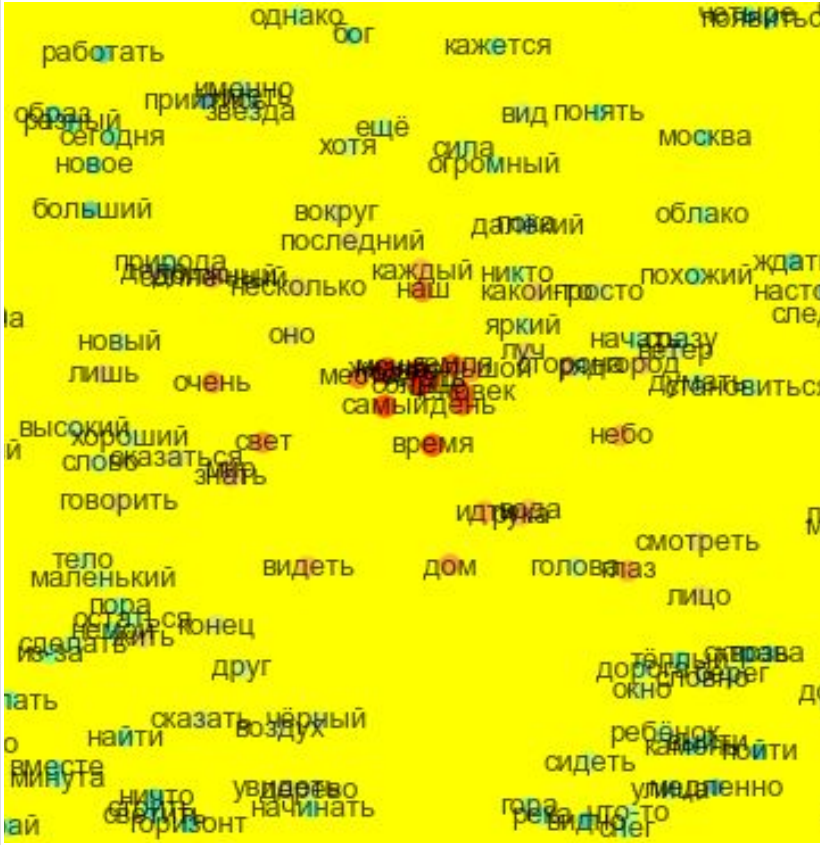
человек, земля, небо, мир, свет, солнечный, лицо, видеть, последний, вокруг, нога, луч, увидеть, ветер, сидеть, высокий, воздух, смотреть

Closeness centrality...
Betweenness centrality...
Eigenvector centrality...



Sun max k-core, colour - degree centrality





Conclusions

- **graphs can be helpful !!!**
- the list of stop-words needs to be extended
- other types of centrality should be tested
- there is a place for further research with other parts of speech and other semantic groups of nouns (abstract, etc.)

Thank you for your attention !