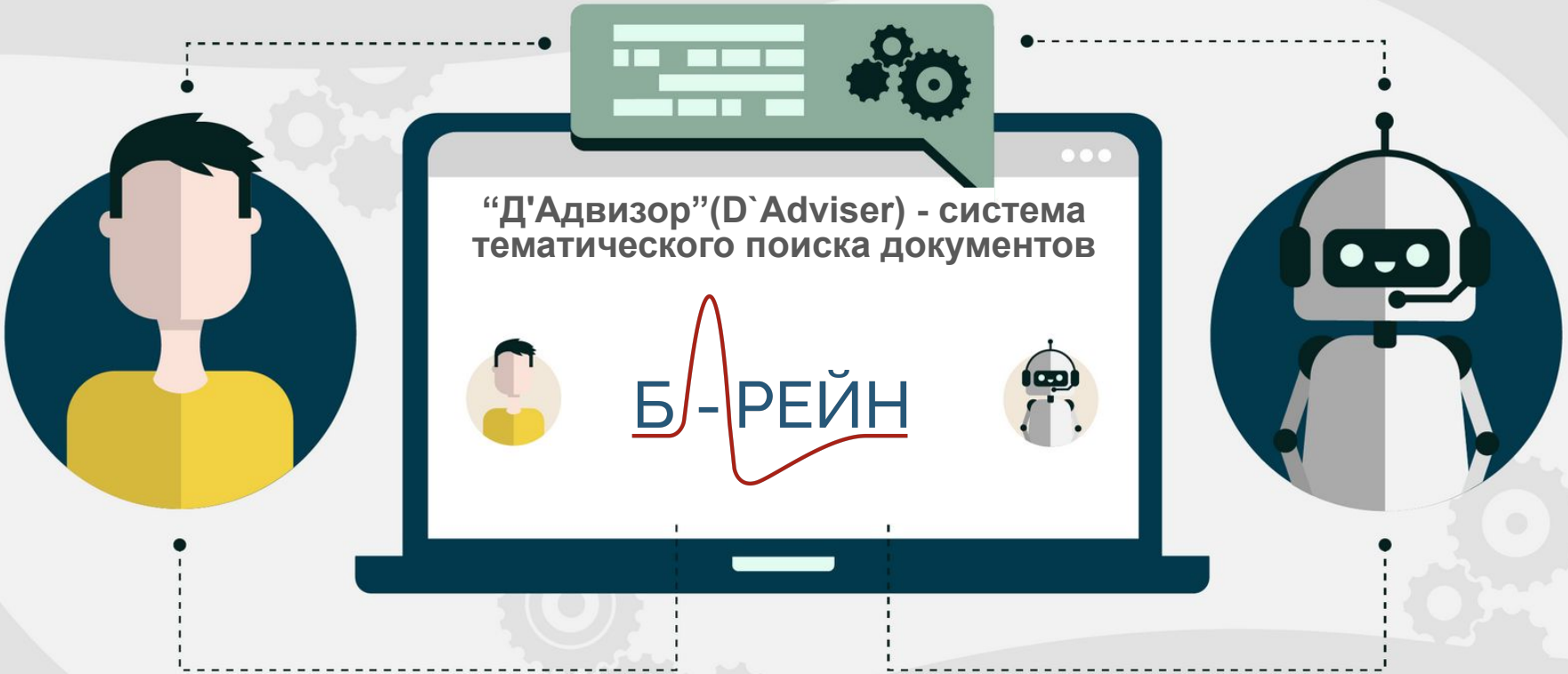
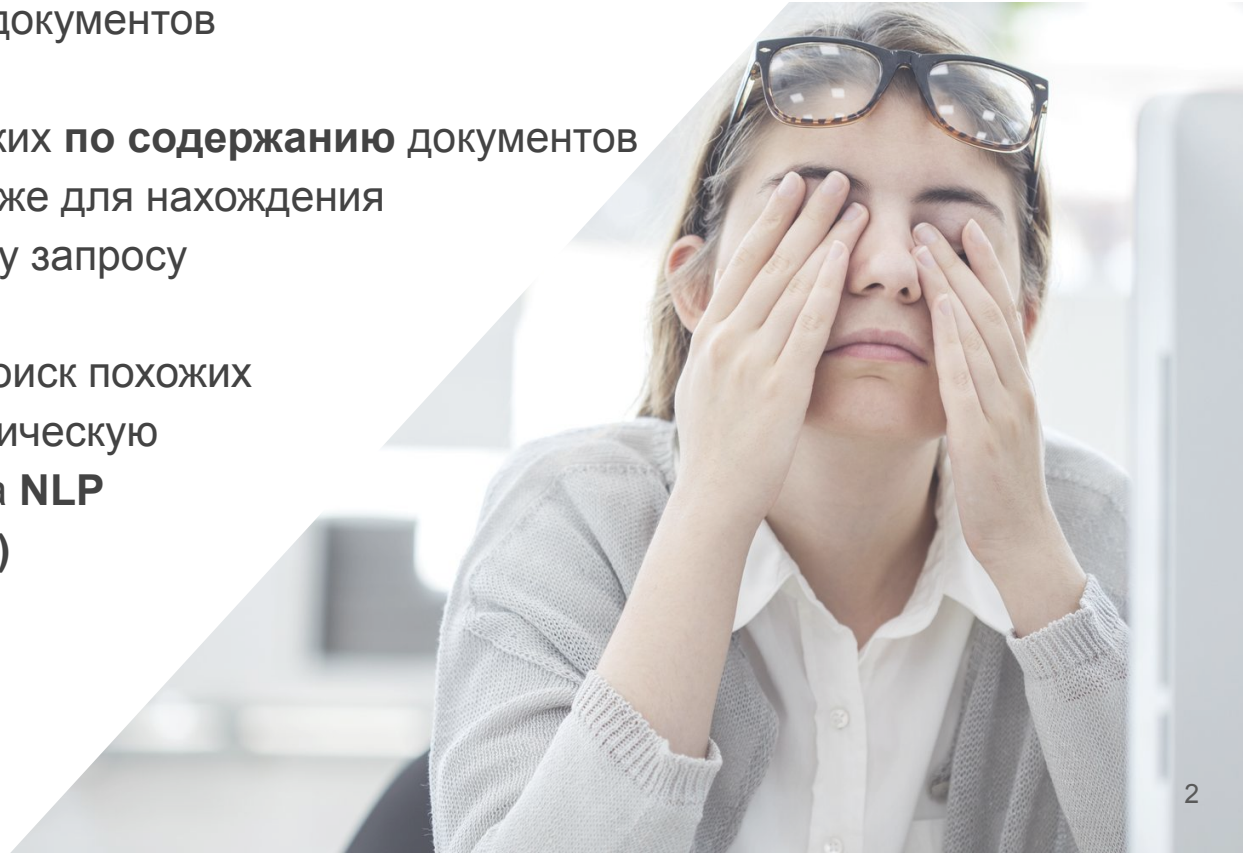

“Д'Адвизор”(D`Adviser) - система тематического поиска документов
Б-РЕЙН

Постановка задачи

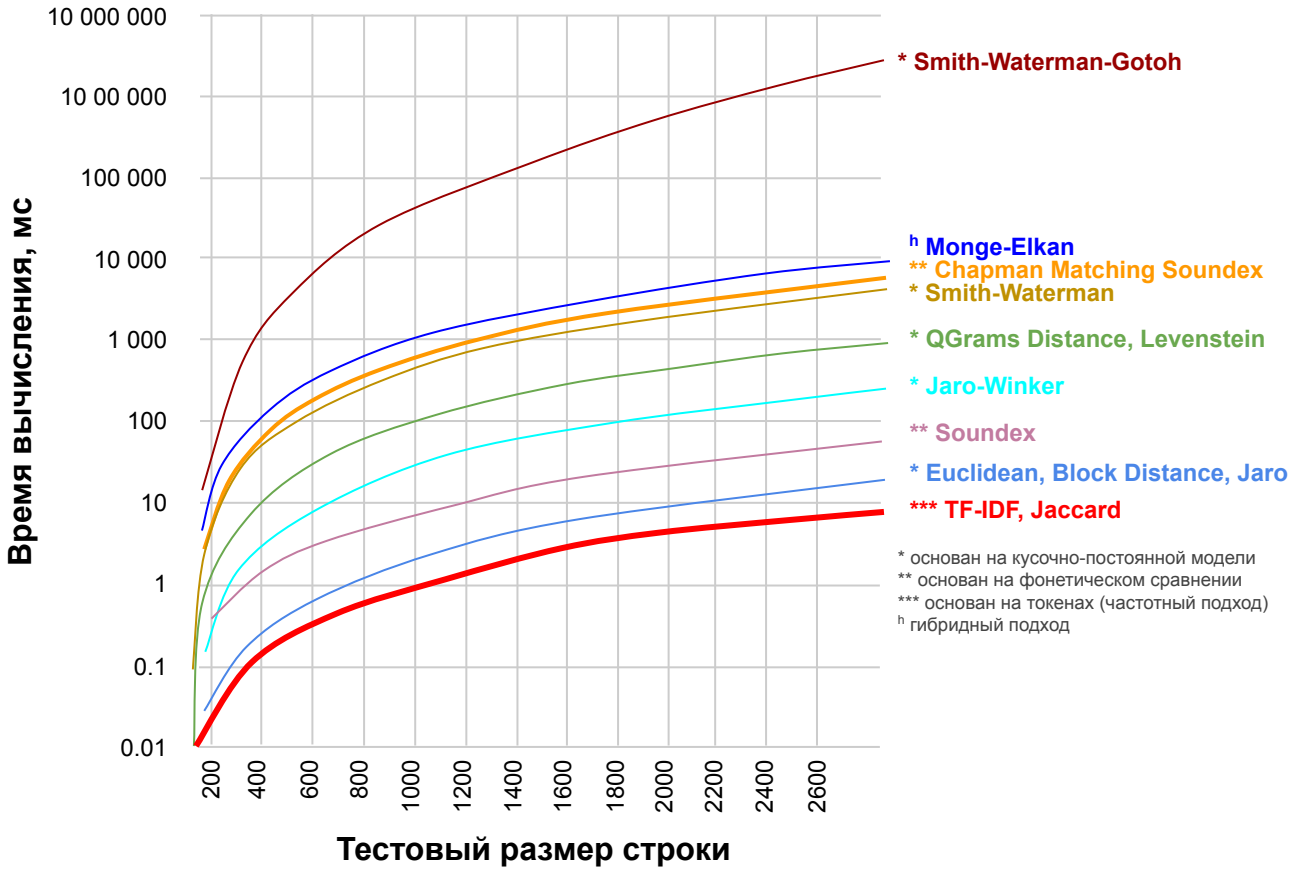
Дано: выборка распознанных документов

Задача: упростить поиск похожих **по содержанию** документов для добавления в связки, а также для нахождения истории переписки по искомому запросу

Решение: автоматизировать поиск похожих документов, используя статистическую обработку естественного языка **NLP (Natural Language Processing)**



Анализ решений



TF-IDF

\+ **высокая скорость** обработки документов

\+ **не требует** больших вычислительных мощностей и времени на обучение

\+ **удобен** в использовании (библиотеки)

\+ **способен** обрабатывать большие по объёму тексты

\- не применяет семантику (груши и яблоки это разные слова, семантика одна - фрукты)

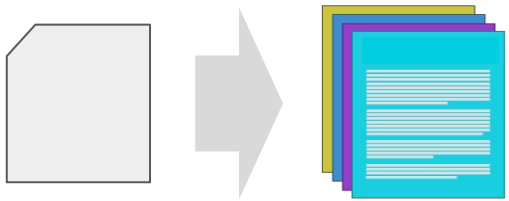
Department of Statistics, Carnegie Mellon University
Cohen, William & Ravikumar, Pradeep & Fienberg, Stephen. (2003). A comparison of string DISTANCE metrics for name-matching tasks. IIWeb. 2003

Natural Language Processing Group, Department of Computer Science,University of Sheffield
www.coli.uni-saarland.de/courses/LT1/2011/slides/stringmetrics.pdf

Similarity Metric Library
https://github.com/magsilva/SimMetrics

Техническое описание решения

Сбор и обработка документов



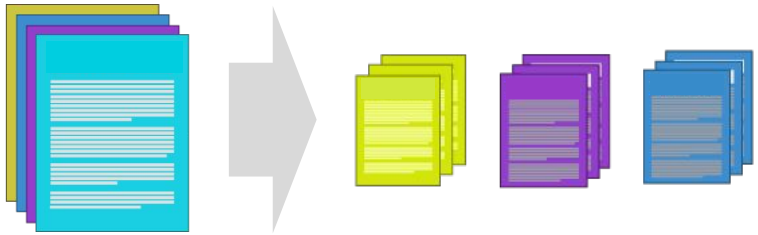
Решения для сегментации на токены	Ошибки на 1000 токенов	Время обработки, секунды
Regexp-baseline	19	**0.5**
SpaCy	17	5.4
NLTK	130	3.1
MyStem	19	4.5
SegTok	12	2.1
SpaCy Russian Tokenizer	**8**	46.4

Препроцессинг (Regex)

- объединение страниц в один документ
- удаление битых текстов (кодировка, OCR)
- разбиение на токены (слова), Regex
- удаление стоп-слов (предлоги, частицы...)

Техническое описание решения

Определение тематики текста через вычисление веса слова



Столько раз слово встретилось в документе

Всего документов

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_j}$$

tf-idf score

всего документов с данным словом

Лемматизация (PyMorphy2)

- приведение слова к словарной форме
- сохранение лемм для каждого документа

Взвешивание (GenSim)

- создание мешка слов и корпуса
- вычисление весов слова, используя информационно-поисковую систему SMART

Centre for Telematics and Information Technology, University of Twente, The Netherlands
Hiemstra, D. A probabilistic justification for using tf×idf term weighting in information retrieval . Int J Digit Libr 3, 131–139 (2000). https://doi.org/10.1007/s007999900025

Техническое описание решения

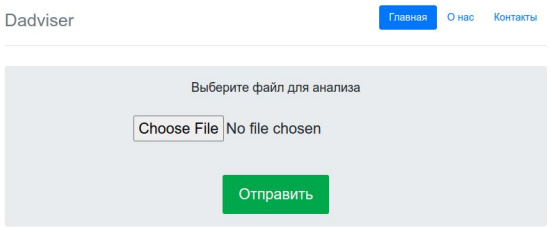
Поиск похожих документов

Загрузка

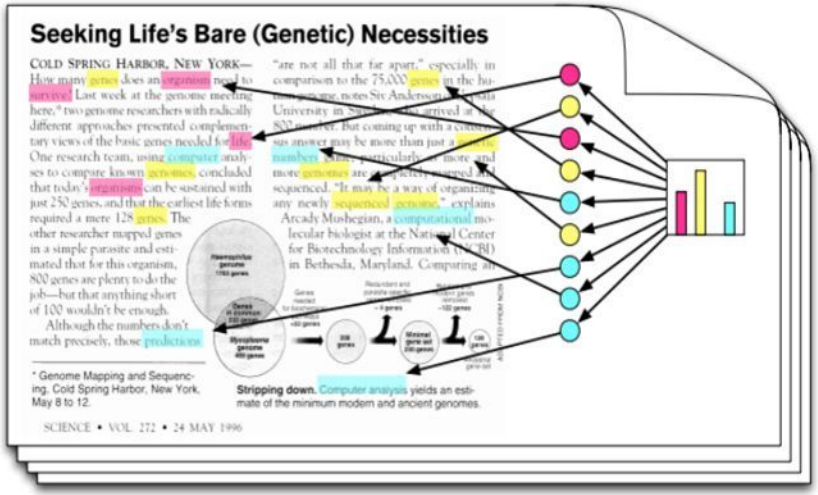
Док.
????

Док.
№183

Результат



*There are words from yellow topic, word from red topic and blue topic. There are no words from green topic. It means doc **is not** green topic*



Policy **183** 96%

Policy **471** 35%

Policy **244** 5.1%

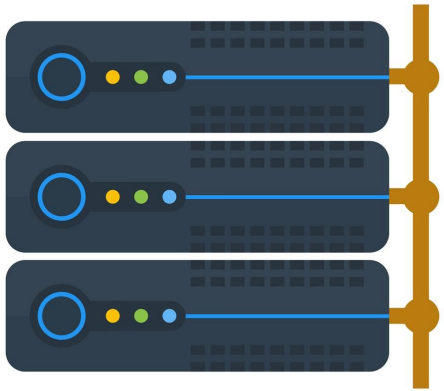
...

Policy **937** 0.5%

Требования к инфраструктуре

Сервер

Минимальные требования:
Core i7, CPU: 6, RAM 8 GB/16 GB,
SSD/HDD 100+ GB



Тестовые данные

совокупный размер выборки документов : 550 MB
количество документов: 36 810
58 миллионов русских слов

Временные показатели

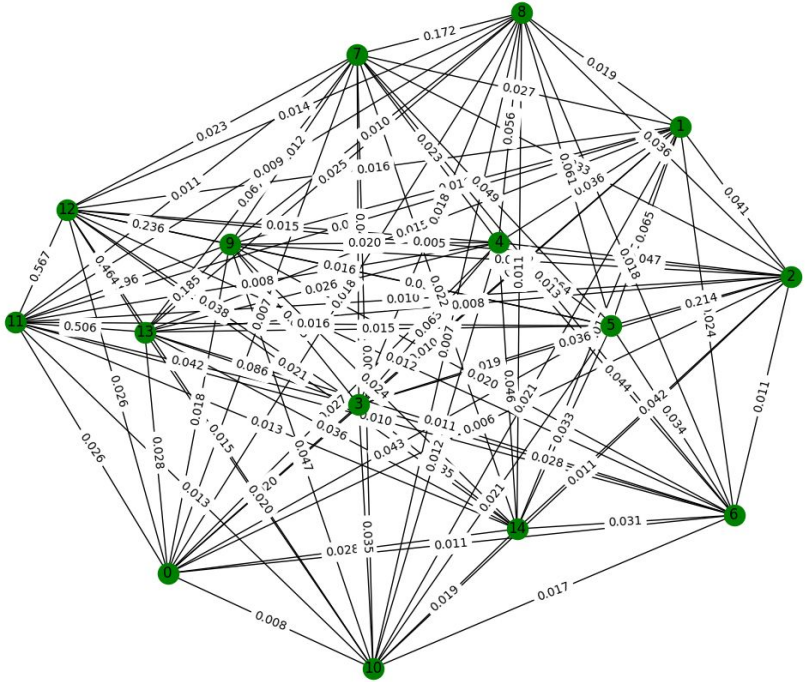
(на тестовых данных)

препроцессинг 1.5 мин
лемматизация 18 мин
* обучение 2 мин

* на большем объеме данных можно обучать 2 модели:
одну с данными за день,
а вторую – с данными за все время

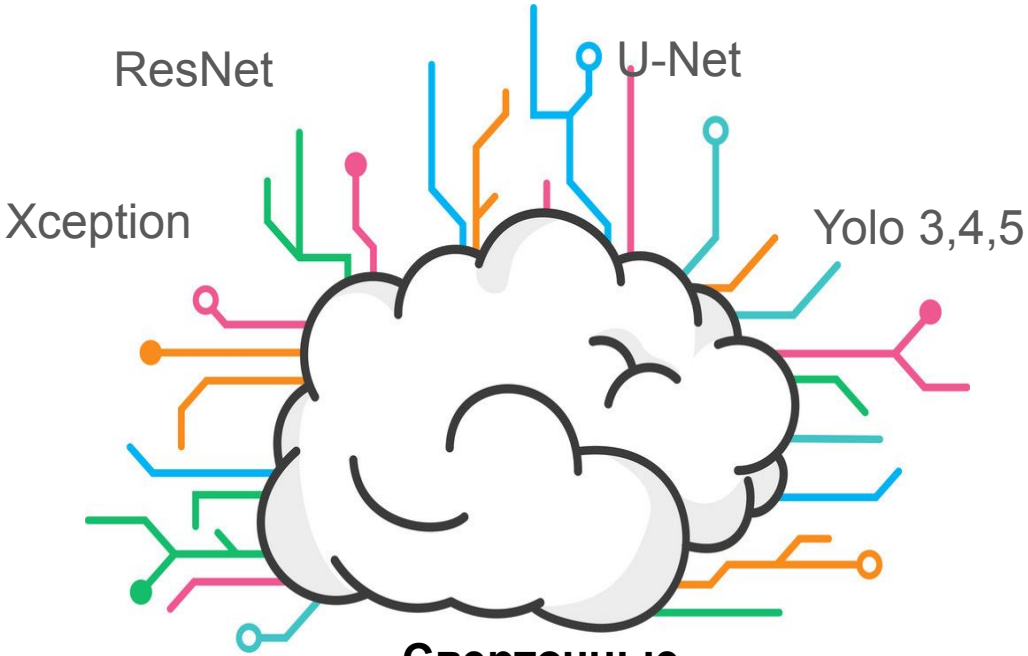
Тестирование решения

- визуальная проверка «схожести документов»
- построение графов похожих документов
  - «если документ А похож на документы В и С, то проверяем, есть ли еще документы, определяемые как похожие на В или С, но непохожие на А»
- проверка взаимной схожести документа
  - «если документ А похож на документы В и С, то проверяем насколько документы В и С похожи на документ А»



Пример графа из 15 документов с указанием схожести

Наши компетенции

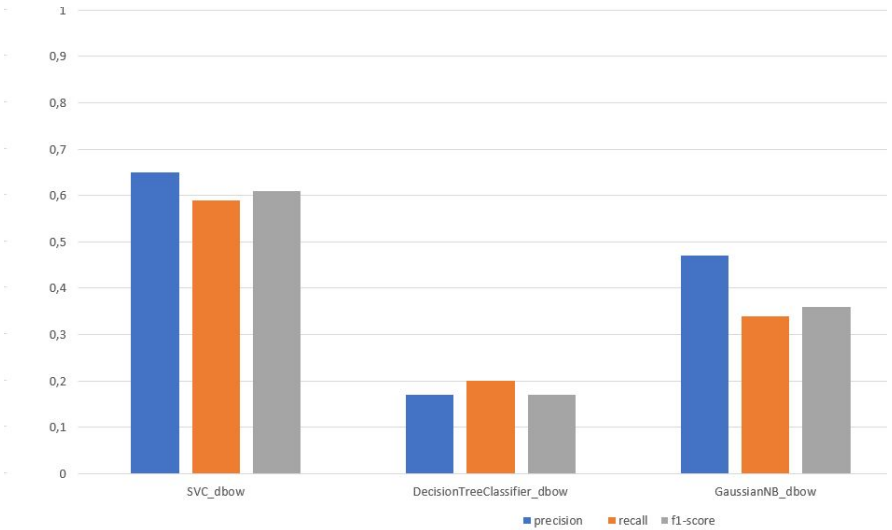


Сверточные нейронные сети (NNs)

Наши возможности в Natural Language Processing

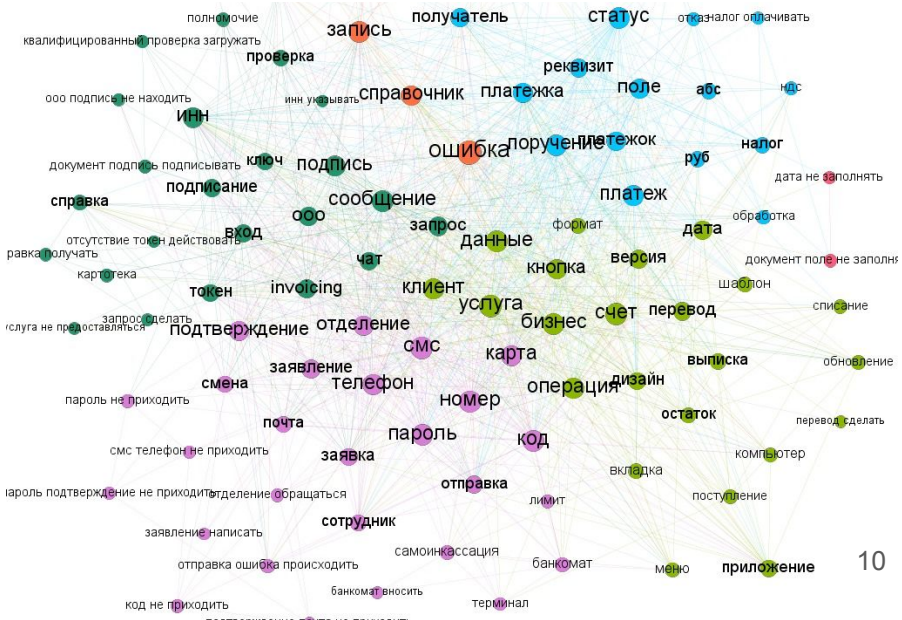
Анализ текста, написанном на естественном языке (русский, английский)

Обработка инцидентов с помощью Word2Vec



Семантический анализ текста (Latent semantic analysis, LSA) на основе анализа tf-idf

Поиск сходных документов



Спасибо за внимание!