

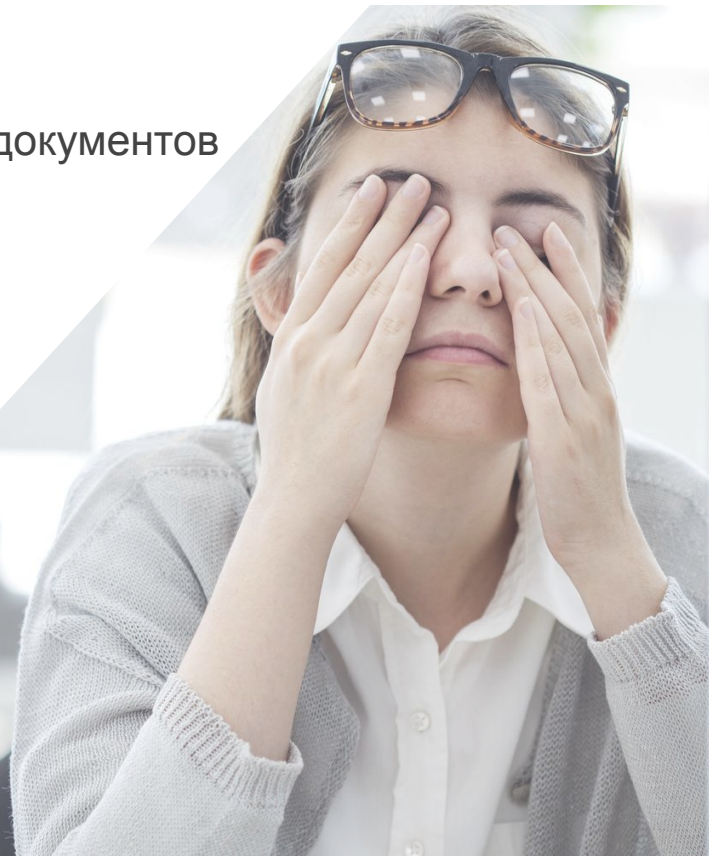


# Постановка задачи

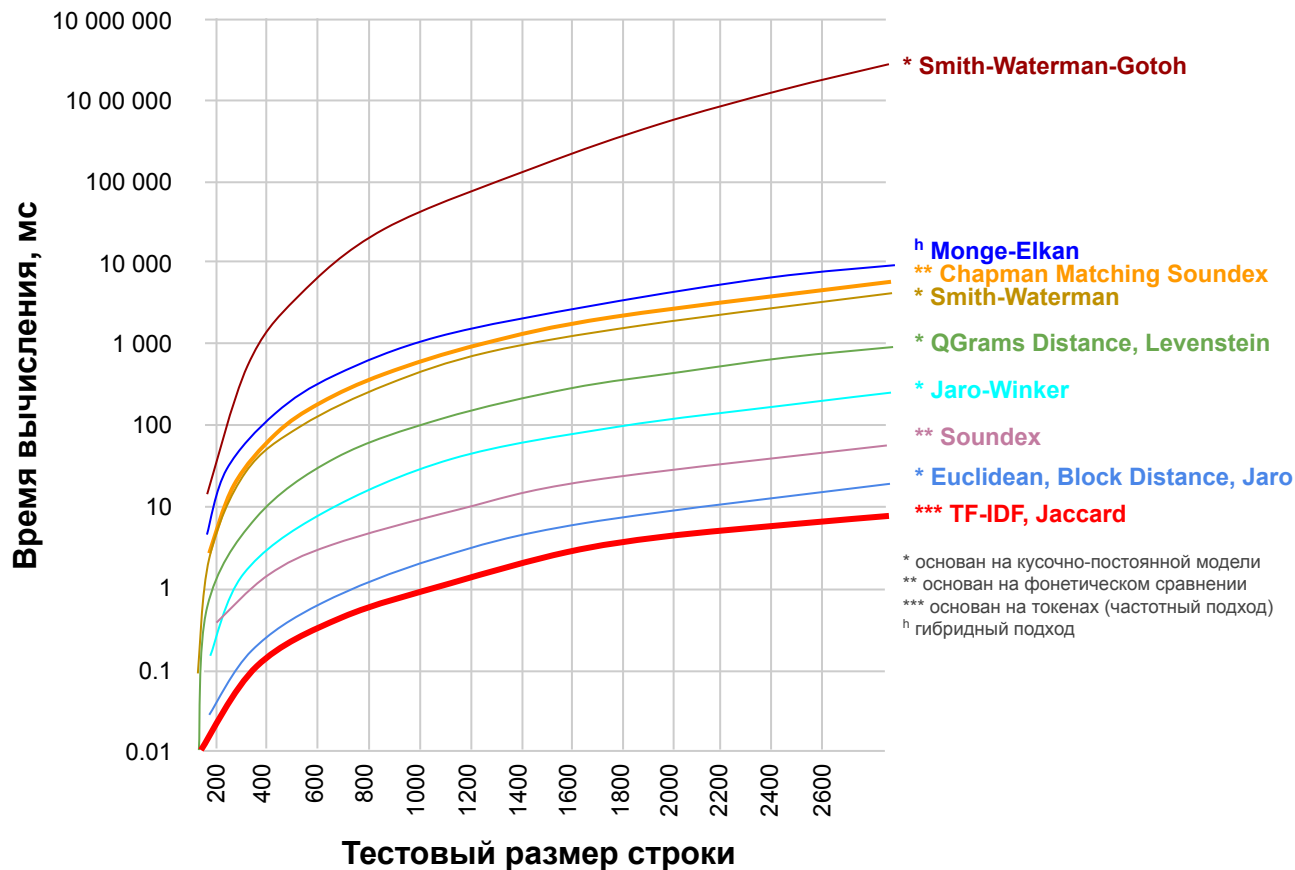
**Дано:** выборка распознанных документов

**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)

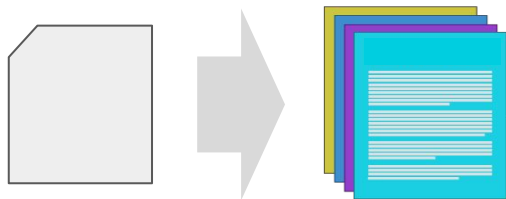
Department of Statistics, Carnegie Mellon University  
Cohen, William & Ravikumar, Pradeep & Fienberg, Stephen. (2003).  
A comparison of string DISTANCE metrics for name-matching tasks.  
IIWeb. 2003

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



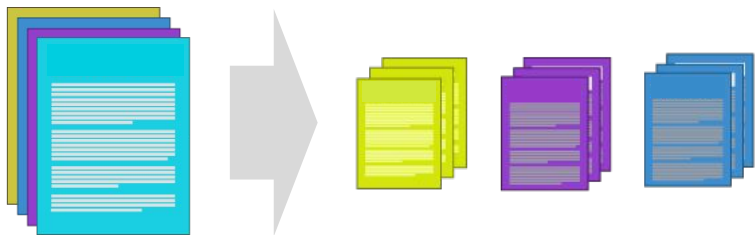
## Препроцессинг (Regex)

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

Решения для сегментации на токены	Ошибки на 1000 токенов	Время обработки, секунды
Regexp-baseline	19	<b>0.5</b>
<a href="#">SpaCy</a>	17	5.4
<a href="#">NLTK</a>	130	3.1
<a href="#">MyStem</a>	19	4.5
<a href="#">SegTok</a>	12	2.1
<a href="#">SpaCy Russian Tokenizer</a>	<b>8</b>	46.4

# Техническое описание решения

Определение тематики текста через вычисление веса слова



# Столько раз слово встретилось  
в документе

# Всего  
документов

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

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



Результат

Dadviser

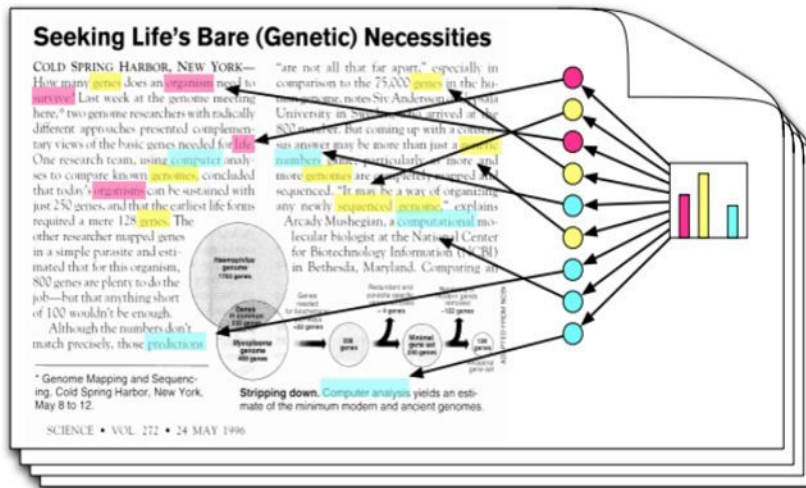
[Главная](#) [О нас](#) [Контакты](#)

Выберите файл для анализа

Choose File | No file chosen

Отправить

В тексте есть слова из **желтой** темы, из **красной** и из **синей**. Но нет слов из **зеленой**. Значит, документ **точно не на зеленую** тему



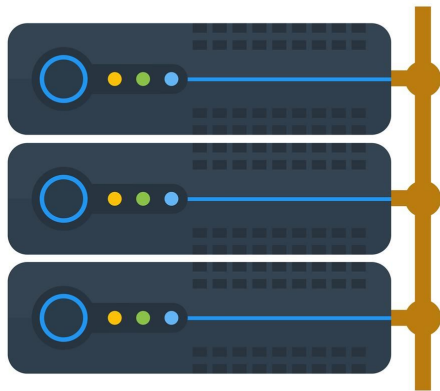
Док. №183	96%
Док. №471	35%
Док. №244	5.1%
...	
Док. №937	0.5%

# 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

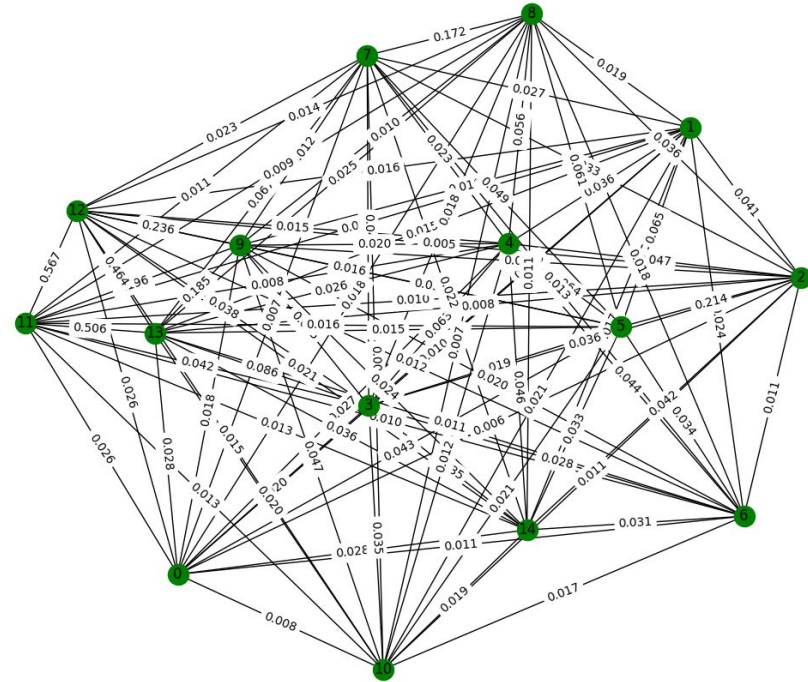
обучать 2 модели:

одну с данными за день,

а вторую – с данными за все время

# Тестирование решения

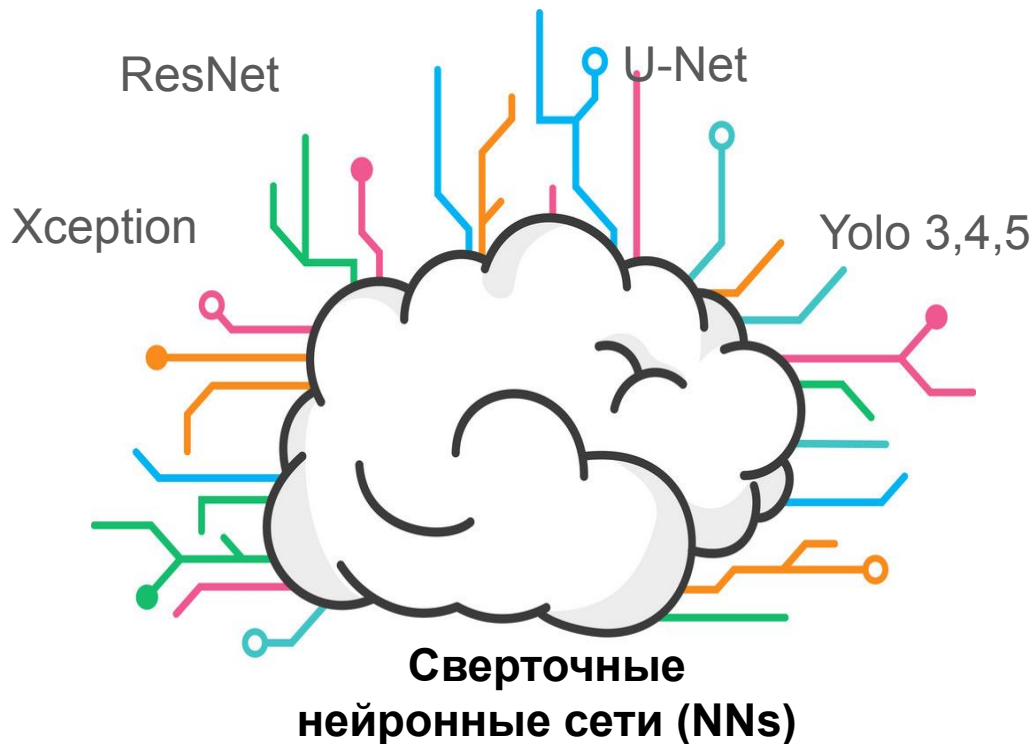
- визуальная проверка «схожести документов»
- построение графов похожих документов
  - «если документ A похож на документы B и C, то проверяем, есть ли еще документы, определяемые как похожие на B или C, но непохожие на A»
- проверка взаимной схожести документа
  - «если документ A похож на документы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



Пример графа из 15 документов с указанием схожести



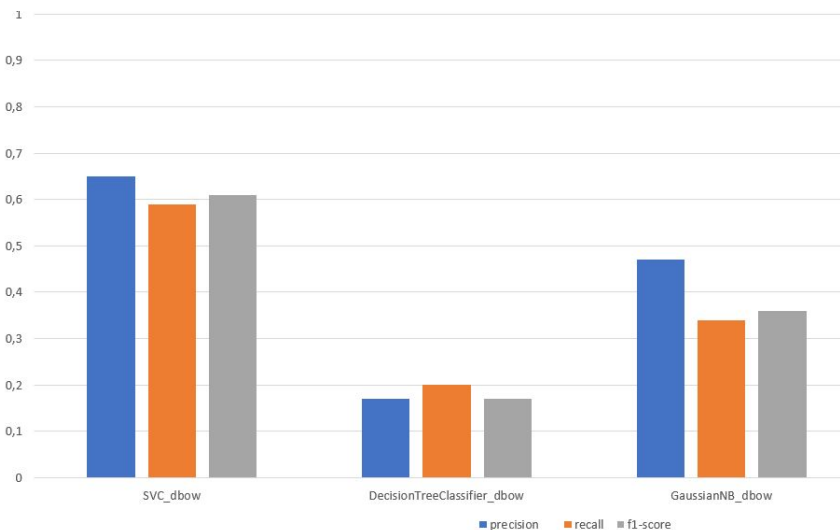
# Наши компетенции



# Наши возможности в Natural Language Processing

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

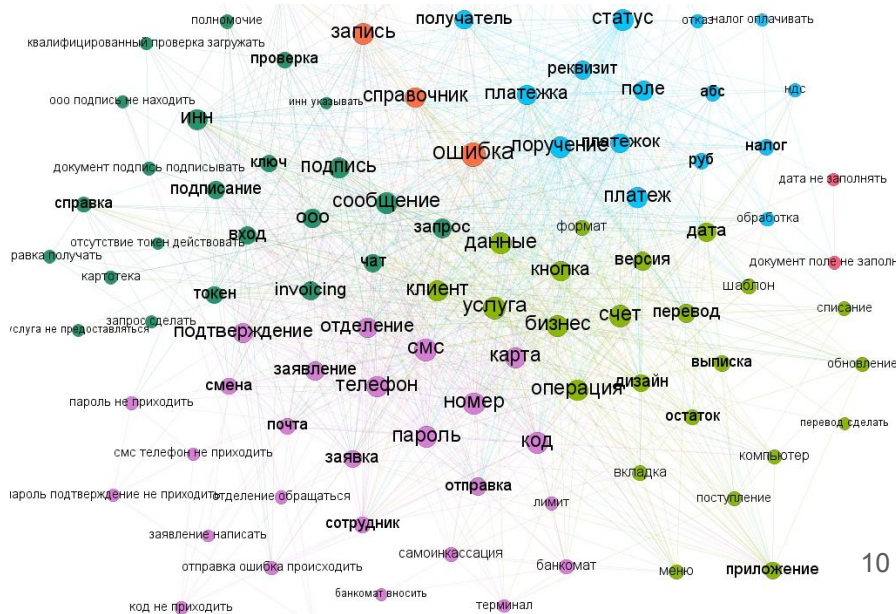
Обработка инцидентов с помощью Word2Vec



Семантический анализ текста (Latent semantic analysis, LSA)

на основе анализа tf-idf

Поиск сходных документов



Спасибо за внимание!

