Практическая работа №1.

Тямгин Иван

Латентно-семантический анализ.

Для обучения поисковой системы были использованы 15 тем по 50-75 документов. Всего 950 документов. Названия документов для каждой темы были получены с помощью поискового запроса в википедии вида: [http://ru.wikipedia.org/w/index.php?title=Служебная:Поиск&limit=**50**&offset=0&profile=default&search=**Таврия**](http://ru.wikipedia.org/w/index.php?title=Служебная:Поиск&limit=50&offset=0&profile=default&search=Таврия)

Затем скачаны по полученным ссылкам средствами языка python.

Были выбраны следующие темы:

* Тригонометрия
* Грипп
* Автомобиль
* Хлеб
* Москва
* Техника
* Таврия
* Биржа
* Депутат
* Газ
* Яндекс
* Водка
* Русский
* Видео
* Собака

Первым шагом были вырезаны из текстов все символы, кроме русских букв. Таким образом слово – это непрерывная последовательность русских букв. Каждое слово поддалось воздействию алгоритма Портера, чтобы одинаковые слова с разными окончаниями не считались как разные.

Была построена матрица 31238x950, которая означала частоту i-го слова в j-м документе. Матрица была нормализована с помощью TF-IDF, чтобы ключевые слова получили большой вес, а предлоги и союзы не значительный.

После чего матрица была разложена с помощью svd по 10-ми признакам. (разложение по 2-3 признакам дали существенно худший результат).

**Дендрограмма**

Далее построена иерархическая кластеризация документов. Расстоянием между двумя документами являлся угол между 10-мерными векторами.

Так как дендрограмма очень большая, приведена её часть:



Также попадались довольно абсурдные объединения документов на первом уровне:



Дерево было построено с помощью стандартной функции **hclust** в *R*. И нарисовано с помощью библиотеки **d3js** в *javascript*.

**Поиск**

Для каждого слова был построен чемпионский список – список документов и расстояние до этого документа. Чем меньше расстояние – тем он релевантнее для данного слова.

Поисковой запрос – неупорядоченный набор слов. Ответ на запрос – упорядоченный список документов по релевантности. Релевантность документа D - это

Примеры запросов:

{Москва, вокзал}

1. [https://ru.wikipedia.org/wiki/Казанский\_вокзал](https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D0%B7%D0%B0%D0%BD%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 91.13%
2. [https://ru.wikipedia.org/wiki/Ярославский\_вокзал](https://ru.wikipedia.org/wiki/%D0%AF%D1%80%D0%BE%D1%81%D0%BB%D0%B0%D0%B2%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 90.95%
3. [https://ru.wikipedia.org/wiki/Ленинградский\_вокзал](https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BD%D0%B8%D0%BD%D0%B3%D1%80%D0%B0%D0%B4%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 90.84%
4. [https://ru.wikipedia.org/wiki/Белорусский\_вокзал](https://ru.wikipedia.org/wiki/%D0%91%D0%B5%D0%BB%D0%BE%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 90.17%
5. [https://ru.wikipedia.org/wiki/Киевский\_вокзал](https://ru.wikipedia.org/wiki/%D0%9A%D0%B8%D0%B5%D0%B2%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 90.02%
6. [https://ru.wikipedia.org/wiki/Курский\_вокзал](https://ru.wikipedia.org/wiki/%D0%9A%D1%83%D1%80%D1%81%D0%BA%D0%B8%D0%B9_%D0%B2%D0%BE%D0%BA%D0%B7%D0%B0%D0%BB) 89.61%
7. [https://ru.wikipedia.org/wiki/Москва-Сити](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B0-%D0%A1%D0%B8%D1%82%D0%B8) 89.37%
8. [https://ru.wikipedia.org/wiki/Москва-Каланчёвская](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B0-%D0%9A%D0%B0%D0%BB%D0%B0%D0%BD%D1%87%D1%91%D0%B2%D1%81%D0%BA%D0%B0%D1%8F) 89.19%
9. [https://ru.wikipedia.org/wiki/Москва\_(гостиница\_в\_Москве)](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B0_(%D0%B3%D0%BE%D1%81%D1%82%D0%B8%D0%BD%D0%B8%D1%86%D0%B0_%D0%B2_%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B5)) 88.91%
10. [https://ru.wikipedia.org/wiki/Памятник\_Кутузову\_(Москва)](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BC%D1%8F%D1%82%D0%BD%D0%B8%D0%BA_%D0%9A%D1%83%D1%82%D1%83%D0%B7%D0%BE%D0%B2%D1%83_(%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B0)) 88.91%

{Яндекс}

1. [https://ru.wikipedia.org/wiki/Яндекс\_Нано](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81_%D0%9D%D0%B0%D0%BD%D0%BE) 99.88%
2. [https://ru.wikipedia.org/wiki/Яндекс.Календарь](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.%D0%9A%D0%B0%D0%BB%D0%B5%D0%BD%D0%B4%D0%B0%D1%80%D1%8C) 99.88%
3. [https://ru.wikipedia.org/wiki/Элементы\_Яндекса](https://ru.wikipedia.org/wiki/%D0%AD%D0%BB%D0%B5%D0%BC%D0%B5%D0%BD%D1%82%D1%8B_%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81%D0%B0) 99.87%
4. [https://ru.wikipedia.org/wiki/Яндекс.Услуги](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.%D0%A3%D1%81%D0%BB%D1%83%D0%B3%D0%B8) 99.85%
5. [https://ru.wikipedia.org/wiki/Яндекс.Открытки](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.%D0%9E%D1%82%D0%BA%D1%80%D1%8B%D1%82%D0%BA%D0%B8) 99.85%
6. [https://ru.wikipedia.org/wiki/Яндекс.Shell](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.Shell) 99.84%
7. <https://ru.wikipedia.org/wiki/Yandex.SpeechKit> 99.84%
8. [https://ru.wikipedia.org/wiki/Яндекс.XML](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.XML) 99.83%
9. [https://ru.wikipedia.org/wiki/Яндекс.Недвижимость](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.%D0%9D%D0%B5%D0%B4%D0%B2%D0%B8%D0%B6%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D1%8C) 99.83%
10. [https://ru.wikipedia.org/wiki/Яндекс.Навигатор](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81.%D0%9D%D0%B0%D0%B2%D0%B8%D0%B3%D0%B0%D1%82%D0%BE%D1%80) 99.83

Более простая версия формулы:

давала на вид более худший результат.