Отчёт по программе

Что делает программа?

Программа собирает статистику употребления токенов разного вида. В моём случае - это были токены:

- время в различных форматах(с AM/PM/am/pm и без, с секундами или без)
- номера телефонов с кодами (все номера телефонов, в том числе телефоны, в которых цифры разделены на группы в любом месте одним пробелом)
- номера пластиковых карт(разной длины: 4 – 3 – 3 – 3, 4 – 6 – 5, 4 – 4 – 4 – 4, 8 - 10. В качестве разделителя можно использовать точку, тире, пробел или писать слитно)
- даты(dd.mm.yyyy или dd-mm-yy)
- адреса электронной почты
- IPv4-адреса
- имена
- организации
- локации

Как результат мы можем получить графики, в которых можно увидеть долю токенов конкретного вида в тексте, далее выводятся графики с частотами самых популярных токенов. Программа осуществляет графематический анализ текста, разделяя его на токены конкретного вида. Также программа выводит всю важную информацию в файл «Tokens_global.txt». Там можно найти все токены, которые программа смогла выделить в файле. Название файла указывается на старте программы

Реализация

Я использовала средства библиотеки Natasha для осуществления графематического анализа текста. После применения метода segment() к объекту класса Doc, мы получаем текст, который разделён на токены в специализированном представлении.

В программе есть функции для формирования списков токенов каждого из выделенных видов.

- def form_span_list(list_spans) для формирования списка токенов-имён. В ней я применяла метод extract_fact(), который сформировывал мне поле fact, в котором находилось имя-токен.
- def form_location_list(list_spans) для формирования списка токенов-локаций. В ней использовались результаты работы метода tag_ner(), который выделяет поле spans, в

котором и можно найти тип конкретного span(a). Я сравнивала type с ‘LOC’ и таким образом формировала список.

- def form_org_list(list_spans) для формирования списка токенов-организаций. Она работает аналогично функции form_location_list.

Для остальных токенов я использовала средства библиотеки nltk. Если быть точнее, использовала метод tokenize класса RegexpTokenizer. Имелись шаблоны-регулярные выражения, при помощи которых и фильтровались нужные токены.

- def make_dict(list) Функция для формирования частотного словаря из списка.
- def find_max(list) Функция для формирования словаря-рейтинга трёх самый популярных элементов в заданном списке.
- def make_list_words(text) Функция для формирования списка из токенов. Именно она осуществляет графематический анализ. Как результат мы получаем список токенов, они неспицифичны, происходила фильтрация на уровне знаков препинания - они в список не добавлялись.

Тесты и выводы

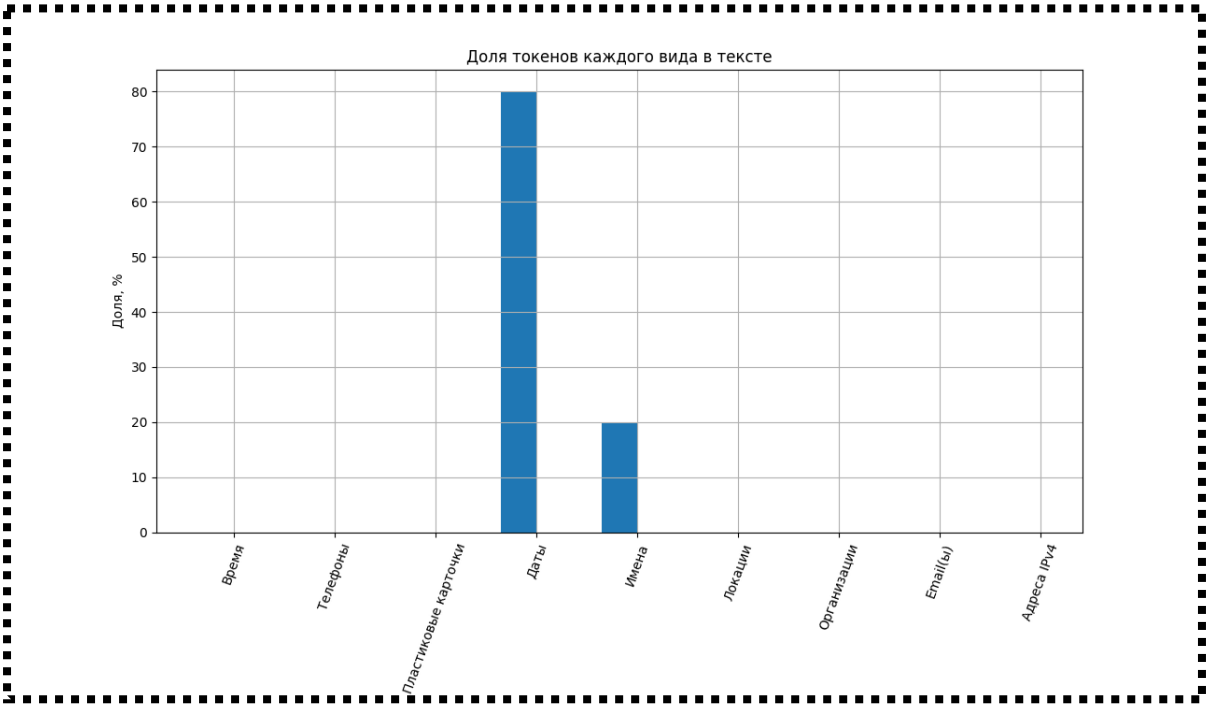
Чтобы протестировать работу программы, я решила для каждого вида токенов найти текст, в котором этот токен будет превалирующим.

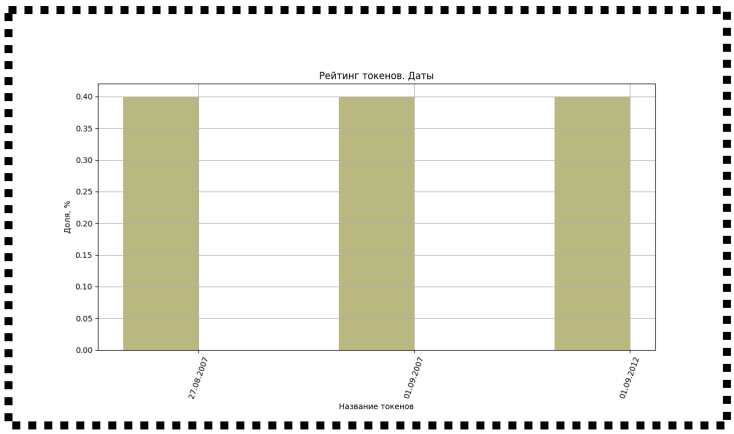
ТОКЕНЫ-ДАТЫ

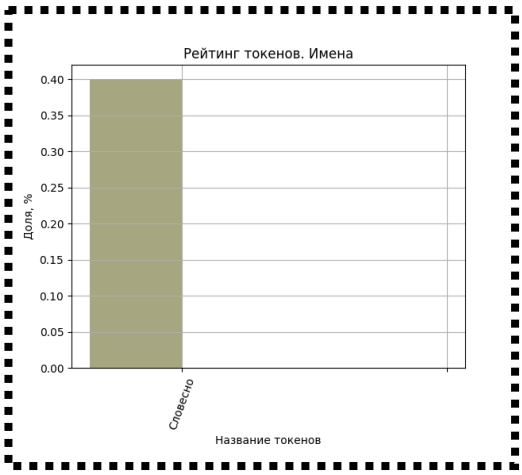
Материал : статья о датах.

Название файла: Dates.txt

Полученные графики:





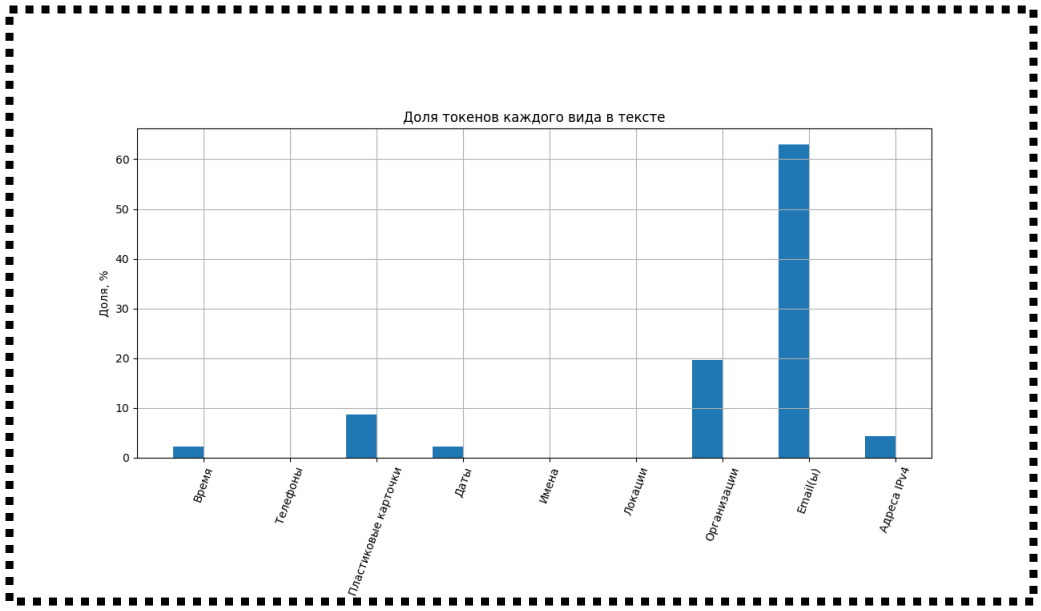


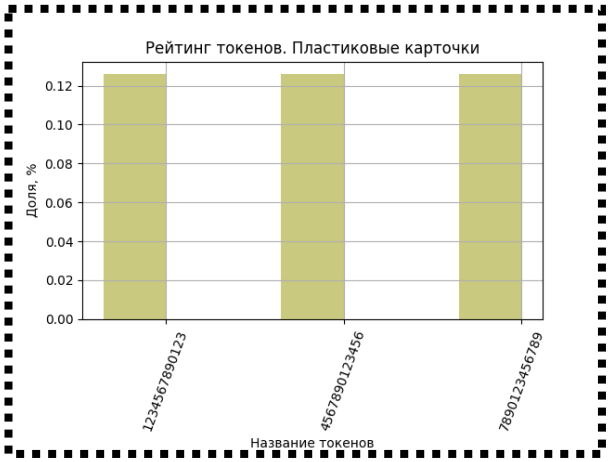
ТОКЕНЫ-Email(ы)

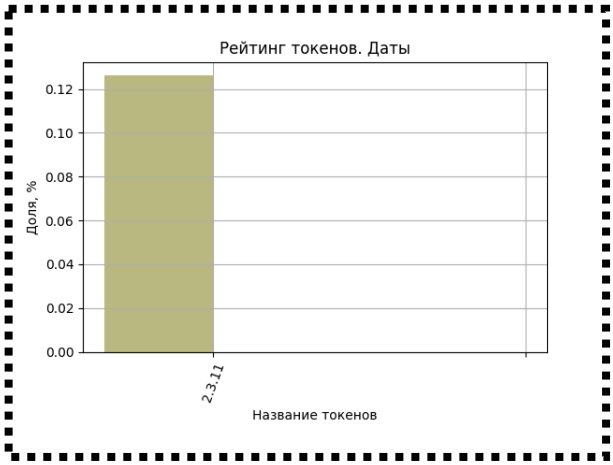
Материал : статья об адресах электронной почты.

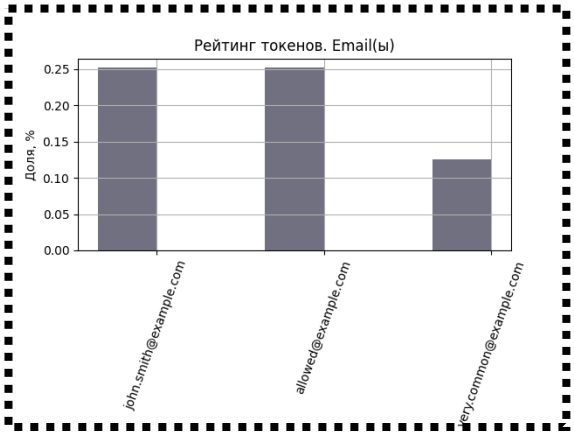
Название файла: Email.txt

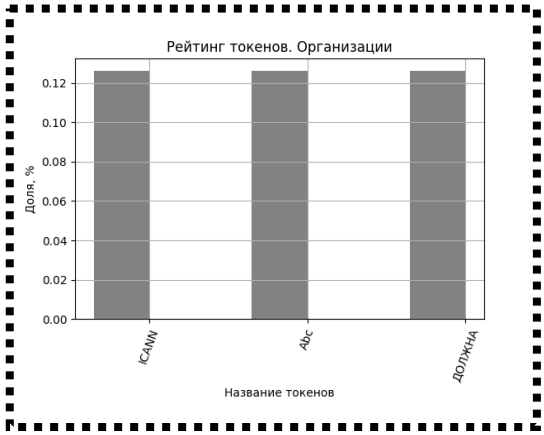
Полученные графики:

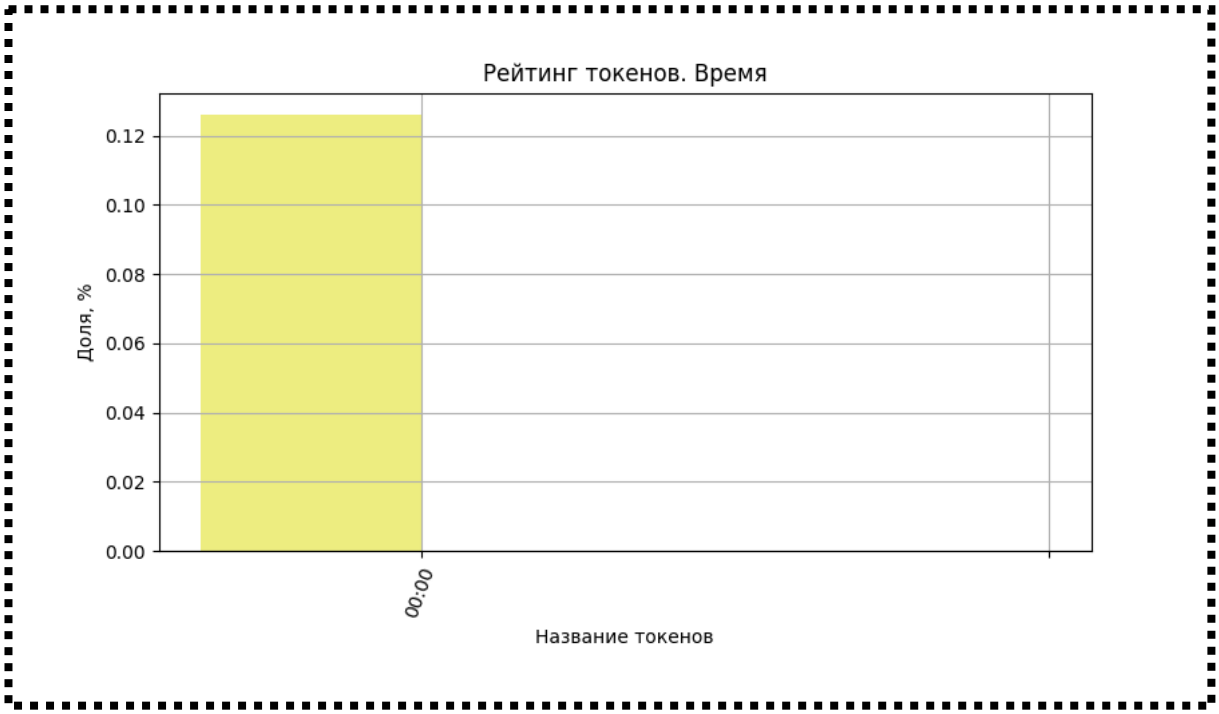










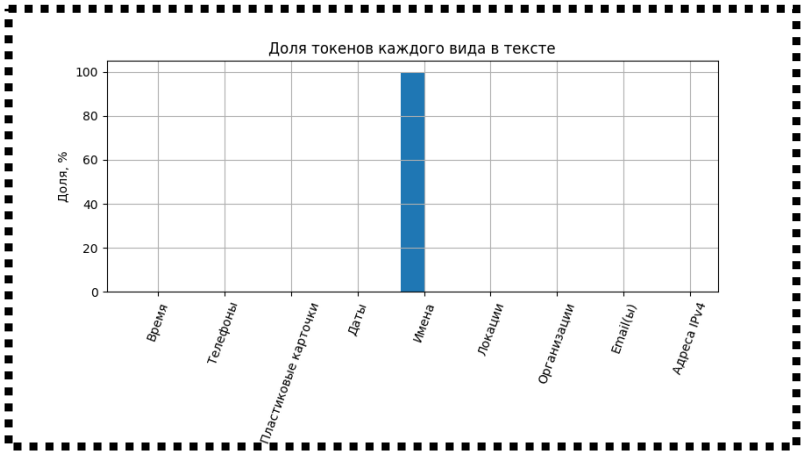


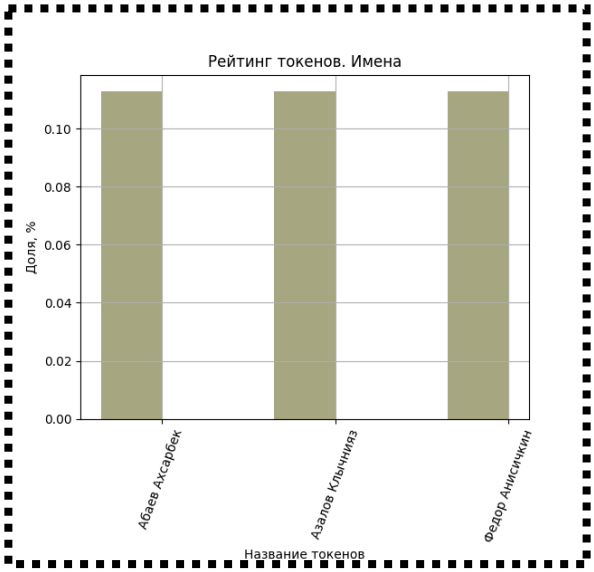
ТОКЕНЫ-ИМЕНА

Материал : список героев Советского Союза

Название файла: Names.txt

Полученные графики:



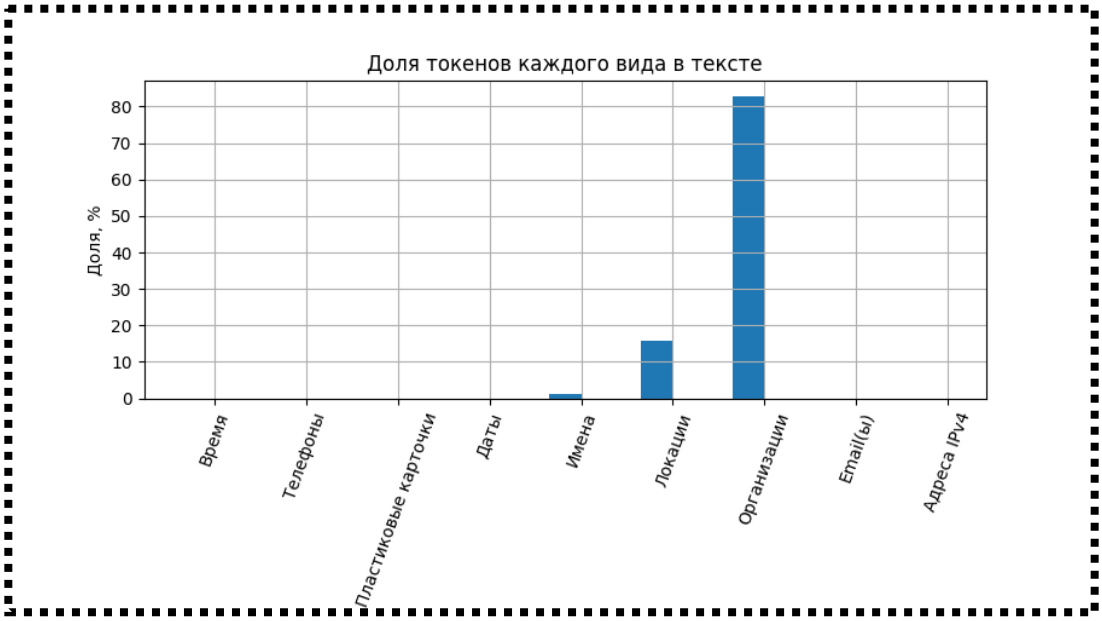


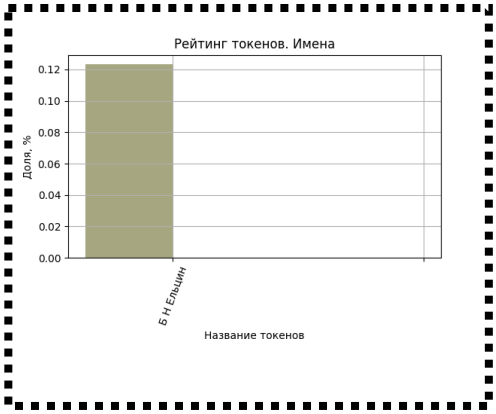
ТОКЕНЫ-ОРГАНИЗАЦИИ

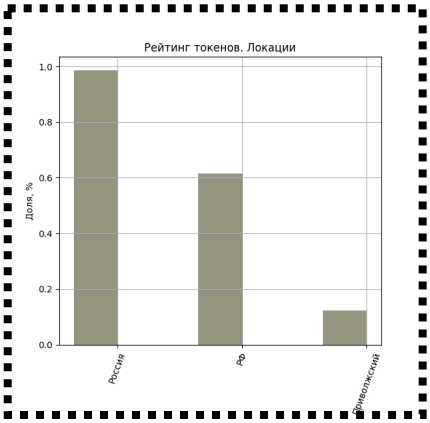
Материал : рейтинг ВУЗов

Название файла: Organisations.txt

Полученные графики:





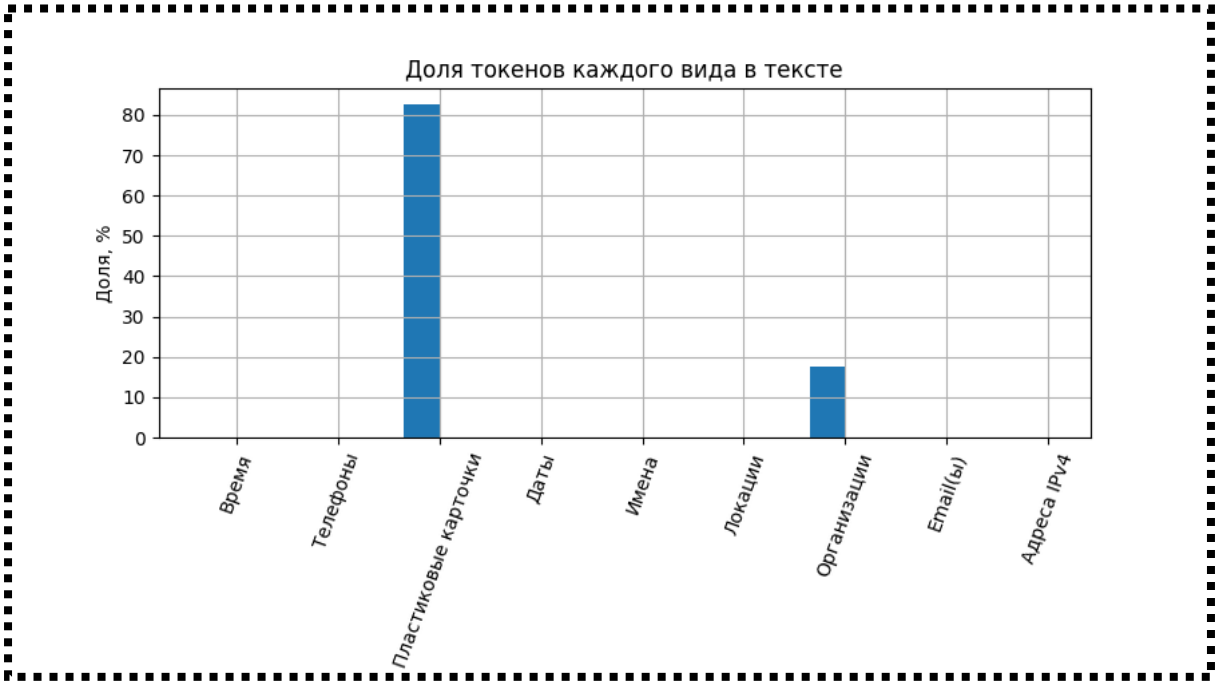


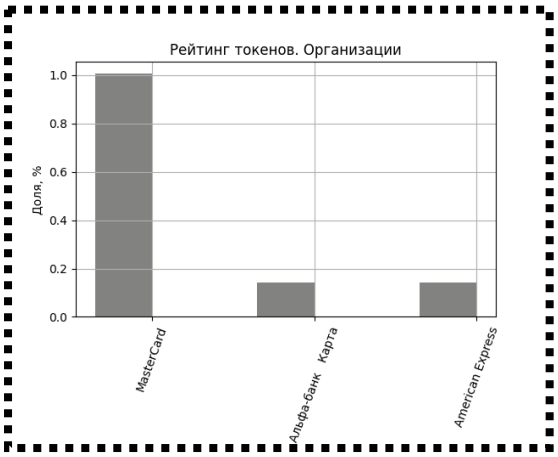
ТОКЕНЫ-ПЛАСТИКОВЫЕ КАРТОЧКИ

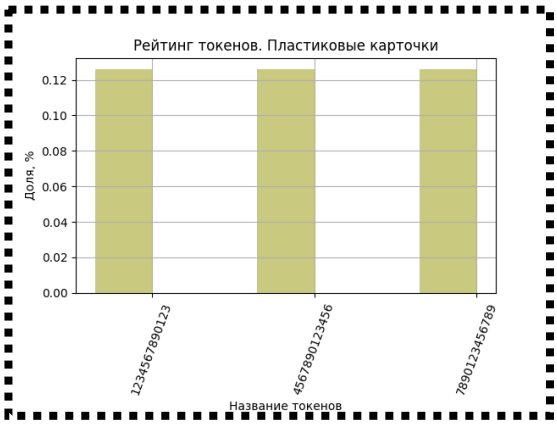
Материал : статья о форматах пластиковых карточек

Название файла: Cards.txt

Полученные графики:





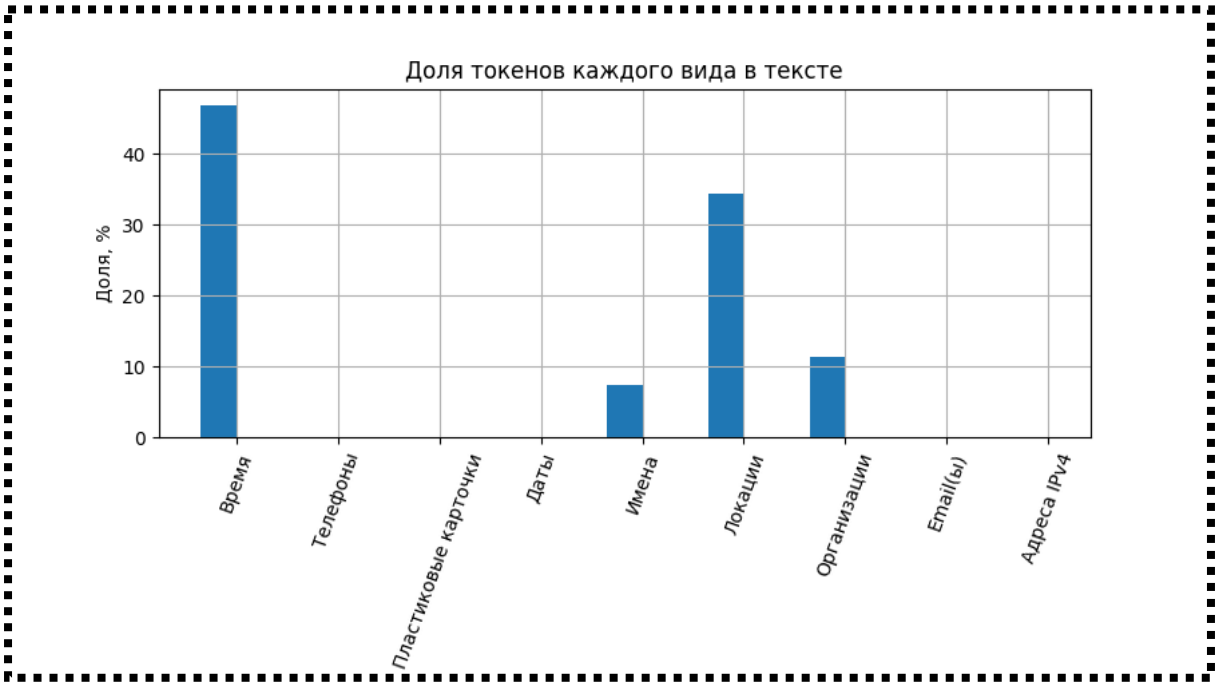


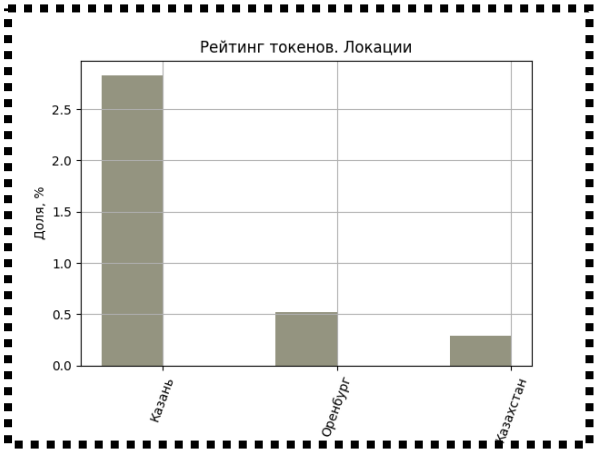
ТОКЕНЫ-ВРЕМЯ

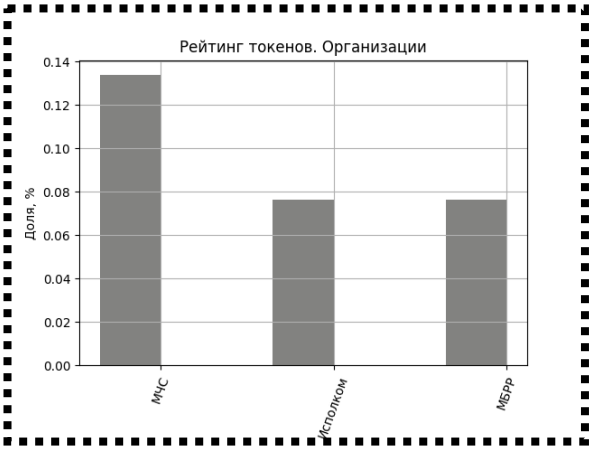
Материал : архив новостей, в котором присутствуют временные метки

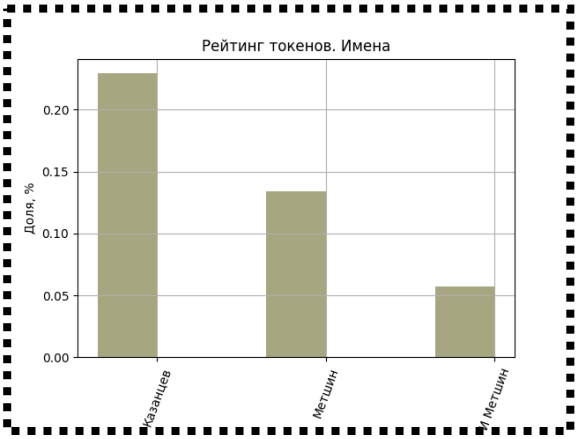
Название файла: Time.txt

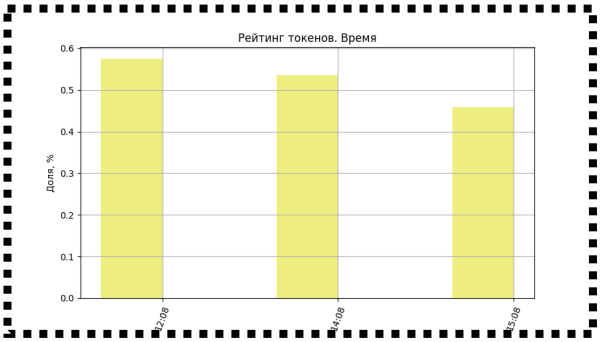
Полученные графики:









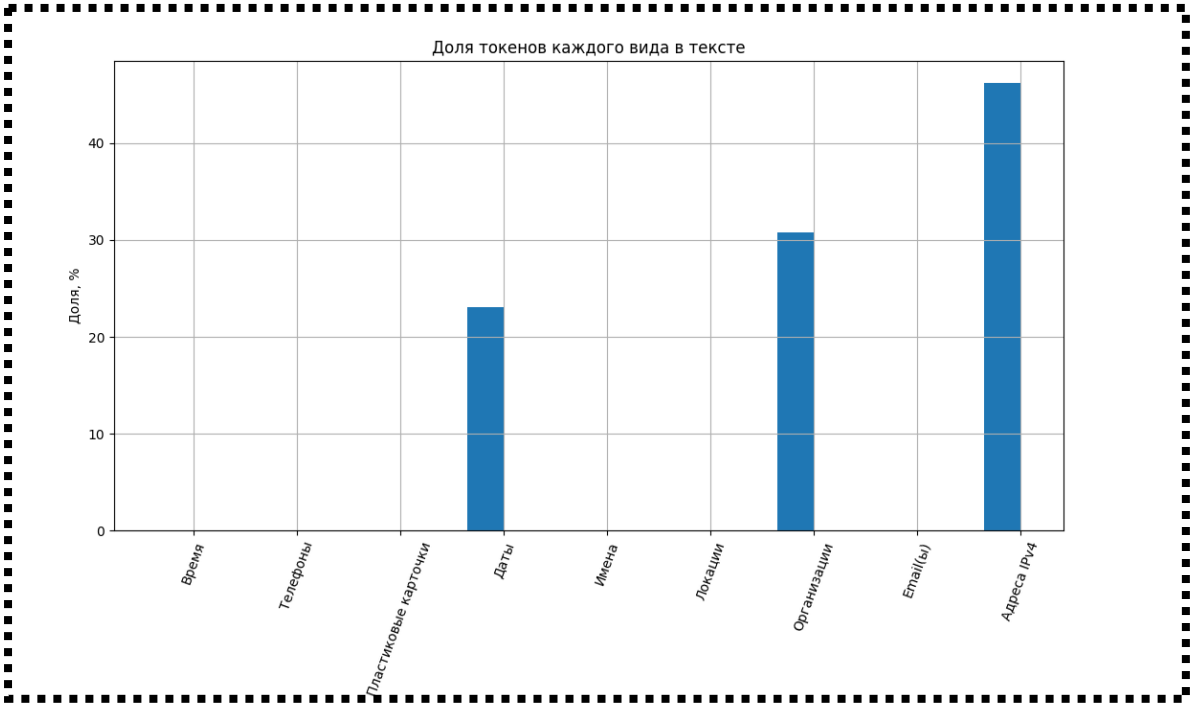


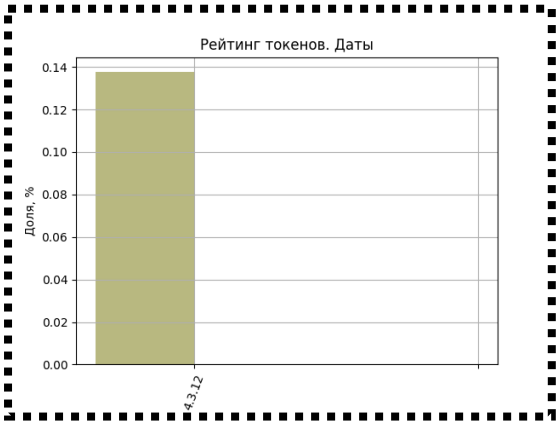
ТОКЕНЫ-адреса IPv4

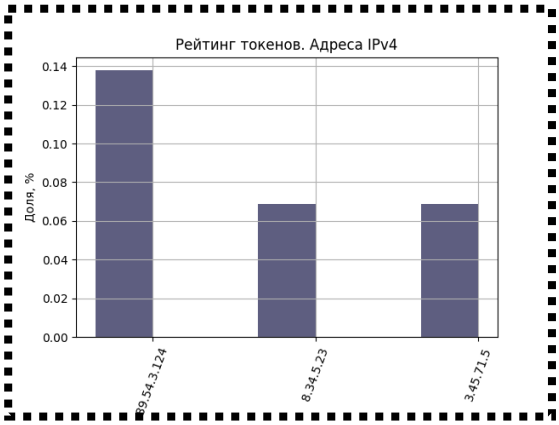
Материал : статья об адресах IPv4

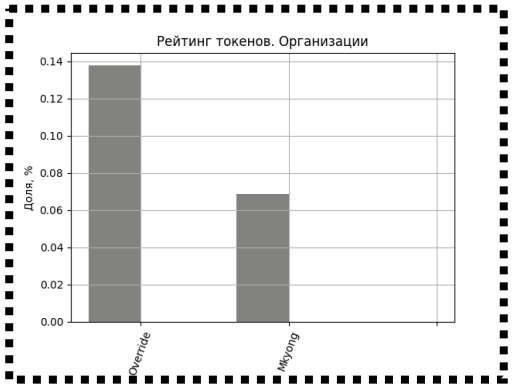
Название файла: IPv4.txt

Полученные графики:







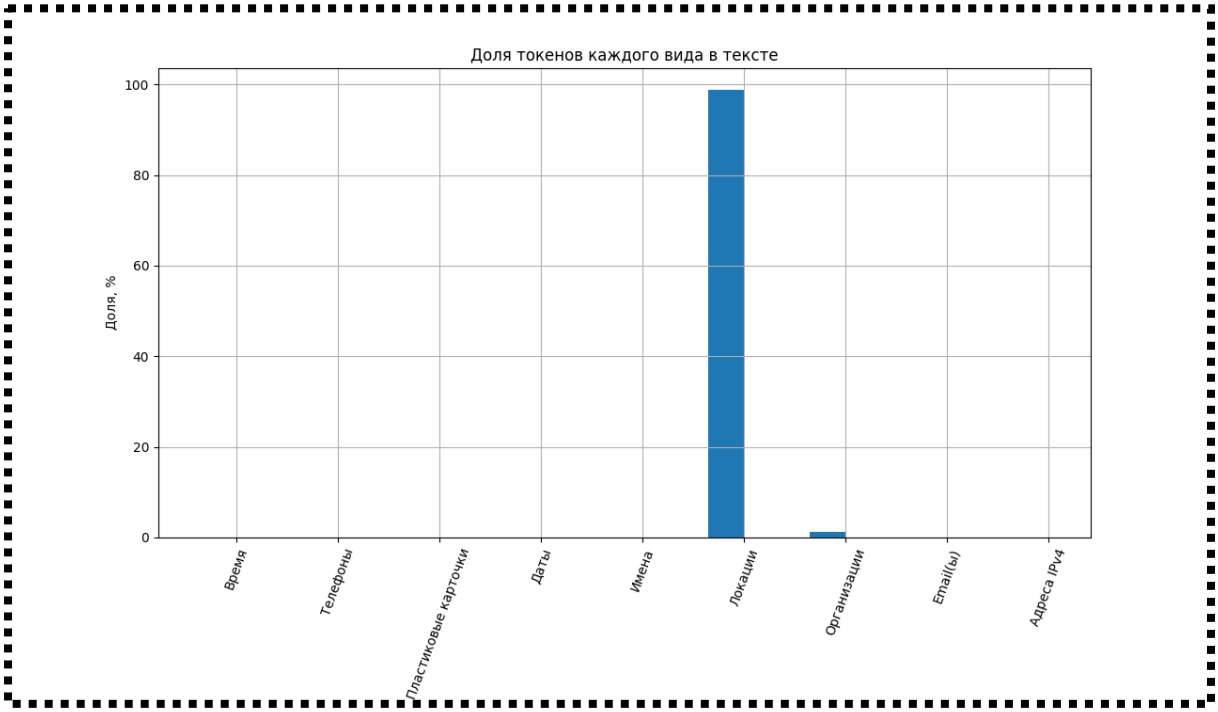


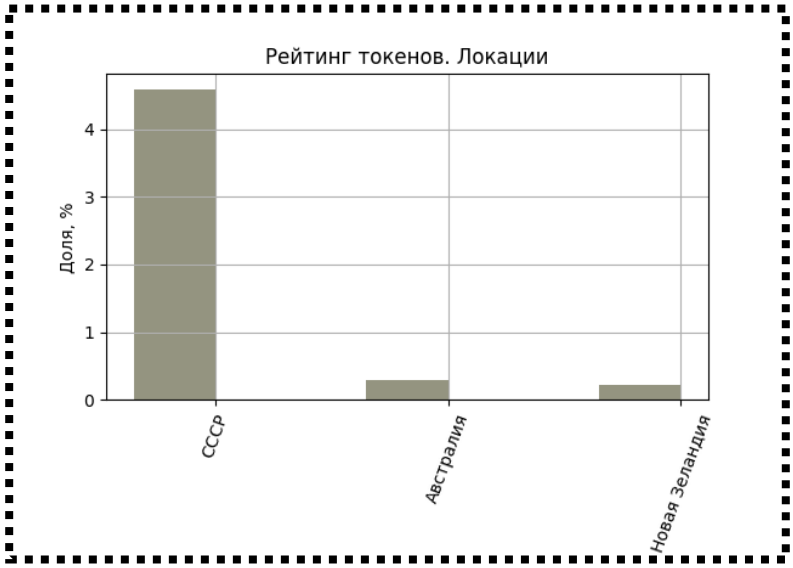
ТОКЕНЫ-ЛОКАЦИИ

Материал : статья об адресах IPv4

Название файла: Locations.txt

Графики:



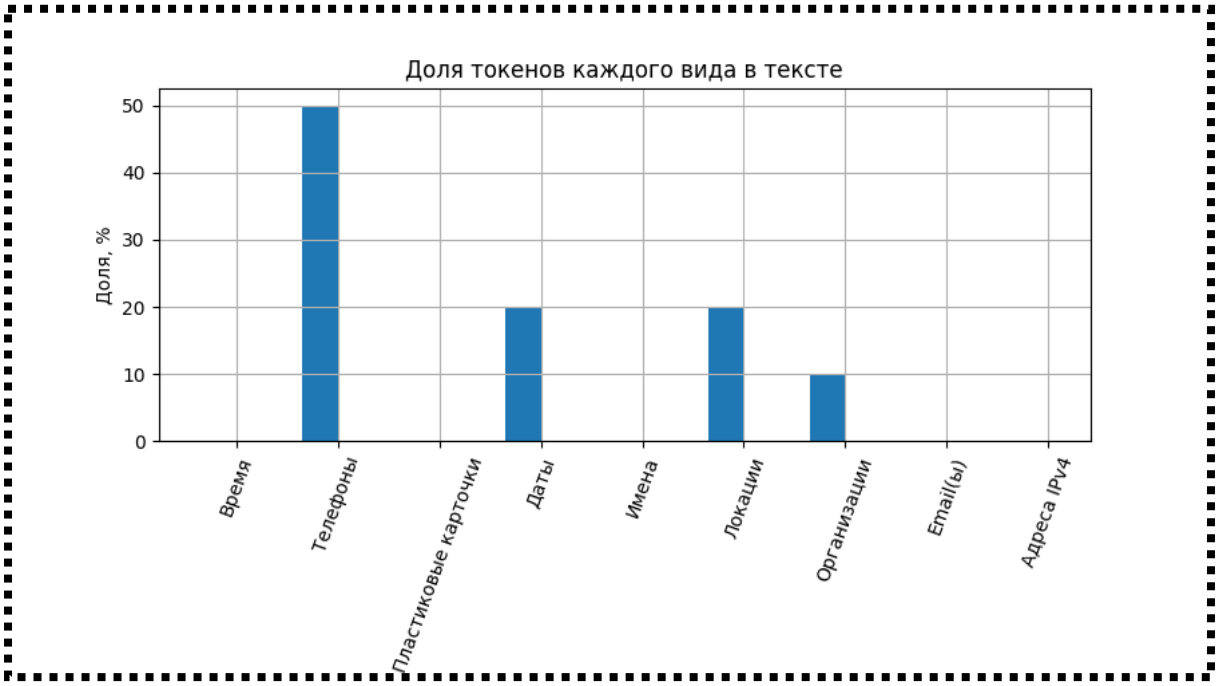


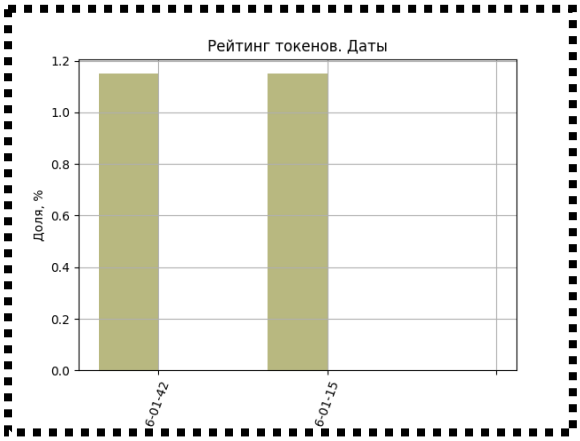
ТОКЕНЫ-ТЕЛЕФОНЫ

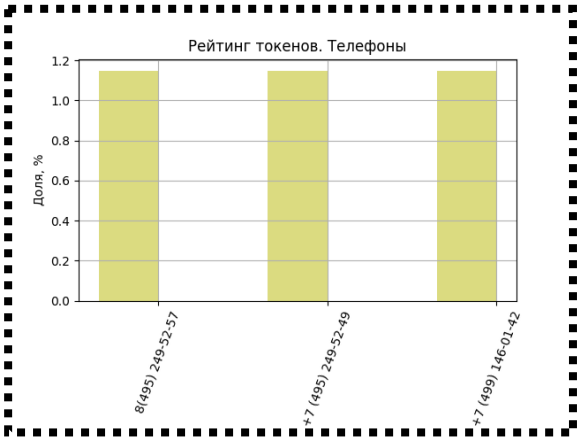
Материал: Контактная информация Финансовой академии

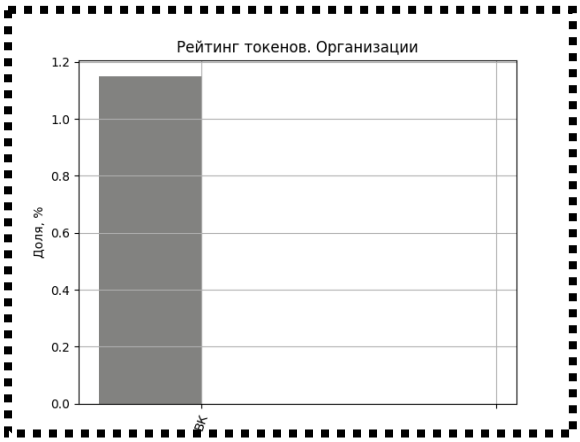
Название файла: Telephones.txt

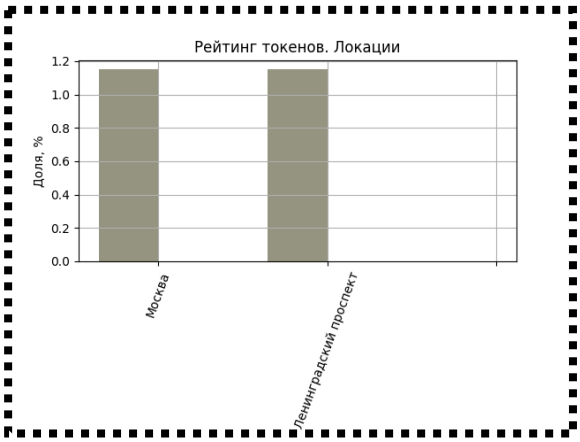
Полученные графики:


Доля токенов каждого вида в тексте
Доля, %
0
10
20
30
40
50
Время
Телефоны
Пластиковые карточки
Даты
Имена
Локации
Организации
Email(ы)
Адреса IPv4


Рейтинг токенов. Даты
Доля, %
0.0
0.2
0.4
0.6
0.8
1.0
1.2
6-01-42
6-01-15


Рейтинг токенов. Телефоны
Доля, %
0.0
0.2
0.4
0.6
0.8
1.0
1.2
8(495) 249-52-57
+7 (495) 249-52-49
+7 (499) 146-01-42


Рейтинг токенов. Организации
Доля, %
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ВК


Рейтинг токенов. Локации
Доля, %
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Москва
Ленинградский проспект

Пример прикладной программы

Программа может быть полезна при обработке заявлений абитуриентов в Московский Университет. Конечно, на сайте вуза выложен образец заявления, но очень часто случаются недоразумения, и абитуриенты заполняют форму неправильно. Например, формат адреса не тот, данные о ФИО неверны, почтовый индекс не указан и т.д. Для того, чтобы заявление абитуриента могло быть корректно зарегистрировано, оно должно быть вручную проверено сотрудниками Центральной Приёмной Комиссии факультета, которыми часто являются студенты факультета, на которое было подано обращение. Ребята выполняют однотипную работу, которая осложняется тем, что нужно обрабатывать и лишнюю информацию, которая вводится по умолчанию. Гораздо легче было бы после конвертации файла в формат txt вывести основные данные в файл, в котором всё быстро посмотреть и при совпадении формата, нажать кнопку подтверждения.

Ректору МГУ имени М.В.Ломоносова
академику В.А. Садовничему

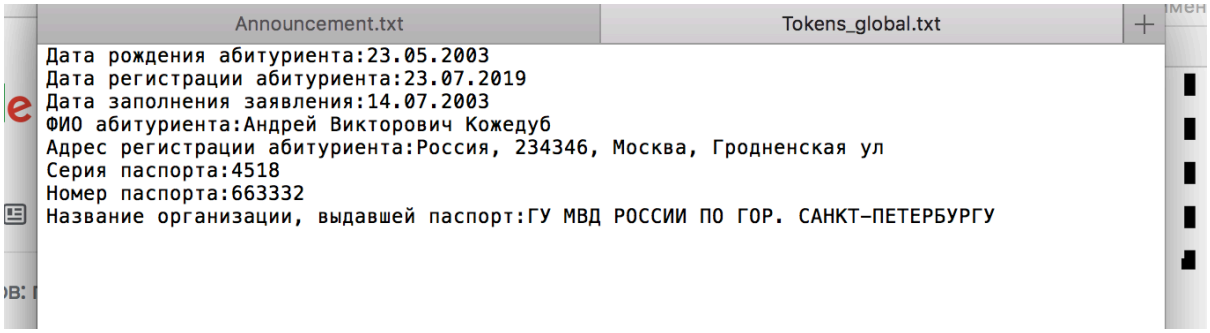
Заявление

Я, ФИО, дата рождения дата рож.,
зарегистрированный по адресу страна, индекс, город, улица, дом
«Паспорт РФ», серия, номер, дата выдачи, организация, код подразделения
,
в соответствии со статьей 9 Федерального закона от 27 июля 2006 г. No 152-ФЗ "О персональных данных" (далее – Закон) даю согласие Федеральному государственному бюджетному образовательному учреждению высшего образования «Московский государственный университет имени М.В.Ломоносова» (119991, Российская Федерация, Москва, Ленинские горы, д. 1) на обработку автоматизированным и неавтоматизированным способами своих персональных данных, необходимых для осуществления образовательной деятельности, в том числе переданных мной в настоящем заявлении, в том числе посредством сети Интернет в Личный кабинет абитуриента, а также полученных в ходе осуществления образовательной деятельности, а именно совершение действий, предусмотренных пунктом 3 статьи 3 Закона, а также передача информации о ходе и результатах рассмотрения заявления о приеме и иных заявлений в федеральную государственную информационную систему "Единый портал государственных и муниципальных услуг (функций)" с целью осуществления образовательной деятельности по образовательным программам высшего образования в соответствии с законодательством об образовании Российской Федерации.

Дата: _____

Подпись: _____

После обработки текста заявления моей программой(текст заявления в файле «Announcement.txt»), можно получить следующий файл, в котором находится информация для проверки:



Как мне кажется, стало намного удобнее.

Библиография

https://fortress-design.com/kak-pisat-daty/ - статья о датах

https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D1%80%D0%B5%D1%81_%D1%8D%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D0%BE%D0%B9_%D0%BF%D0%BE%D1%87%D1%82%D1%8B - статья об адресах электронной почты

https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B8%D1%81%D0%BE%D0%BA_%D0%93%D0%B5%D1%80%D0%BE%D0%B5%D0%B2_%D0%A1%D0%BE%D0%B2%D0%B5%D1%82%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D0%A1%D0%BE%D1%8E%D0%B7%D0%B0 - список Героев Советского Союза

https://www.kommersant.ru/doc/5393508 - рейтинг ВУЗОВ

https://ria.ru/20220501/-архив новостей с временными метками

https://habr.com/ru/post/679008/ - статья об Ipv4 адресах

http://www.fa.ru/priemka/Pages/contacts.aspx - контактная информация финансового университета

https://mattweb.ru/moj-blog/raznoe/item/142-30-primerov-regulyarnykh-vyrazhenij - сайт, который дал полезные регулярные выражения для IPv4-адресов, адресов электронной почты

Автор регулярных выражений:

Номера телефонов - Дмитрий Моисеев

Даты - Алексей Волков

Номера пластиковых карточек - Тимур Пак