# Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

# «Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

Лабораторная работа №4 по дисциплине:

«Компьютерная лингвистика»

**Выполнила**: студентка 2 курса ИКТ

# группы K3242 Плотская Дарья Александровна

**Проверила**: Чернышева Анастасия Вадимовна

# Санкт-Петербург 2021

**Цель**: произвести предобработку текстов собранного

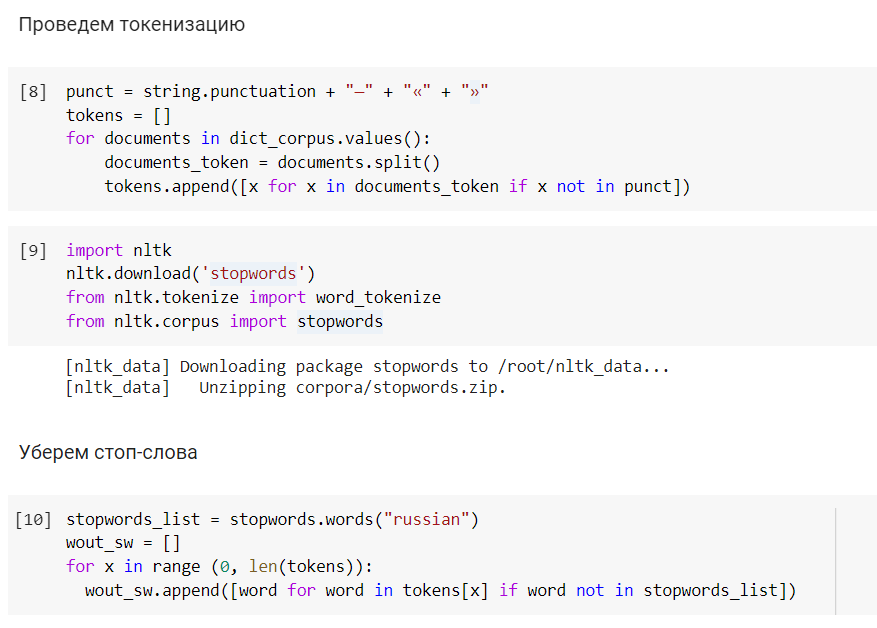
корпуса. Сделать частеречную разметку, подсчитать средние доли для разных частей речи в разных документах, найти нетипичные документы (доля отклоняется более чем на 2 стандартных отклонения).

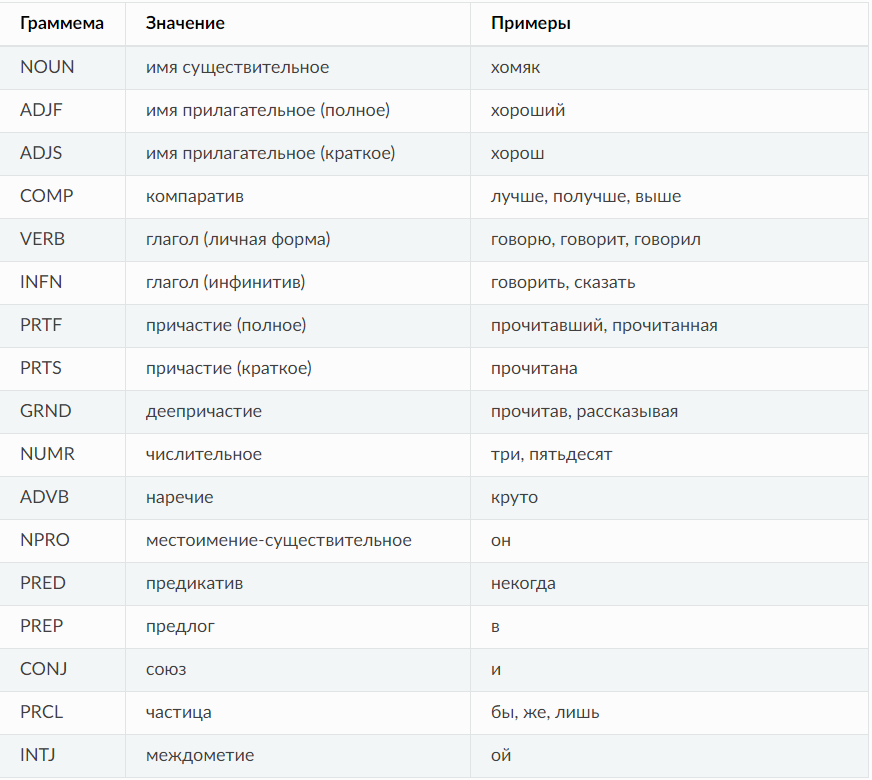
Ссылка на коллаб: <https://colab.research.google.com/drive/1Ftj09V6v7pIealjPUxV56HrU81ey7Ha1#scrollTo=rrGY6kONfL0y>

Импортируем необходимые библиотеки, открываем наш корпус

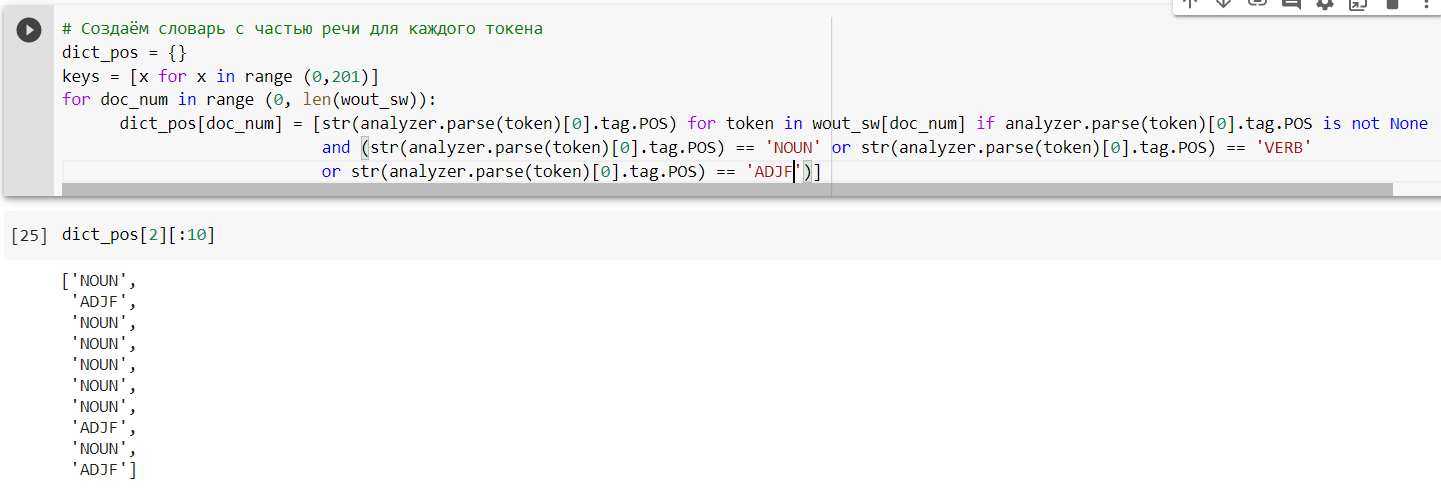


Проведем токенизацию, уберем стоп-слова

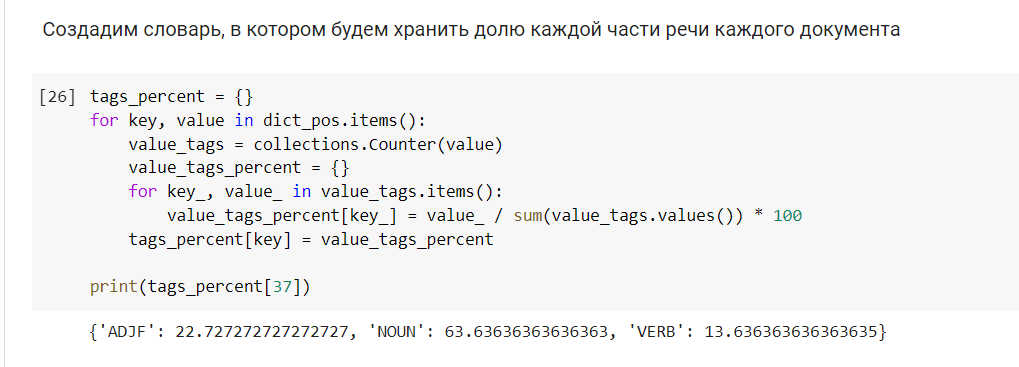
Перед тем, как проводить частеречную разметку, посмотрим, какие граммемы знает python. Всего их 17. Мы будем анализировать только существительные (NOUN), глаголы (VERB) и полные прилагательные (ADJF).



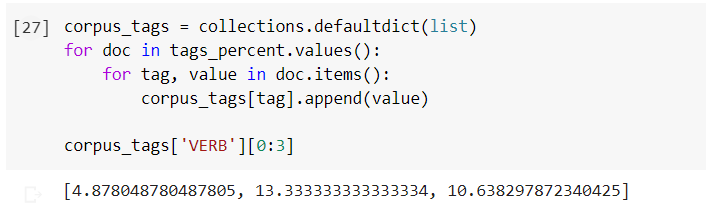
Создадим новый корпус, где будем хранить часть речи для каждого токена. Отбираем только 3 части речи, упомянутые выше.



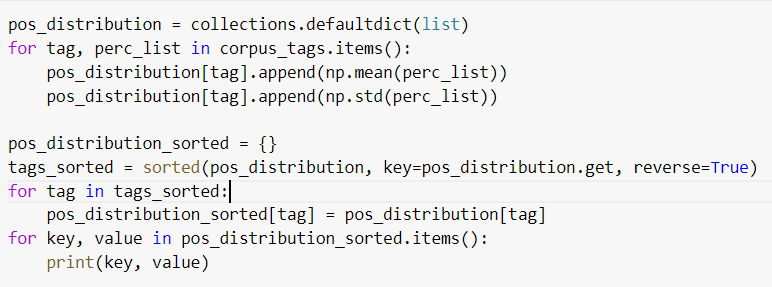
Создадим словарь tags\_percent, в который мы будем записывать долю в процентах существительных, глаголов и прилагательных каждого документа. Пользуемся Counter из библиотеки collections, чтобы создать автоматический словарь с ключом часть речи, а значением – ее доля в документе.



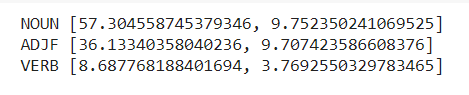
Создаем словарь corpus\_tags (defaultdict означает, что если ключ не найден в словаре, то вместо KeyError создается новая запись). Он содержит в ключе часть речи, а в значении – доля этой части речи в каждом документе.



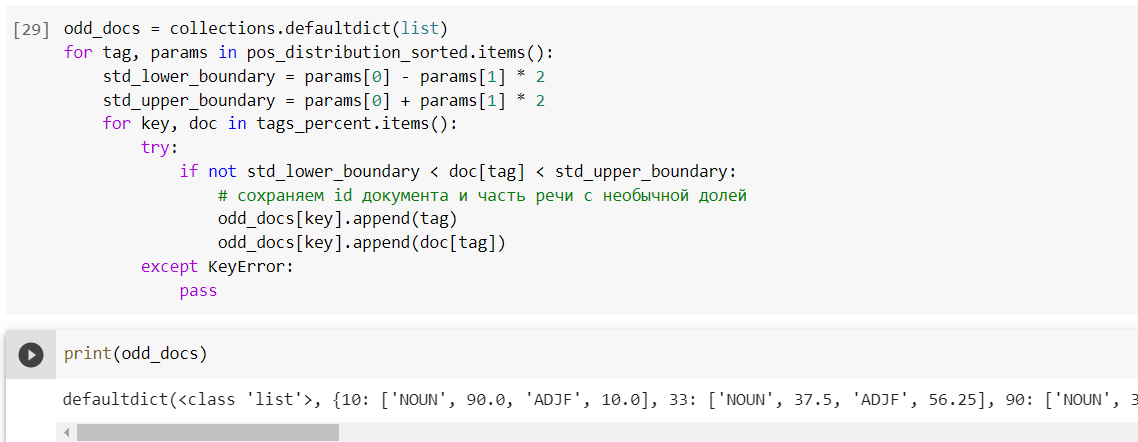
Создаем словарь pos\_distribution, в котором мы будем хранить в ключе часть речи, а в значении – среднее значение доли этой части речи во всех документах и стандартное отклонение доли этой части речи во всех документах.



Результат:

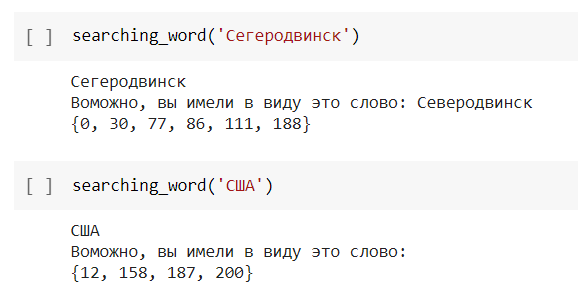


Создаем словарь odd\_docs, где будем хранить в ключе id документа, у которого доля в документе нестандартная, а в ключе – часть речи и ее долю в данном документе. Определяем верхнюю и нижнюю границы для оценки доли части речи в документе. Если доля части речи не находится между ними, то добавляем ее в odd\_docs. Выводим результат



Видим, что, например, в 10 документе у нас 90% существительных. Это действительно нестандартная доля части речи.

Результат работы программы:



Выведем результат для 104 документа:

