Лабораторная работа 1

Цель: освоить основные инструменты для предварительной обработки и первичного анализа текстовых данных

Задания:

1) Настройка виртуальной среды разработки conda
    a. Если на вашем ноутбуке уже установлена Miniconda или альтернативный инструмент создания и настройки виртуальной среды разработки Python (Anaconda, virtualenv, venv, poetry, uv,…), в котором можно открывать, редактировать и запускать код из файла jupyter notebook, можно переходить сразу к пункту 2.

    b. Скачать и установить Miniconda, следуя инструкции "Quickstart install instructions" для вашей системы на официальном сайте https://www.anaconda.com/docs/getting-started/miniconda/install#quickstart-install-instructions
    Выбор обусловлен тем, что Anaconda является стандартным и простым в использовании инструментом создания виртуальной среды для проектов в области анализа данных. Но вы можете использовать и другие инструменты на ваше усмотрение.

    c. Найти в списке программ и запустить Anaconda Prompt (на Mac OS обычный терминал)

    d. После успешной установки в терминале создать и активировать среду с tensorflow для CPU, загрузить в неё нужные библиотеки:

    ```
    conda create -n lab_env tensorflow
    conda activate lab_env
    conda install numpy pandas matplotlib seaborn scikit-learn
    pip install pymorphy3 nltk
    conda install jupyter
    ```

2) Выполните задания в ноутбуке Семинар_2_Лабораторная.ipynb, ориентируясь на Ваш вариант (смотрите таблицу в конце файла). Данные для обработки находятся в файле Lesson_1_user_requests.csv.

У каждого студента свой вариант, номер варианта совпадает с вашим номером в списке группы. Вариант определяет:

- Целевую переменную, которую модель должна будет предсказывать: «sphera» или «categoriya»

- Атрибут num_words класса Tokenizer(), определяющий количество самых часто встречающихся слов из сформированного словаря, которые будут использоваться при токенизации текста

- Какую технику нормализации необходимо применить: лемматизация или стемминг

- Необходимо ли исключать некоторые классы из процесса обучения

- Необходимо ли задавать дополнительные условия при разбиении данных на тренировочную и тестовую выборки

- Значение параметра «mode» функции texts_to_matrix() для векторизации текста

3) Покажите полностью выполненный ноутбук с заполненными пропусками и ответьте на вопросы. После этого мы отметим, что вы сдали лабораторную работу.

Вариант		Цель	num_words	Лемматизация/ Стемминг	Дополнительная работа над целевой переменной	Дополнительная работа над разбиением на тренировочную и тестовую выборки	mode в texts_to_matrix()
1	Акимов Александр Алексеевич	sphera	60000	Лемматизация	исключить классы, в которых меньше 200 обращений	нет	binary
2	Андреев Иван Александрович	categoriya	10000	Лемматизация	исключить классы, в которых меньше 300 обращений	нет	count
3	Бадертдинов Дильназ Дарсинович	sphera	1000000	Лемматизация	нет	обеспечить пребывание всех классов в тестовой выборке	tfidf
4	Баранов Владимир Георгиевич	categoriya	30000	Лемматизация	нет	обеспечить сохранение распределения по классам в тренировочной и тестовой выборках	freq
5	Басманов Дмитрий Алексеевич	sphera	1000	Лемматизация	исключить классы, в которых меньше 500 обращений	нет	binary
6	Вохмин Артур Евгеньевич	categoriya	80000	Лемматизация	нет	обеспечить пребывание всех классов в тестовой выборке	count
7	Горбачев Богдан Альбертович	sphera	90000	Лемматизация	нет	обеспечить пребывание всех классов в тестовой выборке	tfidf
8	Дойков Андрей Кириллович	categoriya	100000	Лемматизация	нет	обеспечить сохранение распределения по классам в тренировочной и тестовой выборках	freq
9	Железняк Николай Игоревич	sphera	110000	Лемматизация	нет	Обеспечить перемешивание данных при формировании выборок	binary
10	Кафтаранов Тимур Дамирович	categoriya	20000	Лемматизация	нет	обеспечить пребывание всех классов в тестовой выборке	count
11	Кобзев Никита Алексеевич	sphera	30000	Лемматизация	нет	Обеспечить перемешивание данных при формировании выборок	tfidf
12	Малышев Андрей Максимович	categoriya	60000	Лемматизация	исключить классы, в которых меньше 200 обращений	нет	freq
13	Милин Владислав Сергеевич	sphera	10000	Лемматизация	исключить классы, в которых меньше 300 обращений	нет	binary
14	Митрошина Екатерина Ильинична	categoriya	1000000	Стемминг	нет	обеспечить пребывание всех классов в тестовой выборке	count
15	Платонов Арсений Александрович	sphera	30000	Стемминг	нет	обеспечить сохранение распределения по классам в тренировочной и тестовой выборках	tfidf
16	Сурдова Елизавета Юрьевна	categoriya	1000	Стемминг	исключить классы, в которых меньше 500 обращений	нет	freq
17	Ткачев Илья Витальевич	categoriya	80000	Стемминг	нет	Обеспечить перемешивание данных при формировании выборок	binary

18	Токмашов Александр Евгеньевич	sphera	90000	Стемминг	нет	обеспечить пребывание всех классов в тестовой выборке	count
19	Ушаков Игнат Михайлович	categoriya	100000	Стемминг	нет	обеспечить сохранение распределения по классам в тренировочной и тестовой выборках	tfidf
20	Фитц Артемий Валерьевич	categoriya	110000	Стемминг	нет	Обеспечить перемешивание данных при формировании выборок	freq
21	Хомякова Ксения Сергеевна	sphera	20000	Стемминг	нет	обеспечить пребывание всех классов в тестовой выборке	binary
22	Цыпленков Константин Алексеевич	categoriya	30000	Стемминг	нет	Обеспечить перемешивание данных при формировании выборок	count
23	Шапошник Даниил Сергеевич	sphera	60000	Стемминг	исключить классы, в которых меньше 200 обращений	нет	tfidf
24	Швидерская Анна Алексеевна	categoriya	10000	Стемминг	исключить классы, в которых меньше 300 обращений	нет	freq
25	Шомахов Альберт Арсенович	sphera	1000000	Стемминг	исключить классы, в которых меньше 500 обращений	нет	binary