Инструкция к сдаче

1. Настоятельно рекомендуем сдавать практическое задание в виде ссылки на личный репозиторий на github.

2. Рекомендуемый способ организации данных в репозитории: создать отдельные папки по темам и помещать в них отдельные файлы для каждой задачи с правильным расширением.

Ссылка на инструкцию по работе с git и сдачу практики:

https://docs.google.com/document/d/1RAT_ukE39iOfbz1xa39QXae2hBUEZ4U6Fko_wFDdrsM/edit

Ссылка на видеокурс по Git:

https://geekbrains.ru/courses/66

Если остались сложности с системой git, то обратитесь к преподавателю или наставнику.

Тема «Создание признакового пространства»

Продолжим обработку данных с Твиттера.

1. Создайте мешок слов с помощью sklearn.feature_extraction.text.CountVectorizer.fit_transform(). Применим его к 'tweet_stemmed' и 'tweet_lemmatized' отдельно.
   - Игнорируем слова, частота которых в документе строго превышает порог 0.9 с помощью max_df.
   - Ограничим количество слов, попадающий в мешок, с помощью max_features = 1000.
   - Исключим стоп-слова с помощью stop_words='english'.
   - Отобразим Bag-of-Words модель как DataFrame. columns необходимо извлечь с помощью CountVectorizer.get_feature_names().

2. Создайте мешок слов с помощью sklearn.feature_extraction.text.TfidfVectorizer.fit_transform(). Применим его к 'tweet_stemmed' и 'tweet_lemmatized' отдельно.
   - Игнорируем слова, частота которых в документе строго превышает порог 0.9 с помощью max_df.
   - Ограничим количество слов, попадающий в мешок, с помощью max_features = 1000.
   - Исключим стоп-слова с помощью stop_words='english'.
   - Отобразим Bag-of-Words модель как DataFrame. columns необходимо извлечь с помощью TfidfVectorizer.get_feature_names().

3. Проверьте ваши векторайзеры на корпусе который использовали на вебинаре, составьте таблицу метод векторизации и скор который вы получили (в методах векторизации по изменяйте параметры что бы добиться лучшего скора) обратите внимание как падает/растёт скор при уменьшении количества фичей, и изменении параметров, так же

попробуйте применить к векторайзерам PCA для сокращения размерности посмотрите на качество сделайте выводы