Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 9**

**по дисциплине «Методы машинного обучения»**

Классификация текста

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Задание**

Для произвольного набора данных, предназначенного для классификации текстов, необходимо решить задачу классификации текста двумя способами:

* Способ 1. На основе CountVectorizer или TfidfVectorizer.
* Способ 2. На основе моделей word2vec или Glove или fastText.
* Сравните качество полученных моделей.

# **Описание данных**

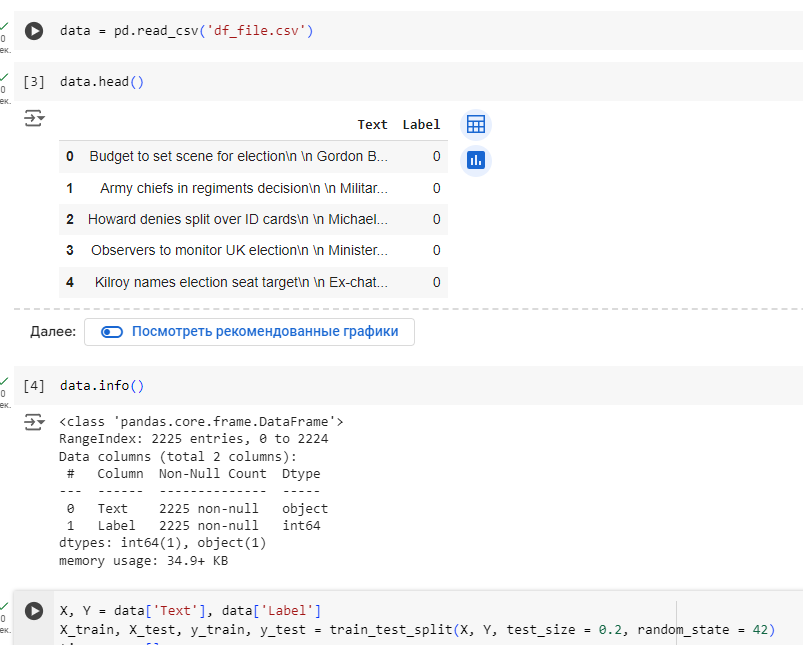
Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения модой.

Будет использован набор данных для классификации текстовых документов, который содержит 2225 текстовых данных и пять категорий документов.

Данные содержат 2 столбца:

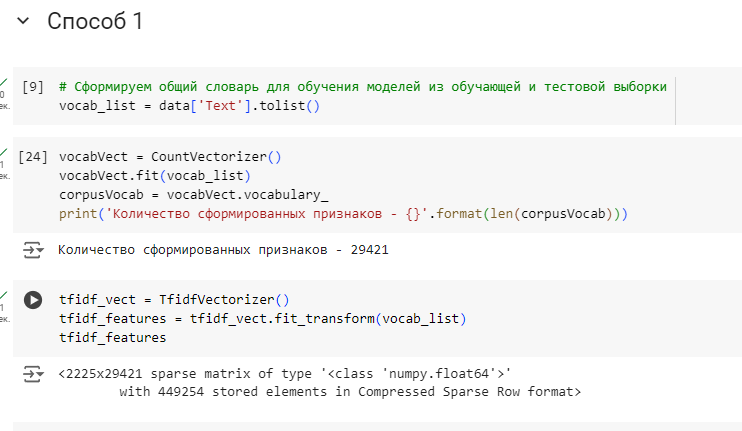
* Текст: Содержит различные категории текстовых данных
* Метка: Содержит метки для пяти различных категорий (0,1,2,3,4)
  1. Политика = 0
  2. Спорт = 1
  3. Технологии = 2
  4. Развлечения = 3
  5. Бизнес = 4

Добавим данные и разделим данные на обучающую и тестовую выборки.



# **Способ 1**

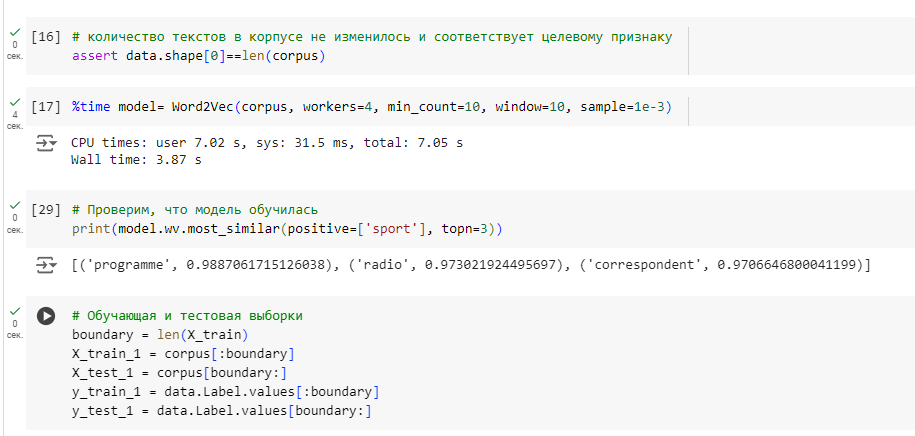
Как первый метод используем TF-IDF, числовой статистический показатель, который отражает важность слова для документа.



1. **Способ 2**

Как второй способ используем Word2Vec, способ векторизации слов текста, использующий нейронные сети.

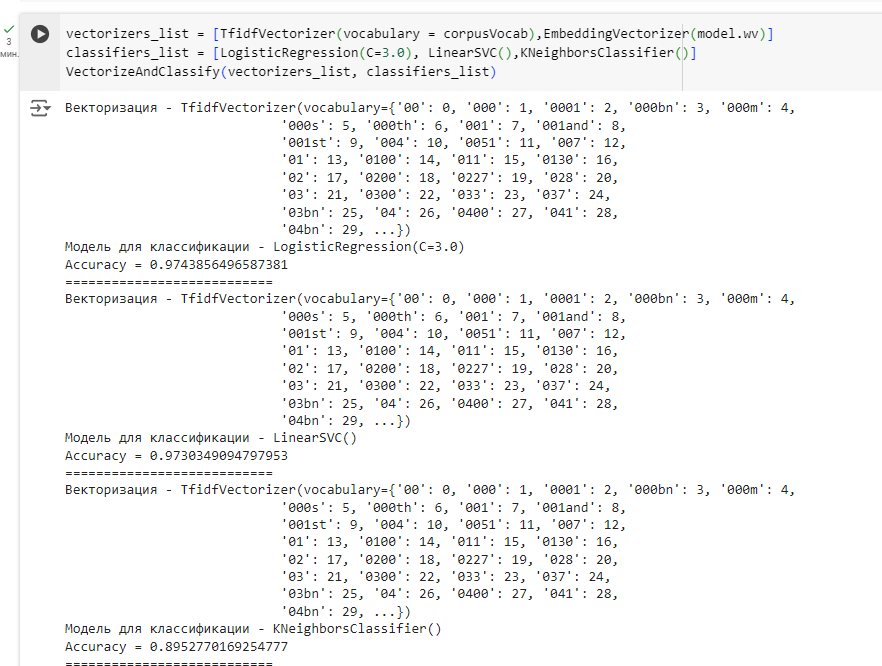




1. **Сравнение качества полученных моделей**

Для числовых признаков были построены гистограммы для обзора их распределений.

Сравним методы векторизации вместе с набором классификаторов: LogisticRegression, LinearSVC и KNeighborsClassifier.



Сведем результаты в таблицу 1.

Таблица 1. Сравнение классификаторов

|  |  |  |
| --- | --- | --- |
| **Векторизация** | **Классификатор** | **Точность** |
| Tfidf | LogisticRegression | 97,44% |
| Tfidf | LinearSVC | 97,30% |
| Tfidf | KNeighborsClassifier | 89,53% |
| Word2Vec | LogisticRegression | 34,47% |
| Word2Vec | LinearSVC | 37,08% |
| Word2Vec | KNeighborsClassifier | 33,93% |

Из таблицы 1 можно отметить, что метод векторизации Tfidf со всеми классификаторами показал качество значительно выше (почти в 3 раза), чем Word2Vec. Наилучшее качество показала комбинация «Tfidf+LogisticRegression», а худшее – «Word2Vec+KNeighborsClassifier».

Для наилучшей комбинации получим оценки качества для классов.



Таким образом, были изучены 2 метода векторизации: Tfidf и Word2Vec. Их использование было сравнено для 3 классификаторов. Результаты лучше у метода Tfidf, а наилучший результат – комбинация «Tfidf+LogisticRegression».