AI NEWS

Команда 62_DS

1. Описание задач

Команды получили неразмеченный датасет с заголовками, текстами и датой выпуска новостей. Были поставлены 2 задачи:

1. Удалить похожие по смыслу новости
2. Классифицировать (без учителя) новости по заданным категориям.

2. Описание алгоритма решения
2.1. Задача удаления дубликатов

Основные этапы по решению:

- чистим тексты (удаляем символы, другие языки и цифры)
- лемматизируем тексты (приводим все слова к инфинитивной форме, инструмент - Mystem)
- векторизуем тексты при помощи tf-idf (подсчитываем частоту встречи слова в конкретном тексте и во всех текстах и создаём вектора по всем текстам, инструмент - sklearn - TfidfVectorizer)
- сравниваем вектора текстов между собой (методом косинусного расстояния: чем ближе вектора друг к другу численно, тем ближе по смыслу, а точнее по одинаковости слов. Инструмент - sklearn - cosine_similarity (0 - тексты не похожи, 1 - тексты одинаковы).
- отбираем похожие тексты и отбрасываем.

2. Описание алгоритма решения
2.2. Задача классификации текста новости

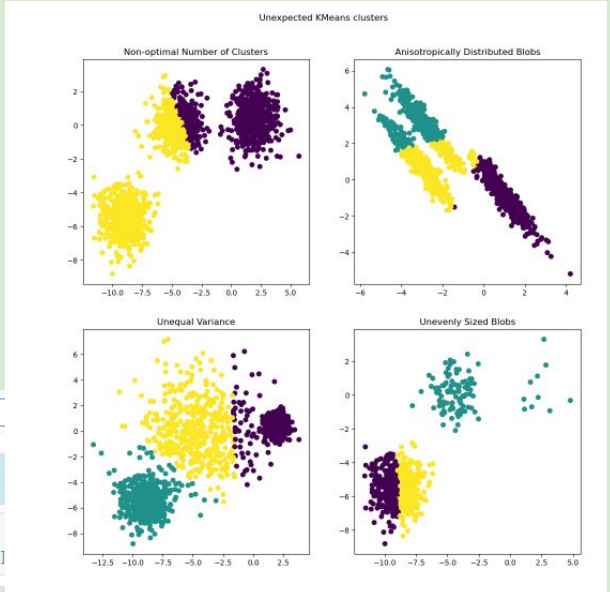
Основные этапы по решению:

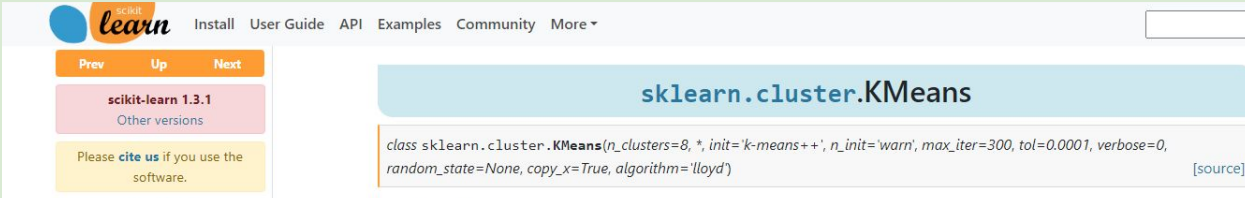
- категории опишем словами-тегами (Шоубиз - это шоубизнес звезда актёр актриса театр скандал и тд)
- векторизуем тексты и слова-теги единым векторизатором (также TF-IDF)
- выделяем вектора текстов и вектора слов - тегов
- вектора слов-тегов используем как центроиды для метода k-средних для обучения без учителя (по сути измеряются расстояния между векторами текстов и векторами центроидов и таким образом производится кластеризация)
- размечаем тексты по категориям (инструмент sklearn KMeans)

3. Техника обучения

Алгоритм очистки дубликатов не использует моделей машинного обучения.

Алгоритм классификации использует модель обучения без учителя методом k-средних при помощи инструмента sklearn KMeans





4. Основные инструменты

```
from pymystem3 import Mystem #импорт класса лемматизатора pymystem3
import re #импорт модуля для работы с регулярными выражениями (для очистки)
from nltk.corpus import stopwords as nltk_stopwords #импорт инструмента для определения стоп-слов
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer #импорт инструмента для расчёта TF-IDF
from sklearn.metrics.pairwise import cosine_similarity #измерение косинусной близости векторов
from sklearn.cluster import KMeans #импорт модели обучения без учителя методомм k-средних
```

5. Результаты на тесте

По результатам проверки Организаторами по критериям точности по первой и второй задачам, а также по времени работы алгоритма, наша Команда заняла 4 место из 40 сданных работ и 100+ начальных участников. Ниже представлена выдержка из лидерборда.

Команда	Скор по дубликатам	Скор по категориям	Итоговая оценка
FAI	0,9285714286	0,5593869732	0,5483308719
Wer-30035	0,8265306122	0,6325136612	0,5027676291
M&L	0,8367346939	0,3941176471	0,4760411168
62_DS	0,8297666939	0,4067817638	0,451384033