2024-11-05

Практика 4. Сбор видеоданных

Написать функцию, которая принимает на вход список URL, выгружает по заданным URL весь видеоконтент, найденный по этому адресу, вместе с метаданными (в `.txt`) и сохраняет его в отдельную папку, названную по URL.

Какая сложность может быть:

1. Адрес прямого видеофайла может быть скрыт — для этого можно воспользоваться специализированными средствами, например, `youtube-dl` GitHub - ytdl-org/youtube-dl: Command-line program to download videos from YouTube.com and other video sites youtube_dl · PyPI — вообще, он поддерживает много сайтов, см. youtube-dl: Supported sites
2. Видео может быть многочастным (в формате `.m3u8`-плейлиста) — нужно разобрать плейлист и скачать каждую из частей видео, а потом склеить
3. Метаданные к видео могут быть разбросаны по странице — в этом случае стоит предполагать, что адрес содержит только 1 видео и пытаться извлекать их из `title`, `description` и т.п.

Какие сайты можно пробовать парсить:

- [Imgur: The magic of the Internet] [rBxB9ih.mp4](например How to wake up at 5am every day - GIF - Imgur -> https://i.imgur.com/rBxB9ih.mp4)
- Reddit (например, https://packaged-media.redd.it/9kr7fm5k16eb1/pb/m2-res_480p.mp4?m=DASHPlaylist.mpd&v=1&e=1690448400&s=a56139280f0853d7bf356fe10e02cf0da081ba56#t=0)
- RuTube
- GIPHY - Be Animated
- https://tenor.com/
- TikTok*
- YouTube*
- Twitch*

- Vimeo*

Лабораторная работа №3. Кластерный анализ веб-данных

1. Написать программу для автоматизированного сбора списка блюд, подаваемых в городе. Это сделать можно путем анализа нескольких веб-сайтов по запросу, например, блюда казани — Яндекс: нашлось 194 тыс. результатов
2. Провести при помощи сбора веб-данных анализ частотности упоминаний каждого блюда в городе на отобранных сайтов (частотность нормализуем, т.е. приводим к виду $[0, 1]$, 0.0 — блюдо не упоминается в контексте города вообще, 1.0 — каждый сайт про кухню города упоминает это блюдо).
3. Сделать это для нескольких городов Пермского края и городов из разных регионов (например, Москва, Санкт-Петербург, Казань, Уфа, Новосибирск, Хабаровск, Петрозаводск и Карелия)
4. Сформировать общий список блюд для всего множества городов
5. Закодировать каждый город: построить вектор, где каждое измерение — блюдо из всего списка блюд, а значение этого измерения — его нормализованная частотность упоминаний в контексте города.
6. Полученные данные кластеризовать методом k-средних, иерархической кластеризацией и спектральной кластеризацией, взяв готовую реализацию из `scikit-learn` 2.3. Clustering — scikit-learn 1.3.2 documentation
7. Визуализировать кластеры городов (как себя проверить: на выходе города, где любят блюда одной кухни, должны оказаться рядом) при помощи `matplotlib` и `seaborn` или иных библиотек