HabrCompaniesParser

Цель:

Продемонстрировать ***NLP (Natural Language Processing)*** - обработку текстов на естественном языке при помощи ***NLTK (Natural Language Toolkit)*** - пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python.

Задачи

Целевой сайт: habr.com

Целевая страница: companies

Дополнительная целевая страница company profile (у каждой компании *собственный* профиль)

1. Веб-скрапинг целевого сайта, а именно целевых страниц.
2. Собрать DataFrame из полученных данных в 1-ом пункте.
3. Провести предобработку данных.
4. NLTK:
    - Стоп-слова
    - Регулярные выражения
    - Стемминг и лемматизация
    - Токенизация:
        - По словам
        - По предложениям
    - Мешок слов
    - TF-IDF

Ход работы

Веб-скрапинг

Используемые технологии: Beautiful Soup, Selenium

Парсинг данных

Некоторые данные на сайте генерируются динамически при помощи **JavaScript**. В таком случае получить эти данные силами **Beautiful Soup** не удастся.

Beautiful Soup исследует исходный код страницы, не исполняя **JavaScript**. Для исполнения **JavaScript** кода и получения страницы, идентичной той, которую видно в браузере, можно использовать **Selenium WebDriver** - это набор драйверов для различных браузеров, снабжающийся библиотеками для работы с этими драйверами.

Основным инструментом будет являться **Selenium**.

В данном случае будут использоваться следующие **локаторы Selenium**: **By.XPATH**, **By.CLASS_NAME**.

Все используемые локаторы

Название поля	**Используемый селектор**	**Технология**
Блок компаний	"tm-companies"	Beautiful Soup
Список всех компаний	"tm-companies__item tm-companies__item_inlined"	Beautiful Soup
Блок краткой информации о кампании	f"//body/div[@id='app']/div[1]/div[2]/main[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[2]/div[{companyID}]/div[1]/div[1]/div[1]"	Selenium
Название компании	"tm-company-snippet__title"	Selenium
Описание компании	"tm-company-snippet__description"	Selenium
Ссылка на профиль компании	"tm-company-snippet__title"	Selenium

Название поля	Используемый селектор	Технология
Блок на цифровые показатели компании	f"//body/div[@id='app']/div[1]/div[2]/main[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[2]/div[{companyID}]/div[2]"	Selenium
Рейтинг компании	f"/html[1]/body[1]/div[1]/div[1]/div[2]/main[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[2]/div[{companyID}]/div[2]/span[1]"	Selenium
Количество подписчиков компании	f"/html[1]/body[1]/div[1]/div[1]/div[2]/main[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[2]/div[{companyID}]/div[2]/span[2]"	Selenium
Блок хабов компании	f"//body/div[@id='app']/div[1]/div[2]/main[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[2]/div[{companyID}]/div[1]/div[2]"	Selenium
Хабы коспании	"tm-companies__hubs-item"	Selenium
Блок отраслей компании	"tm-company-profile__categories"	Selenium
Отрасли компании	"tm-company-profile__categories-wrapper"	Selenium
О компании	"//body/div[@id='app']/div[1]/div[2]/main[1]/div[1]/div[1]/div[2]/div[1]/div[1]/div[2]/section[1]/div[1]/div[1]/dl[2]/dd[1]/span[1]" "//body/div[@id='app']/div[1]/div[2]/main[1]/div[1]/div[1]/div[2]/div[1]/div[1]/div[2]/section[1]/div[1]/div[1]/dl[3]/dd[1]/span[1]"	Selenium

Хранение данных

О каждой компании будет собран **Json**:

```
{
    "Name": "companyName",
    "Description": "companyDescription",
    "About": "companyAbout",
    "Industries": "industries",
    "Rating": "companyRating",
    "Subscribers": "companySubscribers",
    "Hubs": "companyHubs",
    "Profile": "companyProfile"
}
```

Атрибут	Тип данных
Name	***str***
Description	***str***
About	***str, Multi-line***
Industries	***List[str]***
Rating	***float***
Subscribers	***int***
Hubs	***List[str]***
Profile	***str***

После завершение работы кода, будет собран файл `summary of companies.json`:

```
{
  "companyName": {
    "Description": "...",
    "About": "...",
    "Industries": [
      "...",
      "...",
       ...
    ],
    "Rating": ...,
```

```
      "Subscribers": ...,
      "Hubs": [
        "...",
        "...",
         ...
      ],
      "Profile": "..."
    },
    ...
  }
```