**> Детали реализации и описание**

В рамках домашней работы необходимо собрать информацию касательно актёров и фильмов, в которых они снимались (можете представить, что пытаетесь создать граф связей актёров со всего мира, выяснить, кто с кем работает чаще и т.д.). Как было сказано на практике, к сожалению, не представляется возможным и практичным со всех точек зрения заниматься парсингом E-commerce сайтов в рамках обучающих курсов.

В качестве решения необходимо предоставить два файла в формате jsonlines (.jl, каждая строка — один json-валидный объект, т.е. словарь с ключами и значениями). Первый файл описывает топ-50 самых популярных актёров на данный момент, второй — фильмы, в которых они снимались. Более формально, необходимо обработать информацию со страницы [по этой ссылке](https://www.imdb.com/search/name/?gender=male%2Cfemale&ref_=nv_cel_m), переходя на персональные страницы, и для каждого актёра указать:

1. Биографию (ключ bio в json). Не нужно рекурсивно переходить по ссылке See full bio », достаточно ограничиться текстом на странице актёра (при этом фразу See full bio » необходимо вырезать из текста, оставив многоточие в конце, см. примеры на сайте). Биография должна содержать в том числе и текст ссылок (так как часто описываются картины, в которых человек принял участие).
2. Дату рождения актёра (ключ born в json). Может быть null, если дата не указана, иначе просто строка. Дата в формате, указанном на сайте (т.е. в формате yyyy-m-d), не требует постобработки.
3. Список фильмов (ключ movies в json) — первые 15 штук (сверху вниз, т.е. от последних к первым в порядке, указанном на сайте. Отдельно сортировать по дате самостоятельно не нужно). Список представляет собой List[str].
4. Имя актера (ключ name). Требуется только "базовое имя", указанное на основной странице либо выделенное крупными буквами на странице актёра (иногда встречаются псевдонимы, и если они выделены — их и нужно указать).
5. Ссылка на персональную страницу (ключ url). https://www.imdb.com/ в начале и закрывающий / в конце обязательны для прохождения проверки.

Как было указано выше, необходимо для первых 15 фильмов в таблице произвести парсинг актёров, принявших участие (можно представить. что это описание рёбер графа):

1. Ссылка на страницу фильма (ключ url). https://www.imdb.com/ в начале и закрывающий / в конце обязательны для прохождения проверки.
2. Название фильма (ключ title).
3. Список актёров (ключ cast) — List[str] с перечислением имён (чтобы они вязались с п.4 предыдущего списка).

По сути требуется начать с корневой ссылки, перейти на страницы актёров, затем для каждого проследовать по ссылкам на 15 фильмов. Для всех текстовых полей не забудьте применить .strip() (в том числе после постобработки, если она производилась).

Есть два варианта получения данных из парсинга. Первый — полноценный, с запуском парсера из консоли, написанием файла настроек и т.д. Он детально описан в инструкции [по ссылке](https://docs.scrapy.org/en/latest/intro/tutorial.html). Он подойдёт для процессов в продакшене и при более сложных пайплайнах (к примеру, при записи в БД и обработке картинок). Второй вариант, предлагаемый в целях простоты, — сохранение прямо из ноутбука. Для этого нужно передать, к примеру, следующие настройки при запуске процесса краулинга:

process = CrawlerProcess(

settings={

"FEEDS": {

"items.json": {"format": "json"},

}

}

)

process.crawl(ImdbSpider)

process.start()

Тогда все объекты будут сохранены в файл items.json, после чего оттуда их можно вручную считать и разбить на два файла для загрузки в систему проверки. О наличии дубликатов в фильмах не беспокойтесь.