# Парсер HTML страниц на Python

1. Написать парсер для извлечения нужной информации из одной произвольной статьи на **e1**. Оформить решение с использованием своих функций.  
   Перечень того, что нужно извлечь:

Таблица 1 Список простейших характеристик

|  |  |  |
| --- | --- | --- |
| **№** | **Характеристики** | **Обозначение** |
| 1 | Количество слов в заголовке | words-in-title |
| 2 | Количество слов в статье | words-in-article |
| 3 | Количество слов в теге description | words-in-description |
| 4 | Количество изображений в статье | images-number |
| 5 | Количество ссылок на другие новости e1.ru | links- number |
| 6 | Количество просмотров | views-number |
| 7 | Количество комментариев | comments- number |

1. В разделе **Новости** <http://www.e1.ru/news/> выберите один интересный вам раздел:

Таблица 2 Категории в разделе Новости

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | Автоновости | 14 | Дороги | 26 | Осень |
| 2 | Бизнес | 15 | Дорожное видео E1.RU | 27 | Мнение |
| 3 | В верхах | 16 | Здоровье | 28 | Недвижимость |
| 4 | Вкусно today | 17 | Интервью | 29 | Неизвестный Свердловск |
| 5 | Город | 18 | Конкурс | 30 | Новости сообщества E1.RU |
| 6 | Дом | 19 | Культура | 31 | Обзор |
| 7 | Образование и работа | 20 | Общество | 32 | Отдых |
| 8 | Подробности | 21 | Покупки | 33 | Происшествия |
| 9 | Расследование | 22 | Репортаж | 34 | Спорт |
| 10 | Стиль | 23 | Страна и мир | 35 | Суды и криминал |
| 11 | Тесты | 24 | Технологии и интернет | 36 | Экономика |
| 12 | Эксклюзив | 25 | Эксперимент | 37 | Фоторепортаж |
| 13 | Лонгриды |  |  |  |  |

1. Извлеките из выбранного раздела статьи за полгода. В зависимости от тематики их может быть больше или меньше (это нормально).
2. Вычислите для каждой статьи указанные выше характеристики.
3. Добавьте к каждой статье характеристику №8 — метку категории. Пусть это будет номер категории в таблице выше.
4. Постройте гистограмму распределения **views-number** и **comments- number** (две гистограммы должно быть). Гистограмма распределения показывает, насколько часто встречаются те или иные значения X (аналог плотности распределения). Каждый столбец гистограммы показывает частоту попадания значения выборки в интервал значений - чем выше столбец, тем вероятнее соответствующие значения показателя.

**Важно.** В будущих занятиях планируется объединить результаты всех вычислений в один набор данных. Поэтому нужно договориться о том, как мы будем хранить все результаты вычислений. Понятно, что нужно это делать единообразно.

Например, можно сделать список словарей. Для каждой статьи значения характеристик записывать в словарь. А словари уже помещать в список. В результате размер списка должен равняться количеству статей в выбранной категории.

**О сроках.** На выполнение это работы отводится 2 недели. Обратите внимание, что к 20 октября у вас должны быть получены все результаты. Чтобы можно было на занятии объединить все расчеты и двигаться дальше.