Лика Капустина и Кирилл Крипайтис.
Проект по курсу "Основы программирования в Python".

Документация: списки и словари

Название списка или словаря	Описание
list_of_pages	Список со ссылками на все 16 страниц с содержанием сайта. Создается с помощью генератора списков (элементы от 2 до 16 включительно); далее на первое место вставляется ссылка на первую страницу (у нее другой синтаксис).
stopwords	Список со словами, которые не учитываются при подсчете самых часто упоминаемых слов в материале подкаста. Создан вручную. Содержит частицы, местоимения, предлоги и так далее.
standard_links	Список со ссылками, которые могут присутствовать на каждой странице. Включает в себя ссылку на главную страницу сайта, социальные сети проекта; расширяется ссылками из list_of_pages (тк они в том или ином количестве тоже присутствуют на каждой странице с ссылками).
author_list, category_list, podcast_list_new	Списки авторов, категорий и упоминаемых в материалах сайта подкастах. Значения внутри списка не уникальны.
author_unique_list, category_unique_list, podcast_unique_list	Списки с уникальными значениями, созданный из предыдущих 3 списков. Нужны для подсчета количества опубликованных авторами материалов, количества опубликованных в категории материалов и количества упоминаний подкастов в материалах сайта.
author_plot, category_plot, podcast_plot	Словари, нужные для построения графиков, содержащие координату x и label, то есть определенного автора, категорию или упоминаемый подкаст.
podcast_list_of_lists	Список, содержащий списки упоминаемых в подкасте материалов (список списков). То есть под индексом i можно найти список подкастов, которые упоминались в i-той статье.

Документация: программы

Название программы	Принцип работы программы
inner_podcast	Программа получает на вход html-текст (soup). Внутри программы создается пустой словарь, куда парами добавляются ключ (название подкаста) и значение (ссылка, i.get('href')). Проверяется, что получаемая ссылка не равна None, что podcast.ru/ есть в ссылке (именно podcast.ru. а не podcasts.ru, это разные домены и на первом хранятся как бы taplink на подкаст, это страница с которой можно перейти на все ссылки на подкаст (яндекс музыка, спотифай, и так далее).

	Возвращается словарь с упомянутыми подкастами и ссылками на них
cleaning	Программа получает на вход строку. Создает новую. Если элемент строки не левая елочка, не правая елочка, и не находится в string_digits (вынесены в качестве отдельной строки), то к новой строке прибавляется этот элемент старой. Возвращает новую строку. По факту, очищает строку от знаков препинания и елочек. Нужна для дальнейшего анализа самых часто упоминавшихся слов.
take_info_from_page	Программа получает на вход ссылку. Переходит на нее и преобразует html.text в soup. Далее в изначально заданные списки (links_to_materials, dates, categories) присваивает соответствующие значения, полученные из html-текста. Они были предварительно найдены на одной странице и перенесены сюда. В программе так же применяется функция cleaning к каждой строке текста внутри, считается длинна материала, считаются самые упоминаемые слова, из них находится 5 самых упоминаемых и они возвращаются строкой с разделителем в виде запятой с пробелом. Используется программа inner_podcast, и возвращает словарь inner_links, в котором находятся все упомянутые в материале подкасты, а уже это значение добавляется в изначально определенный словарь.
take_info_from_lenta_page	Программа получает на вход одну из 16 страниц с материалами podcasts.ru, получает с нее html.text, и находит там все ссылки (те ссылки, которых нет в standard_links + еще несколько условий). Этот список сохраняется в links_on_the_page, а уже по этому списку пускается цикл for, который применяет функцию take_info_from_page к каждой странице на странице. И засыпает на 1 секунду.
main	Главная функция, объединяющая в себе все остальные. По сути это цикл: она обращается к кааждой из 16 ссылок на страницы с материалами podcasts.ru, и применяет функцию take_info_from_lenta_page, котрая в свою очередь применяет функцию take_info_from_page, и так далее.
authors_list, category_list, podcast_list	Программа, которая обращается к pandas-датафрейму и возвращает из него конкретную колонку: список авторов, список категорий или список словарей по упоминанию подкастов в материалах
authors_unique, category_unique podcast_unique	Программа, которая возвращает уникальные значения в колонке. В случае podcast_unique механизм устроен сложнее и, хотя сама функция тоже просто возвращает уникальные значения, но до этого следует podcast_unique_create, которая создает список уникальных значений с помощью сет.
search_podcast	Функция, которая выполняет поиск материалов сайта, в которых упоминается заданный подкаст. Принцип работы такой: пользователь вводит название подкаста,

	программа обращается к списку списков podcast_list_of_lists и печатает название подкаста, материал, в котором он упоминается и ссылку на этот материал, повторяя это несколько раз.