Пакет «Робот» реализован на языке *Python 2.7* и состоит из двух программ:

1. Webdizer.

Программа обладает функционалом просто поискового робота. Входные данные задаются в файле *«settings.ini»*.

settings.ini:

user\_agent::Mozilla/5.0 (iPhone; CPU iPhone OS 7\_0 like Mac OS X) AppleWebKit/546.10 (KHTML, like Gecko) Version/6.0 Mobile/7E18WD Safari/8536.25

url\_list::url\_list.txt

depth::1

save\_method::page

save\_to::C:\

Параметры

*User\_agent* — задает user agent, который будет помещаться в запросы на загрузку веб-страниц

*Url\_list* — указывает на файл, содержащий ссылки для обхода. Требование к файлу — каждая ссылка должна находиться на новой строчке.

*Depth* — задает глубину обхода ссылок. Глубина задается цифрами.

*Save\_method* — задает один из двух вариантов сохранения ссылок: в виде списка (значение url) и в виде html-страниц (значение page).

*Save\_to* — указывает директорию для записи файлов страниц/списков ссылок.

Выходные данные — набор директорий, содержащих в себе html-страницы или текстовые файлы со ссылками (в зависимости от входных данных). Название директории — дата начала работы программы, доменное имя сайта.

Название файла страницы — ссылка на страницу, обработанная хэш-функцией MD5. При сохранении страниц создается текстовый файл *«url+md5.txt»*, содержащий в себе названия всех файлов, загруженных в эту директорию и ссылки, с которых они были скачаны.

После окончания обхода ссылок составляется файл базовых метрик.

Название файла — «*дата начала работы программы\_metrics.txt*»

=============BASIC\_\_CRAWLER\_\_METRICS=============

Start time: 2015-05-06 16:48:52.521000

Finish time: 2015-05-06 16:49:14.439000

Crawled time: 0:00:21.918000

Number of collected links: 3708

Of them unique, related to inputed sites: 22

Longest link(89 symbols): http://economics.lb.ua/state/2015/05/05/303925\_ukraina\_ostanovila\_revers\_gaza\_cherez.html

Shortest link(20 symbols): economics.lb.ua/news

Amount of downloaded data: 1438935 bytes

=====Link, times it is quoted=====

(u'http://economics.lb.ua/business/2015/05/05/303909\_uzbekskiy\_aviaperevozchik.html', 1)

(u'http://economics.lb.ua/finances/2015/04/22/302564\_mifi\_ekonomii\_kommunalke.html', 1)

(u'http://economics.lb.ua/state', 1)

(u'http://economics.lb.ua/state/2015/05/05/303998\_naftogaz\_perechislil\_gazpromu.html', 1)

(u'http://economics.lb.ua/business/2015/05/05/303996\_aktsioneri\_priznali\_effektivnoy.html', 1)

(u'economics.lb.ua/state', 1)

(u'http://economics.lb.ua/state/2015/05/05/303951\_gazprom\_trebuet\_ukraini.html', 1)

(u'http://economics.lb.ua/state/2015/05/06/303999\_shimkiv\_potribni\_i\_liderstvo\_i.html', 1)

(u'http://economics.lb.ua/state/2015/05/05/303925\_ukraina\_ostanovila\_revers\_gaza\_cherez.html', 1)

(u'http://economics.lb.ua/finances/2015/05/05/303971\_kabmin\_dal\_nasirovu\_tri\_mesyatsa.html', 1)

(u'http://economics.lb.ua/state/2015/05/05/303929\_glavoy\_fiskalnoy\_sluzhbi\_stal.html', 1)

(u'http://economics.lb.ua', 2)

(u'http://economics.lb.ua/business', 1)

(u'http://economics.lb.ua/state/2013/11/20/242069\_arenda\_zhilya\_kieve\_varshave.html', 1)

(u'http://economics.lb.ua/finances', 1)

(u'http://economics.lb.ua/business/2013/08/08/217888\_dorogo\_li\_zhit\_irlandii.html', 1)

(u'economics.lb.ua/finances', 1)

(u'http://economics.lb.ua/business/2015/05/05/304005\_aeroport\_borispol\_nazval\_samie.html', 1)

(u'http://economics.lb.ua/state/2014/05/08/265112\_dorogo\_li\_zhit\_latvii.html', 1)

(u'economics.lb.ua/business', 1)

Значения метрик.

Группа времени *(Start time, Finish time, Crawled time)*, данные о загруженной информации *(Amount of downloaded data)*, группа ссылок *(«Number of collected links», «Of them unique, related to inputed sites»)* позволяют оценить объем обработанных данных и производительность скрипта.

Метрики *(«Longest link», «Shortest link»)* представляют интерес при обзоре большой группы сайтов.

Группа *«Link, times it is quoted»* содержит в себе список пройденных ссылок и количество упоминаний ссылки в скаченных документах.

Варианты поставки :

* в виде скриптa на *Python 2.7* (для работы необходима библиотека *BeautifulSoup 3(4))*;
* в виде набора библиотек и файла robot.exe с иконкой.

1. Analyzer

Программа обрабатывает коллекцию скаченных документов. При работе используется файл параметров «*directory.ini»*.

directory.ini:

read\_from::C:\Users\KR2142\Desktop\PyScripts\

read\_from — параметр указывает расположение папок файлов для анализа. Следует указывать абсолютный путь.

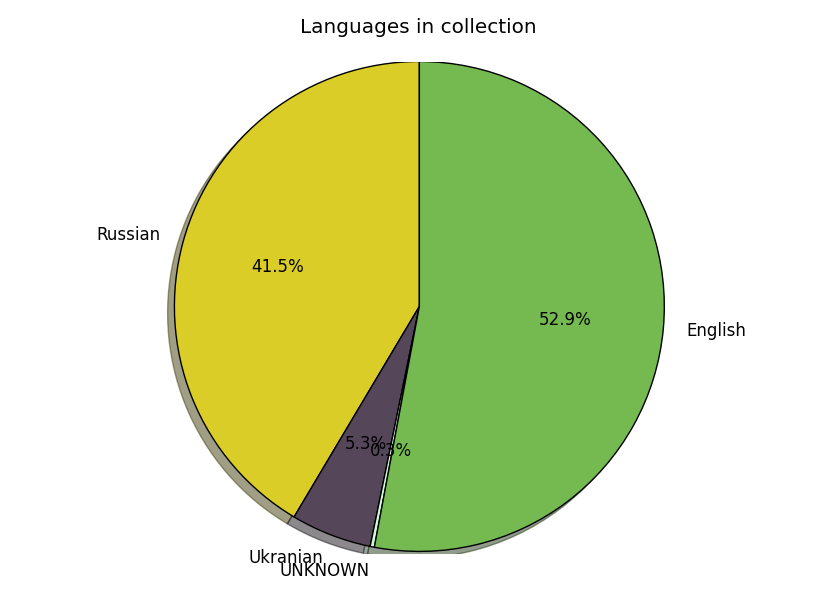


Рисунок 1. Пример диаграммы, построенной программой Analyzer

В ходе работы программа строит круговую диаграмму языков, из которых составлены страницы коллекции, и создает файл отчета *«текущая дата-hard\_metrics.txt»*.

C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-16.06.03\_economics.lb.ua\13e10c5a4360dcd4bcdd7c02598c9d57.html

Language: Russian

Number of links: 179

Size of file: 69921

Amount of words: 1455

==============================================================

…

==============================================================

C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\fe02ac123d4bd9d02b0c55768f0082e8.html

Language: English

Number of links: 48

Size of file: 106268

Amount of words: 68

==============================================================

Start time: 2015-05-06 22:08:16.397000

Finish time: 2015-05-06 22:09:01.743000

Time spent on analysis: 0:00:45.346000

File with highest # of links(207): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\a4eb3a98a4d494c32553006c0b18e21e.html

File with lowest # of links(0): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\c9058d383cb1fde01bf5b4320072d3a7.html

File with highest # of words(4571): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\1ecdc4ec5c01cf589ff207a2b3cb3301.html

File with lowest # of words(2): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\c9058d383cb1fde01bf5b4320072d3a7.html

Biggest file in the collection(412950): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\a4eb3a98a4d494c32553006c0b18e21e.html

Smallest file in the collection(1059): C:\Users\KR2142\Desktop\PyScripts\robot2\06.05.2015-21.29.14\_9gag.com\c9058d383cb1fde01bf5b4320072d3a7.html

Languages in the collection with # of files: Counter({'English': 189, 'Russian': 148, 'Ukranian': 19, 'UNKNOWN': 1})

Sizes (bytes) in the collection with # of files: Counter({61367L: 8, 77370L: 4, 77338L: 3, 70171L: 3, 70581L: 3, 80737L: 3, 80234L: 3, 92654L: 3, 89543L: 3, 77293L: 3, 106502L: 2, 74250L: 2, 101909L: 2, 70167L: 2, 49674L: 2, 49736L: 2, 74826L: 2, 84568L: 2, 80003L: 2, 77448L: 2, 98449L: 2, 108709L: 2, 57002L: 2, 106167L: 2, 92380L: 2, 92384L: 2, 70389L: 2, 101114L: 2, 107392L: 2, 83204L: 2, 54614L: 2, 83345L: 2, 92572L: 2, 49570L: 2, 95146L: 2, 83016L: 2, 74142L: 2, 49622L: 2, 55266L: 2, 69629L: 2, 77818L: 2, 55291L: 2, 73727L: 2, 55297L: 1, 54787L: 1, 80900L: 1, 108039L: 1, 51208L: 1, 105993L: 1, 74146L: 1, 55313L: 1, 49683L: 1, 49174L: 1, 50204L: 1, 85021L: 1, 55327L: 1, 85025L: 1, 96694L: 1, 79907L: 1, 80422L: 1, 98855L: 1, 70184L: 1, 74281L: 1, 107565L: 1, 109106L: 1, 75187L: 1, 76340L: 1, 49205L: 1, 56756L: 1, 80223L: 1, 83518L: 1, 107111L: 1, 76353L: 1, 72459L: 1, 50244L: 1, 69046L: 1, 106268L: 1, 56903L: 1, 83529L: 1, 109132L: 1, 108621L: 1, 109134L: 1, 77391L: 1, 108626L: 1, 106456L: 1, 99912L: 1, 77241L: 1, 57435L: 1, 74332L: 1, 55482L: 1, 108638L: 1, 43109L: 1, 109664L: 1, 77371L: 1, 106082L: 1, 50275L: 1, 70244L: 1, 95334L: 1, 49766L: 1, 49767L: 1, 56936L: 1, 95338L: 1, 74859L: 1, 78444L: 1, 79983L: 1, 89716L: 1, 50805L: 1, 56950L: 1, 77436L: 1, 77438L: 1, 50303L: 1, 60255L: 1, 107142L: 1, 89735L: 1, 77630L: 1, 95683L: 1, 77456L: 1, 107666L: 1, 76947L: 1, 49302L: 1, 86127L: 1, 107381L: 1, 104225L: 1, 81058L: 1, 109220L: 1, 98982L: 1, 77479L: 1, 57512L: 1, 77340L: 1, 90795L: 1, 108205L: 1, 108539L: 1, 89264L: 1, 76978L: 1, 80756L: 1, 77498L: 1, 73915L: 1, 80572L: 1, 107709L: 1, 89547L: 1, 107637L: 1, 107713L: 1, 76483L: 1, 57543L: 1, 95945L: 1, 77514L: 1, 50382L: 1, 76497L: 1, 1059L: 1, 108758L: 1, 106682L: 1, 51416L: 1, 49881L: 1, 77531L: 1, 85213L: 1, 107995L: 1, 69855L: 1, 108772L: 1, 86309L: 1, 108775L: 1, 77545L: 1, 69867L: 1, 50413L: 1, 80936L: 1, 80626L: 1, 50931L: 1, 76532L: 1, 101110L: 1, 30847L: 1, 87850L: 1, 69886L: 1, 88320L: 1, 78297L: 1, 70406L: 1, 95495L: 1, 49928L: 1, 77578L: 1, 109323L: 1, 73944L: 1, 84757L: 1, 412950L: 1, 49943L: 1, 70677L: 1, 83741L: 1, 106270L: 1, 69921L: 1, 108611L: 1, 50139L: 1, 69876L: 1, 107814L: 1, 77607L: 1, 49450L: 1, 86315L: 1, 86319L: 1, 55603L: 1, 50996L: 1, 88136L: 1, 69428L: 1, 50399L: 1, 49980L: 1, 80190L: 1, 51007L: 1, 51008L: 1, 70368L: 1, 107330L: 1, 80709L: 1, 102097L: 1, 108872L: 1, 78811L: 1, 54755L: 1, 80720L: 1, 107858L: 1, 80723L: 1, 106638L: 1, 108375L: 1, 45912L: 1, 48986L: 1, 108381L: 1, 51038L: 1, 51039L: 1, 107360L: 1, 80329L: 1, 55139L: 1, 51569L: 1, 51060L: 1, 54645L: 1, 70377L: 1, 83007L: 1, 98794L: 1, 61310L: 1, 57727L: 1, 80256L: 1, 51073L: 1, 77188L: 1, 74117L: 1, 81286L: 1, 82825L: 1, 107927L: 1, 80269L: 1, 41872L: 1, 51090L: 1, 103383L: 1, 31124L: 1, 93590L: 1, 83863L: 1, 50589L: 1, 49566L: 1, 87456L: 1, 80801L: 1, 51107L: 1, 100824L: 1, 108455L: 1, 106923L: 1, 110508L: 1, 69647L: 1, 95150L: 1, 79795L: 1, 76212L: 1, 70582L: 1, 55736L: 1, 54713L: 1, 50106L: 1, 74683L: 1, 70590L: 1, 95679L: 1, 51139L: 1, 70596L: 1, 108485L: 1, 55750L: 1, 50252L: 1, 80331L: 1, 76962L: 1, 70359L: 1, 106448L: 1, 83883L: 1, 51668L: 1, 98261L: 1, 31914L: 1, 77272L: 1, 104409L: 1, 85209L: 1, 108507L: 1, 106490L: 1, 49630L: 1, 80863L: 1, 79841L: 1, 70627L: 1, 70629L: 1, 29635L: 1, 29671L: 1, 76264L: 1, 77289L: 1, 80362L: 1, 98798L: 1, 110581L: 1, 76278L: 1, 110072L: 1, 98986L: 1})

Метрики.

Метрики, относящиеся к странице *(Language, Number of links, Size of file, Amount of words)*, ко всей коллекции *(File with highest # of links, File with lowest # of links, File with highest # of words, File with lowest # of words, Biggest file in the collection, Smallest file in the collection)* могут быть основой для дальнейших исследований.

Группа времени *(Start time, Finish time, Crawled time)* позволяет оценить объем обработанных данных и производительность скрипта.

*«Languages in the collection with # of files»* позволяют определить численность языковых групп коллекции.

*«Sizes (bytes) in the collection with # of files»* позволяют определить численность размерных групп коллекции.

Варианты поставки — в виде скриптa на *Python 2.7* (для работы необходимы библиотеки *BeautifulSoup 3(4), matplotlib, guess\_language* (экспериментальная библиотека, определяющая язык текста на основе n-грамм)), в виде набора библиотек и файла «*analyzer.exe»* с иконкой.