In [1]:

```
import os
import logging
from time import sleep
from heapq import heappop, heappushpop

import networkx as nx

from link_parser import LinkParser

logging.basicConfig(filename="/media/ssd/simple.wiki/spider.log",
                    format="%(asctime)s - %(levelname)s - %(message)s",
                    datefmt='%I:%M:%S %p',
                    filemode='w',
                    level=logging.INFO)
```

На этапе фильтрации при помощи регулярного выражения отбираются url-ы, ведущие на уникальные вики-страницы. Эти страницы имеют доменное имя simple.wikipedia.org и лежат по пути wiki. По интересующему пути simple.wikipedia.org/wiki/ также лежат служебные страницы, начинающиеся на File:, Help:, Special: и т.д., которые отбрасываются во время фильтрации (при этом нельзя отбрасывать любой url с символом ':', т.к. существуют такие страницы, как "ISO_3166-2:BR" или "Star_Trek:_Phase_II"). Ещё одним особым случаем является url по вышеуказанному пути, чья уникальная часть начинается на символ '#' (например, simple.wikipedia.org/wiki/#blabla) - такой url редиректит на главную страницу. Его также считаем невалидным (*встретился после посещения половины всех страниц при самом первом заходе).

На этапе нормализации из url-а отбрасывается часть, начинающаяся с хештега, т.к. она содержит в себе лишь информацию об определённом положении на странице. Также отсекается внутренний путь страницы, идущий за слешем после названия страницы.

При обходе поддерживаются множества url-ов to_visit и visited (в первом не могут содержаться объекты из второго). Также во время обхода строится ориентированный граф из страниц для дальнейшего вычисления PageRank-а.

In [3]:

```
%%time
link_parser = LinkParser()
to_visit = {"https://simple.wikipedia.org/wiki/Main_Page"}  # starting url
visited = set()
graph = nx.DiGraph()
bad_urls = []

def spider(delay=0.1):
    global link_parser, to_visit, visited, graph
    while len(to_visit):
        url = to_visit.pop()
        logging.info("%s: %s", len(visited), url)
        graph.add_node(url)
        visited.add(url)
        try:
            links = link_parser.get_links(url)
        except Exception as e:
            print e
            logging.error(e)
            bad_urls.append(url)
            continue
        to_visit = to_visit.union(links - visited)
        graph.add_edges_from(zip([url] * len(links), links))
        sleep(delay)

spider(0.01)
```

```
URL u'https://simple.wikipedia.org/wiki/Seljuk_Sultanate_of_R\xfbm'
 contains non-ASCII characters
[Errno socket error] [Errno 104] Connection reset by peer
[Errno socket error] [Errno -3] Temporary failure in name resolution
[Errno socket error] [Errno -3] Temporary failure in name resolution
[Errno socket error] [Errno -3] Temporary failure in name resolution
[Errno socket error] [Errno -3] Temporary failure in name resolution
URL u'https://simple.wikipedia.org/wiki/Berg\xfcn' contains non-ASCI
I characters
URL u'https://simple.wikipedia.org/wiki/Kingdom_of_France_(1791\u201
31792)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/History_of_Sweden_(800\u2013
1521)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Guimar\xe3es' contains non-A
SCII characters
URL u'https://simple.wikipedia.org/wiki/Republic_of_Lithuania_(1918
\u20131940)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/History_of_Sweden_(1772\u201
31809)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/B\xfcndnis_90' contains non-
ASCII characters
URL u'https://simple.wikipedia.org/wiki/Lithuanian_Soviet_Socialist_
Republic_(1918\u20131919)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Republic_of_the_Congo_(L\xe9
opoldville)' contains non-ASCII characters
[Errno socket error] [Errno 101] Network is unreachable
URL u'https://simple.wikipedia.org/wiki/History_of_the_Philippines_
(1946\u20131965)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Duchy_of_Limburg_(1839\u2013
1867)' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/General_classification_in_th
e_Vuelta_a_Espa\xf1a' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Denmark\u2013Norway' contain
s non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Electorate_of_W\xfcrttember
g' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Emirate_of_C\xf3rdoba' conta
ins non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Free_People%27s_State_of_W\x
fcrttemberg' contains non-ASCII characters
URL u'https://simple.wikipedia.org/wiki/Free_Peoples%27_State_of_W\x
fcrttemberg' contains non-ASCII characters
```

Как видно, у небольшого числа урлов оказалась проблема с кодировкой, дропнем их из графа. И посмотрим сколько страниц нашли.

In [6]:

```
graph.remove_nodes_from(bad_urls)
```

In [7]:

```
graph.number_of_nodes()
```

Out[7]:

135318

135318 из 123771 уникальных страниц. Предполагаю, что причина превосходства числа найденных страниц связанно с наличием редиректов, которые не предлагалось обрабатывать.

Посчитаем PageRank и найдём 20 топовых страниц.

In [17]:

```
%%time
pagerank = nx.pagerank(graph)
```

```
CPU times: user 1min 38s, sys: 924 ms, total: 1min 39s
Wall time: 1min 38s
```

In [18]:

```
top20 = [0.] * 20
for url, rank in pagerank.iteritems():
    heappushpop(top20, (rank, url))
```

In [19]:

```
while top20:
    rank, url = heappop(top20)
    print rank, url
```

```
0.0011737667132 https://simple.wikipedia.org/wiki/Germany
0.001187208145 https://simple.wikipedia.org/wiki/Association_footbal
l
0.00119092877372 https://simple.wikipedia.org/wiki/Movie
0.00119112473273 https://simple.wikipedia.org/wiki/Government
0.00120016830024 https://simple.wikipedia.org/wiki/Europe
0.00120024361172 https://simple.wikipedia.org/wiki/Television
0.00124078709979 https://simple.wikipedia.org/wiki/City
0.00133306869528 https://simple.wikipedia.org/wiki/England
0.00138645648289 https://simple.wikipedia.org/wiki/Wikimedia_Commons
0.00142378722101 https://simple.wikipedia.org/wiki/Country
0.00150814436333 https://simple.wikipedia.org/wiki/International_Sta
ndard_Book_Number
0.00154114677339 https://simple.wikipedia.org/wiki/Geographic_coordi
nate_system
0.00158951577699 https://simple.wikipedia.org/wiki/United_Kingdom
0.00161383908321 https://simple.wikipedia.org/wiki/English_language
0.00171625717977 https://simple.wikipedia.org/wiki/Japan
0.00194699256684 https://simple.wikipedia.org/wiki/Definition
0.0022335082908 https://simple.wikipedia.org/wiki/France
0.00421364219491 https://simple.wikipedia.org/wiki/United_States
0.00523352769261 https://simple.wikipedia.org/wiki/Multimedia
0.0403967868691 https://simple.wikipedia.org/wiki/Main_Page
```