PySpark configuration

In [1]:
```
import os

root_dir = os.path.dirname(os.path.abspath('PageRank_IMDB.jpynb'))

content_dir = os.path.join(root_dir, "content/")
if not os.path.isdir(content_dir):
    os.mkdir(content_dir)

kaggle_dir = os.path.join(root_dir, ".kaggle/")
if not os.path.isdir(kaggle_dir):
    os.mkdir(kaggle_dir)

variables_dir = os.path.join(content_dir, "variables/")
if not os.path.isdir(variables_dir):
    os.mkdir(variables_dir)
```

In [2]:
```
# DO NOT RUN ON DEBIAN VM, JDK IS PREINSTALLED
!sudo apt-get install openjdk-11-jdk-headless -qq > /dev/null
```

In [3]:
```
!wget -q https://archive.apache.org/dist/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz  #https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz #http://www-eu.apache.org/dist/spark/spark
!tar xf spark-3.2.0-bin-hadoop3.2.tgz
!pip install -q findspark
```

In [4]:
```
!pip install py4j
```

```
Requirement already satisfied: py4j in /opt/conda/lib/python3.7/site-packages (0.10.9.2)
```

In [5]:
```
import gc
import json
import zipfile
import pickle
import pandas as pd
import numpy as np
import networkx as nx
import matplotlib.pyplot as plt
import sys

os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = os.path.join(root_dir, "spark-3.2.0-bin-hadoop3.2")

import findspark
findspark.init("spark-3.2.0-bin-hadoop3.2")# SPARK_HOME
from pyspark.sql import SparkSession

def getsize(obj):
    print('{:.2f} MB'.format(sys.getsizeof(obj)/(2**20)))
```

In [6]:
```
#!pip install pyspark
```

In [7]:
```
#from pyspark import SparkContext
#from pyspark.sql import SparkSession
```

In [8]:
```
#sc = SparkContext("local", "amd")
```

In [10]:
```
spark = SparkSession.builder.master("local[*]").config("spark.driver.memory", "30g").getOrCreate()
sc = spark.sparkContext
```

Load variables

In [11]:
```
with open(os.path.join(variables_dir, 'actors.pkl'), 'rb') as inpt:
    actors = pickle.load(inpt)
```

In [12]:
```
with open(os.path.join(variables_dir, 'connection_matrix.pkl'), 'rb') as inpt:
    connection_matrix = pickle.load(inpt)
```

PageRank

In [13]:
```
def l2distance(v, q):
    if len(v) != len(q):
        raise ValueError('Cannot compute the distance'
                         ' of two vectors of different size')
    return np.sqrt(sum([(q_el - v_el)**2 for v_el, q_el in zip(v, q)]))
```

In [14]:
```
def get_page_rank(n, connection_matrix, beta,
                  checkpoint_pr = None, verbose=False, tolerance=10e-5, max_iterations=100):

    links_RDD = sc.parallelize(connection_matrix, numSlices=1000).cache()
    telep = (1.-beta)/n

    if(verbose):
        print('RDD created')

    if checkpoint_pr is None:
        page_rank = np.ones(n)/n
    else:
        page_rank = checkpoint_pr

    old_page_rank = np.ones(n)

    if(verbose):
        print('Start: ', page_rank, '\n ---------------------- \n')

    iteration = 0
    while l2distance(old_page_rank, page_rank) >= tolerance and \
            iteration < max_iterations:
        old_page_rank = page_rank
        page_rank_values = (links_RDD
                            .map(lambda t: (t[0], beta*t[2]*page_rank[t[1]]))
                            .reduceByKey(lambda a, b: a+b)
                            .map(lambda x: (x[0], x[1]+telep))
                            .sortByKey()
                            .collect()
                           )

        if(verbose):
            print(f'Map and reduce step {iteration+1} completed.')
            #print(f'Size of page_rank_values = {sys.getsizeof(page_rank_values)/1024} MiB')

        out_nodes = [n for n, r in page_rank_values]
        if len(out_nodes) < n:
            missing_nodes = list()
            c = 0
            for i in out_nodes:
                while i > c:
                    missing_nodes.append(c)
                    c = c+1
                c = c+1
            if c > i:
                missing_nodes = missing_nodes + list(range(c,n))
            page_rank_values = page_rank_values + list(zip(missing_nodes, [telep]*len(missing_nodes)))

        page_rank = np.array([c for (i, c) in sorted(page_rank_values, key = lambda x: x[0])])

        if verbose:
            print(page_rank)

        with open(os.path.join(variables_dir, 'page_rank.pkl'), 'wb') as outp:
            pickle.dump(page_rank, outp)

        if verbose:
            print("Written: ", os.path.join(variables_dir, 'page_rank.pkl'))

        iteration += 1

    print('{} iterations'.format(iteration))

    return page_rank
```

In [17]:
```
with open(os.path.join(variables_dir, 'page_rank.pkl'), 'rb') as inpt:
    page_rank = pickle.load(inpt)
```

In [18]:
```
page_rank = get_page_rank(n=len(actors), connection_matrix=connection_matrix, beta=0.9, checkpoint_pr=page_rank,
                          verbose=True, tolerance=10e-10, max_iterations=50)
```

```
RDD created
Start:  [2.63615019e-06 3.96164657e-06 2.28161628e-06 ... 4.58057492e-07
 1.39410144e-07 4.58057492e-07]
 ----------------------

Map and reduce step 1 completed.
[2.63615201e-06 3.96164753e-06 2.28161721e-06 ... 4.58057413e-07
 1.39410161e-07 4.58057413e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 2 completed.
[2.63615350e-06 3.96164835e-06 2.28161802e-06 ... 4.58057347e-07
 1.39410177e-07 4.58057347e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 3 completed.
[2.63615493e-06 3.96164904e-06 2.28161873e-06 ... 4.58057292e-07
 1.39410192e-07 4.58057292e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 4 completed.
[2.63615611e-06 3.96164963e-06 2.28161935e-06 ... 4.58057246e-07
 1.39410205e-07 4.58057246e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 5 completed.
[2.63615713e-06 3.96165013e-06 2.28161989e-06 ... 4.58057209e-07
 1.39410218e-07 4.58057209e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 6 completed.
[2.63615801e-06 3.96165055e-06 2.28162036e-06 ... 4.58057178e-07
 1.39410229e-07 4.58057178e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 7 completed.
[2.63615877e-06 3.96165090e-06 2.28162077e-06 ... 4.58057152e-07
 1.39410239e-07 4.58057152e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 8 completed.
[2.63615943e-06 3.96165120e-06 2.28162113e-06 ... 4.58057131e-07
 1.39410248e-07 4.58057131e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 9 completed.
[2.63615999e-06 3.96165145e-06 2.28162144e-06 ... 4.58057114e-07
 1.39410257e-07 4.58057114e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 10 completed.
[2.63616049e-06 3.96165166e-06 2.28162171e-06 ... 4.58057099e-07
 1.39410264e-07 4.58057099e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 11 completed.
[2.63616091e-06 3.96165183e-06 2.28162195e-06 ... 4.58057088e-07
 1.39410271e-07 4.58057088e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 12 completed.
[2.63616128e-06 3.96165198e-06 2.28162216e-06 ... 4.58057078e-07
 1.39410278e-07 4.58057078e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 13 completed.
[2.63616160e-06 3.96165210e-06 2.28162234e-06 ... 4.58057070e-07
 1.39410283e-07 4.58057070e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 14 completed.
[2.63616187e-06 3.96165221e-06 2.28162250e-06 ... 4.58057064e-07
 1.39410288e-07 4.58057064e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 15 completed.
[2.63616211e-06 3.96165229e-06 2.28162264e-06 ... 4.58057059e-07
 1.39410293e-07 4.58057059e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 16 completed.
[2.63616231e-06 3.96165236e-06 2.28162276e-06 ... 4.58057055e-07
 1.39410297e-07 4.58057055e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 17 completed.
[2.63616249e-06 3.96165242e-06 2.28162286e-06 ... 4.58057052e-07
 1.39410301e-07 4.58057052e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 18 completed.
[2.63616264e-06 3.96165247e-06 2.28162296e-06 ... 4.58057049e-07
 1.39410304e-07 4.58057049e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 19 completed.
[2.63616277e-06 3.96165251e-06 2.28162304e-06 ... 4.58057047e-07
 1.39410307e-07 4.58057047e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl

Map and reduce step 20 completed.
[2.63616289e-06 3.96165254e-06 2.28162311e-06 ... 4.58057046e-07
 1.39410310e-07 4.58057046e-07]
Written:  /home/jupyter/content/variables/page_rank.pkl
20 iterations
```

In [ ]:
```

```