HW 5 - Page Rank

MIDS w261: Machine Learning at Scale | UC Berkeley School of Information | Fall 2018

In Weeks 8 and 9 you discussed key concepts related to graph based algorithms and implemented SSSP.
In this final homework assignment you'll implement distributed PageRank using some data from Wikipedia. By the end of this homework you should be able to:

- ... **compare/contrast** adjacency matrices and lists as representations of graphs for parallel computation.
- ... **explain** the goal of the PageRank algorithm using the concept of an infinite Random Walk.
- ... **define** a Markov chain including the conditions underwhich it will converge.
- ... **identify** what modifications must be made to the web graph inorder to leverage Markov Chains.
- ... **implement** distributed PageRank in Spark.

Please refer to the `README` for homework submission instructions and additional resources.

Cmd 2

Notebook Set-Up

Before starting your homework run the following cells to confirm your setup.

Cmd 3

```
# imports
import re
import ast
import time
import numpy as np
import pandas as pd
import seaborn as sns
import networkx as nx
import matplotlib.pyplot as plt
```

```
/databricks/python/lib/python3.7/site-packages/networkx/classes/graph.py:23: DeprecationWarning: Using or importing the ABCs from
'collections' instead of from 'collections.abc' is deprecated, and in 3.8 it will stop working
  from collections import Mapping
```

Command took 1.01 seconds -- by tonydisera@ischool.berkeley.edu at 3/17/2020, 7:38:50 PM on w261-homework-HC-section1

Cmd 4

Run the next cell to create your directory in dbfs

You do not need to understand this scala snippet. It simply dynamically fetches your user directory name so that any files you write can be saved in your own directory.

Cmd 5

```
# RUN THIS CELL AS IS
# This code snippet reads the user directory name, and stores is in a python variable.
# Next, it creates a folder inside your home folder, which you will use for files which you save inside this notebook.
username = dbutils.notebook.entry_point.getDbutils().notebook().getContext().tags().apply('user')
userhome = 'dbfs:/user/' + username
print(userhome)
hw5_path = userhome + "/HW5/"
hw5_path_open = '/dbfs' + hw5_path.split(':')[-1] # for use with python open()
dbutils.fs.mkdirs(hw5_path)
```

```
dbfs:/user/tonydisera@ischool.berkeley.edu
Out[2]: True
```

Command took 0.09 seconds -- by tonydisera@ischool.berkeley.edu at 3/17/2020, 7:38:50 PM on w261-homework-HC-section1

Cmd 6