Homework 3

Released: Oct. 17

Due Date: Oct.25

Name: Marlene Lin

Collaborators: Very fine caffeine

This homework focuses on topics related to classes, inheritance, exceptions, and iterators.

I encourage collaborating with your peers, but the final text, code, and comments in this homework assignment should still be written by you.

Submission instructions:

- Convert this notebook into a pdf file and submit it on GradeScope under "HW3 - PDF".
- Compress files `PageRank.py` , and `utils.py` into one `HW3.zip` , and submit it to Gradescope under "HW3 - Autograder". Make sure the python files are at the top of the directory when you zip them, and don't change their name.

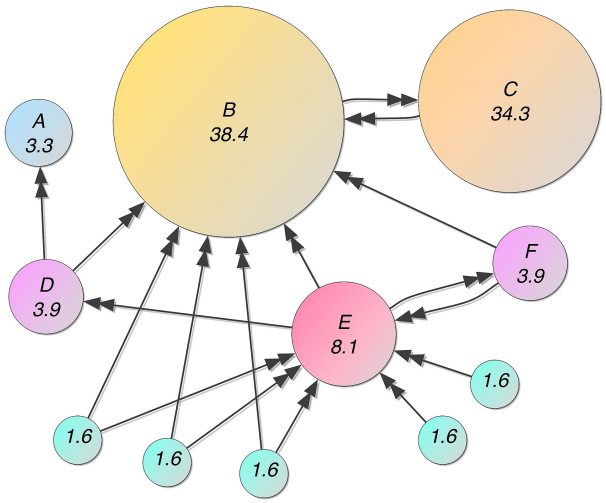
PageRank

What is the most important website on the internet? Who is the "key player" on a sports team? Which countries are the most central players in the world economy? There is no one correct answer to any of these questions, but there is a most profitable one. PageRank is an algorithm for ranking individual elements of complex systems, invited by Sergey Brin and Larry Page. It was the first and most famous algorithm used by the Google Search engine, and it is fair to say that the internet as we know it today would not exist without PageRank.

In this assignment, we will implement PageRank. There are many good ways to implement this algorithm, but in this assignment we will use our newfound skills with object-oriented programming and iterators.

How it works

For the purposes of this example, let's assume that we are talking about webpages. PageRank works by allowing a "random surfer" to move around webpages by following links. Each time the surfer lands on a page, it then looks for all the links on that page. It then picks one at random and follows it, thereby arriving at the next page, where the process repeats. Because the surfer moves between linked pages, PageRank expresses an intuitive idea: **important pages are linked to other important pages.** This diagram from Wikipedia gives a nice illustration. Note that more important webpages (higher PageRank) tend to be connected to other important webpages.

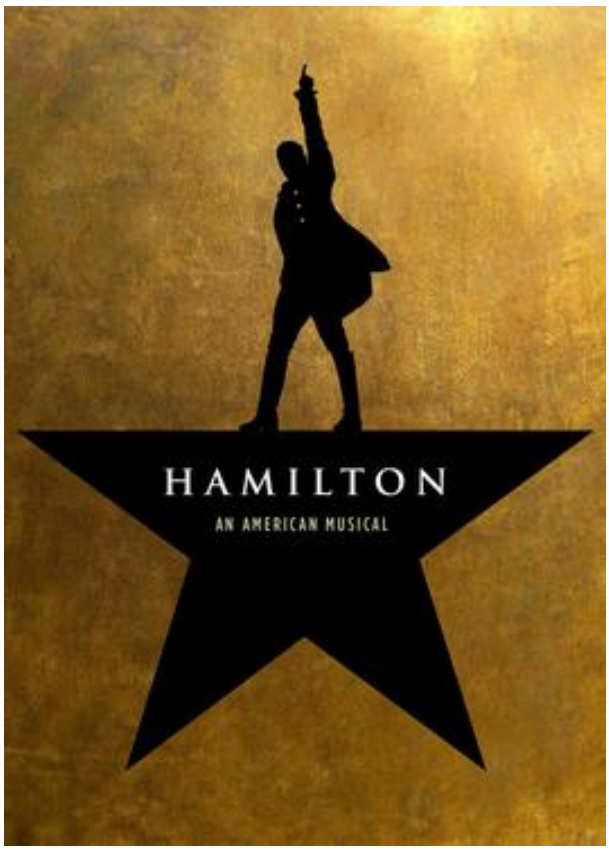


A schematic for PageRank.

(There's a small issue related to webpages that don't have any links, which we'll address later on in the assignment. This is the "damping factor" in the diagram, which comes into play when we implement "teleportation" below.)

Data

This data set comes from the hit Broadway musical "Hamilton."



The Hamilton data set

The good folks at The Hamilton Project analyzed the script for us, obtaining data on **who talks about whom** in each of the show's songs. When character A mentions character B, we'll think

of this as a *link* from A to B, suggesting that B might be important.

Listening to the soundtrack while working is strongly recommended.

In [1]:
```
# import functions and classes from other files.
from utils import retrieve_data, read_data, describe, data_to_dictionary
from PageRank import PageRankDiGraph, PageRankIterator
import numpy as np
import random
```

Problem 1

Part (a): Call `retrieve_data` and `read_data` with proper arguments

In `utils.py`, take a look at the two functions that are already defined. The first one `retrieve_data` retrieves the data from the internet and saves it to your local computer, while the second `read_data` reads in the data from the local copy, producing a list of tuples.

The Hamilton dataset lives at the following URL:

`https://philchodrow.github.io/PIC16A/homework/HW3-hamilton-data.csv`

Each row corresponds to a "link" between objects, and the pairs have format `mentioner, mentioned`.

The cell below

- sets the variable `url`,
- calls `retrieve_data` and `read_data` with proper arguments
- saves the return value of `read_data` in a variable called `data`.

You don't need to change anything in this cell, but read and make sure you understand what the lines are doing. It's not important for you to understand the code inside these functions right now -- we'll discuss them in a coming week.

In [2]:
```
url = "https://philchodrow.github.io/PIC16A/homework/HW3-hamilton-data.csv"

fname = "data.csv"
retrieve_data(url, fname)
data = read_data(fname)

#fname = "data.csv"
#data = read_data(fname)
```

Part (b): Define `describe` in utils.py

This would also be a good time to inspect the data to make sure you understand how it is structured. Write a function `describe` that describes the meaning of the `n` th row of `data`. Running `describe(data, 5)` on the Hamilton data set should print the following:

```
"Element 5 of the Hamilton data set is ('burr', 'betsy'). This means
that Burr mentions Betsy in a song."
```

Please attend to capitalization and formatting. While the standard string concatenation operator `+` is completely fine for this task, the fancy `str.format()` function may make your code somewhat simpler. This page has some useful examples in case you'd like to try this.

In [3]:
```
# test your describe function here
describe(data,0)
```

```
Element 0 of the Hamilton data set is ('burr', 'hamilton'). This means that Burr
mentions Hamilton in a song.
```

Part (c): Define `data_to_dictionary` in utils.py

Write a function called `data_to_dictionary` that converts the data into a dictionary such that:

1. There is a single key for each character in Hamilton.
2. The value corresponding to each key is a list of the characters/airports to which that key links. The list should contain repeats if there are multiple links.

Here's an example of the desired behavior on a fake data set.

```
data = [("a", "b"),
        ("a", "b"),
        ("a", "c"),
        ("b", "c"),
        ("b", "a")]

data_to_dictionary(data)

# output
{"a" : ["b", "b", "c"], "b" : ["a", "c"]}
```

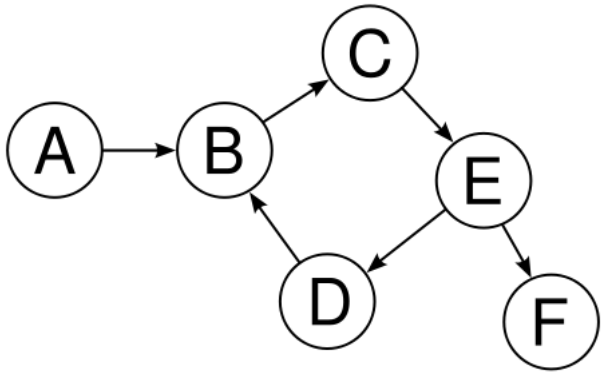
In [4]:
```
# run this code to test

toy_data = [("a", "b"), ("a", "b"), ("a", "c"), ("b", "c"), ("b", "a")]
D = data_to_dictionary(toy_data)
print(D)
```

```
{'a': ['b', 'b', 'c'], 'b': ['c', 'a']}
```

Problem 2: Define `PageRankDiGraph` class in PageRank.py

A **directed graph**, or DiGraph, is just a set of arrows ("edges") between objects ("nodes"). It is a natural way to represent data that represents one-way relationships, such as links from one webpage to another or mentions of one character by another. We already saw a directed graph above when we introduced the idea of PageRank. Here's a paired-down example.



Example of a directed graph.

Implement a `PageRankDiGraph` class according to the following specs. https://docs.python.org/3/reference/datamodel.html#special-method-names might be helpful. Let `G`, `G1`, `G2` be instances of `PageRankDiGraph`.

- The `__init__` method should accept one argument, `data`, which you can expect to be a list of tuples like our Hamilton data. It should save `data` as `self.data`, then construct an instance variable `self.link_dict` which is simply the output of `data_to_dictionary` applied to the argument `data`.
- `get_nodes` returns a list of all nodes in the graph.
- Running `{character name} in G` should return `True` if the character is a node in the graph, `False` otherwise.
- `print(G)` should print `PageRankDiGraph with {number of nodes} nodes and {number of edges} edges.`
- `G1 + G2` returns a new instance of `PageRankDiGraph` that contains both the edges of `G1` and `G2`.
- Define a method `self.linked_by(x)` which, when called, returns the list of characters that `x` connects to. Hint: It should be a one liner using `self.link_dict`.

Example:

```
D = PageRankDiGraph(data)
print(D.linked_by('peggy'))

# output
['peggy', 'schuylerSis']
```

In [5]:

```
G = PageRankDiGraph(data)
print("(1)\n",'len of the data is',len(G))
print("(2)\n", 'the nodes in the data is ', G.get_nodes())
print("(3)\n","hamilton is in the data: ", "hamilton" in G)
print("(4)\n", 'people mentioned by hamilton: ', G.linked_by('hamilton'))
print("(5)\n",'Description of G: ',G)
```

```
# test different methods of PageRankDiGraph here
```

```
(1)
 len of the data is 293
(2)
 the nodes in the data is  ['peggy', 'schuylerSis', 'paine', 'philipS', 'angelic
a', 'admiralHowe', 'knox', 'lafayette', 'madison', 'philipH', 'marthaWashingto
n', 'mulligan', 'rochambeau', 'doctor', 'eacker', 'kingLouis', 'kingGeorge', 'we
eks', 'theodosiaMother', 'generalMontgomery', 'company', 'jay', 'burr', 'green',
'reynolds', 'conway', 'seabury', 'theodosiaDaughter', 'generalMercer', 'hamilto
n', 'jAdams', 'washington', 'maria', 'sally', 'franklin', 'jefferson', 'betsy',
'ensemble', 'eliza', 'ness', 'lee', 'sAdams', 'women', 'men', 'pendleton', 'laur
ens']
(3)
 hamilton is in the data:  True
(4)
 people mentioned by hamilton:  ['burr', 'angelica', 'philipH', 'lafayette', 'el
iza', 'laurens', 'mulligan', 'washington', 'eliza', 'lee', 'laurens', 'conway',
'hamilton', 'washington', 'lee', 'laurens', 'burr', 'washington', 'hamilton', 'b
urr', 'lee', 'burr', 'eliza', 'peggy', 'angelica', 'hamilton', 'laurens', 'mulli
gan', 'lafayette', 'burr', 'kingGeorge', 'burr', 'lafayette', 'laurens', 'burr',
'hamilton', 'reynolds', 'eliza', 'angelica', 'philipH', 'eliza', 'eacker', 'phil
ipH', 'eliza', 'reynolds', 'jefferson', 'madison', 'burr', 'reynolds', 'washingt
on', 'jefferson', 'washington', 'kingLouis', 'lafayette', 'burr', 'burr', 'angel
ica', 'maria', 'reynolds', 'angelica', 'madison', 'jefferson', 'eliza', 'schuyle
rSis', 'jAdams', 'jefferson', 'washington', 'madison', 'jefferson', 'hamilton',
'philipH', 'eliza', 'burr', 'jefferson', 'jAdams', 'burr', 'hamilton', 'burr',
'laurens', 'washington', 'eliza']
(5)
 Description of G:  PageRankDiGraph with 46 nodes and 293 edges.
```

In [ ]:

```
```

Problem 3: Define `PageRankIterator` class in PageRank.py

Define a `PageRankIterator` class that iterates through a `PageRankDiGraph` via a custom `__next__` method.

When initialized, this class should create instance variables to store:

- `graph` , the `PageRankDiGraph` instance, given as input
- `iteration_limit` , an integer given as input,
- `jump_prob` , a number between 0 and 1 (inclusive), given as input,
- a counter `iter` , starting at 0, to log the number of steps taken,
- `current_state` variable whose value is one of the keys of the `link_dict` of the `PageRankDiGraph` . You can choose its initial value arbitrarily; in my code I chose `self.current_state = "hamilton"` .

Your `__init__` method should check that the input `graph` is an instance of `PageRankDiGraph` , and raise `TypeError` if it is not. Hint: which one of these would be useful? https://docs.python.org/3/library/functions.html

We are going to use iteration to implement the PageRank algorithm. This means we are going to imagine a surfer who is following the links in our data set. **Implement the following two methods:**

1. `follow_link()` .
   A. Pick a random new character mentioned by the current character, or new airport served by the current airport. Let's call this `next_state` .
   B. If `next_state != current_state` , set `current_state` to `next_state` .
   C. Otherwise (if `next_state == current_state`), teleport (see below).
   D. You might run into `KeyError` s, in which case you should again teleport (use a `try-except` block).
2. `teleport()` .
   A. Set the current state to a new state (key of the link dict) completely at random.

Hint: use `random.choice` from the `random` module to choose elements of lists.

Finally, **implement** `__next__()` . `__next__()` should do `follow_link()` with `jump_prob` % probability, and do `teleport()` with `1-jump_prob` % probability. Then return the `current_state` . You should also define a custom `StopIteration` condition to ensure that only as many steps are taken as the `iteration_limit` supplied to the `PageRankDiGraph` initializer.

1. To do something with 85% probability, use the following:

```
if random.random() < 0.85:
    # do the thing
else:
    # do the other thing
```

Example Usage

After you define your class, run the following code and show that it works. Note: your precise sequence may be different from mine.

```
I = PageRankIterator(G, 5, 0.6)
for x in I:
    print(x)
```

```
following link : current state = burr
following link : current state = washington
following link : current state = burr
following link : current state = hamilton
teleporting    : current state = washington
```

I have added printed messages here for you to more clearly see what should be happening, but it is not necessary for you to do this. It is sufficient for your output to look like:

```
burr
washington
burr
```

```
hamilton
washington
```

In [6]:

```
I = PageRankIterator(G, 5, 0.6)
for x in I:
    print(x)
```

```
doctor
hamilton
lafayette
men
peggy
jefferson
Reach iteration limit
```

Problem 4: Implement class `IterablePageRankDiGraph` from scratch in PageRank.py.

`IterablePageRankDiGraph` is a subclass of `PageRankDiGraph`, and should inherit all of `PageRankDiGraph`'s methods. You need to define three methods, `__init__`, `__str__`, and `__iter__` such that:

- At `__init__`, in addition to what initializing `PageRankDiGraph` requires, it should also get `iteration_limit` (default 20) and `jump_prob` (default 0.75), and save corresponding instance variables.
- If `IG` is an instance of `IterablePageRankDiGraph`, `print(IG)` should print `IterablePageRankDiGraph with {number of nodes} nodes and {number of edges} edges.`
- `__iter__` returns a new instance of `PageRankIterator` initialized with the `IterablePageRankDiGraph` instance, and its `iteration_limit` and `jump_prob` values.

If successful, the following cell should run without throwing any errors.

In [7]:

```
from PageRank import IterablePageRankDiGraph

# test IterablePageRankDiGraph here
G = IterablePageRankDiGraph(data)
print("(1)\n",'len of the data is',len(G))
print("(2)\n", 'the nodes in the data is ', G.get_nodes())
print("(3)\n","hamilton is in the data: ", "hamilton" in G)
print("(4)\n", 'people mentioned by hamilton: ', G.linked_by('hamilton'))
print("(5)\n",'Description of G: ',G)
```

```
(1)
 len of the data is 293
(2)
 the nodes in the data is  ['peggy', 'schuylerSis', 'paine', 'philipS', 'angelic
a', 'admiralHowe', 'knox', 'lafayette', 'madison', 'philipH', 'marthaWashingto
n', 'mulligan', 'rochambeau', 'doctor', 'eacker', 'kingLouis', 'kingGeorge', 'we
eks', 'theodosiaMother', 'generalMontgomery', 'company', 'jay', 'burr', 'green',
```

```
'reynolds', 'conway', 'seabury', 'theodosiaDaughter', 'generalMercer', 'hamilto
n', 'jAdams', 'washington', 'maria', 'sally', 'franklin', 'jefferson', 'betsy',
'ensemble', 'eliza', 'ness', 'lee', 'sAdams', 'women', 'men', 'pendleton', 'laur
ens']
(3)
 hamilton is in the data:  True
(4)
 people mentioned by hamilton:  ['burr', 'angelica', 'philipH', 'lafayette', 'el
iza', 'laurens', 'mulligan', 'washington', 'eliza', 'lee', 'laurens', 'conway',
'hamilton', 'washington', 'lee', 'laurens', 'burr', 'washington', 'hamilton', 'b
urr', 'lee', 'burr', 'eliza', 'peggy', 'angelica', 'hamilton', 'laurens', 'mulli
gan', 'lafayette', 'burr', 'kingGeorge', 'burr', 'lafayette', 'laurens', 'burr',
'hamilton', 'reynolds', 'eliza', 'angelica', 'philipH', 'eliza', 'eacker', 'phil
ipH', 'eliza', 'reynolds', 'jefferson', 'madison', 'burr', 'reynolds', 'washingt
on', 'jefferson', 'washington', 'kingLouis', 'lafayette', 'burr', 'burr', 'angel
ica', 'maria', 'reynolds', 'angelica', 'madison', 'jefferson', 'eliza', 'schuyle
rSis', 'jAdams', 'jefferson', 'washington', 'madison', 'jefferson', 'hamilton',
'philipH', 'eliza', 'burr', 'jefferson', 'jAdams', 'burr', 'hamilton', 'burr',
'laurens', 'washington', 'eliza']
(5)
 Description of G:  PageRankDiGraph with 46 nodes and 293 edges.
```

Problem 5

Part (a): Compute PageRank

Finally, we are ready to compute the PageRank in our data set. Initialize a `PageRankDiGraph` with a large iteration limit (say, 1,000,000). We will let our surfer randomly move through the data set this many times. The number of times that the surfer visits state `x` is the PageRank score of `x`.

Create a `dict` which logs how many times a given state appears when iterating through the `PageRankDiGraph`.

Your Solution

In [8]:

```
# write your code here
AHgraph = IterablePageRankDiGraph(data,iteration_limit = 1000000)
#AHgraph = PageRankIterator(AHgraph)
print(AHgraph.link_dict.keys())
count = 0
log = {}
for i in AHgraph:
    if i == "hamilton":
        count += 1
    log[i] = log.get(i, 0) + 1
print(count == log["hamilton"])
```

```
dict_keys(['burr', 'hamilton', 'ensemble', 'company', 'men', 'women', 'angelic
a', 'eliza', 'washington', 'mulligan', 'lafayette', 'laurens', 'kingGeorge', 'je
fferson', 'madison', 'philipH', 'lee', 'peggy', 'seabury', 'reynolds', 'docto
r'])
Reach iteration limit
True
```

Part (b): Display Your Result

Use your favorite approach to show the results in sorted format, descending by PageRank score. The entries at the top should be the entries with highest PageRank. You may show either the complete list or just the top 10.

Consider using the sort method of list https://docs.python.org/3/library/stdtypes.html#list.sort with an appropriate lambda expression - `L.sort(key = lambda ...)`

Check your code by comparing your top 10 to mine. Because we are using a randomized algorithm, your results will not agree exactly with mine, but they should be relatively close. If your top 10 list is very different, then you might want to revisit your previous solutions.

My top 10 were:

```
[('hamilton', 101962),
 ('burr', 61948),
 ('washington', 57775),
 ('jefferson', 45480),
 ('jAdams', 32970),
 ('eliza', 32941),
 ('angelica', 31817),
 ('schuylerSis', 25556),
 ('madison', 24702),
 ('lafayette', 23070)]
```

What are the most important elements in the data set? Does it change with different jump probabilities in the iterator? Does it change with different initial states?

In [13]:
```
# write your code here
mention = list(log.items())
mention.sort(key = lambda x: x[1])
```

- The most important elements in the data set is the numbers of occurences of each of the Hamilton character names in all tuples. It is also the 'how connected' (not sure of the term in graph theory) the nodes are in this undirected and unweighted graph. "degree"
- It does not change with the jump probabilities, that is whether teleport/follow link, the elements with most occurences still show up in the top list.

In [22]:
```
mention[-1:-11:-1]
```

Out[22]:
```
[('hamilton', 154049),
 ('burr', 93902),
 ('washington', 88640),
 ('jefferson', 69249),
 ('eliza', 50290),
 ('angelica', 47799),
 ('madison', 36962),
 ('lee', 35050),
 ('lafayette', 34954),
 ('kingGeorge', 32600)]
```

In [ ]: