Web Science & Web Technology: Assignment II

Denis Helic Elisabeth Lex Fiona Draxler Thomas Wurmitzer

14.03.2016, `#0828412`

Deadlines

- The deadline is 2016-05-23 23:59:59.
- The hard deadline is 2016-05-24 23:59:59. Committing files to the repository after the deadline will result in deduction of points (see Introduction slides)!

Prerequisites

Before you can start solving the tasks for assignment **A2** you'll need to ensure that you've installed Python 2.7.x, Python 3.x on your development machine. Solving the tasks using Python 3 *should* probably work, however please note that we do not officially support it and thus cannot offer support if something goes awry.

Assuming you've successfully setup your Python environment, you'll also need the following external modules/libraries. To avoid version related problems please try to stick as close to the recommended versions below or use `pip` to resolve them.

Additional Dependencies

Besides a working Python environment you'll need the following external Python libraries:

- `numpy` >= 1.10
- `networkx` >= 1.11
- `matplotlib` >= 1.5

If you have `pip` installed you can fetch and install those external dependencies in one go by running `pip install -r requirements.txt` inside the assignment folder (`YMMV`).

Introduction

The goal of this exercise is to dive deeper into the topics of link analysis and the concept of strong and weak ties discussed in the previous lectures.

The main focus of this assignment is on understanding and implementing **Hubs & Authorities (HITS)**, **PageRank** and trying out **community detection** as discussed in the lectures.

The workload for this assignment is (slightly) higher than the first one!

This assignment is split into three tasks which can be solved and submitted individually. Each of the those tasks comes with a `skeleton` that you *should* use as your implementation's foundation and a set of tests that you *can* use to check your deliverables and prepare your solutions for submission.

Folder Structure

Below is a simple visualization of the directory structure you should find inside the working copy of your repository after **A2** has been handed out.

- **a1/** - previous assignment A1
- **a2/**
  - `Makefile` - See *Makefile*
  - **`README.pdf`**
  - `communities.py`
  - `hits.py`
  - `pagerank.py`
  - `requirements.txt` - required external libraries (see *Prerequisites*).
  - `valar-morghulis.gml.gz`
  - **submission/**
  - **tests/**
    - `__init__.py`
    - `check_communities_submission.py`
    - `check_hits_submission.py`
    - `check_pagerank_submission.py`
    - `test_communities.py`
    - `test_hits.py`
    - `test_pagerank.py`

Makefile

The `Makefile` packaged with your assignment skeleton provides you with a few convenience functions for building, verifying and testing your implementation and submission.

- Issuing `make` or `make all` inside your assignment folder will execute all tasks one by one. To manually run a specific task say `pagerank`, issue `make pagerank.json` or simply throw the source file at the interpreter by using `python pagerank.py`.
- `make check` will tell you if the results written by a certain or all tasks comply with the assignment requirements i.e. that the files inside the `submission` folder are valid `JSON` files containing the proper `n`-tuple and that `non-zero` plots exist as well. This might help you prepare and verify your deliverables and make your repository ready for submission.
- `make clean` will delete the contents of the `submission` folder as well as *clutter* like `*.pyc` files from your repository.

Tasks

The provided skeletons for each task already provide a starting point for your solution by providing a predefined entry point and means to persist your results to disk. You **must not modify** the header of the 'perform' function in any of the provided task skeletons. Make sure your results are stored properly and in the required order, form and location otherwise you might receive no points for this task!

In this assignment you have to solve the following tasks.

1. `community.py` (10 Points)
2. `pagerank.py` (10 Points)
3. `hits.py` (10 Points)

A detailed description and task-related `TODOs` can be found in the provided task skeletons.

To run any of the tasks a simple `make <taskname>.json` or `python <taskname>.py` in the `a2` folder should suffice. To check your solutions for errors and prepare your repository for submission read the **Testing** and **Submission** section in this document.

If you encounter any kind of problem a short search or consulting the Python documentation, which usually provides detailed documentation including simple examples on each subject, might help.

In any other case, please post general questions regarding the assignment tasks to the `tu-graz.lv.web-science`. If you have a particular, specific problem, you can contact us per e-mail as well. Answering delays might be longer, though.

Testing

This assignment comes with a set of test cases to check if the submission folder contains the corresponding `*.png` and (valid) `*.json` files. However, try not to solely rely on these automated means for testing your solution.

Besides running `make check` to discover and run *all* testcases, they can also be fired up individually by passing the name of the wanted testcase inside the `tests` folder/module (and without the file extension) e.g. to check that the required files for the `pagerank.py` task in the `submission` folder exist:

```
% python -m unittest tests.check_pagerank_submission
...
----------------------------------------------------------------------
Ran 3 tests in 0.001s

OK
```

To verify your implementation against a series of (simple) testcases run:

```
% python -m unittest tests.test_pagerank
....
----------------------------------------------------------------------
Ran 4 tests in 0.010s

OK
```

You can add an additional `-v` before the test case to make the output a little more verbose.

Compliance with the required structure of your submission directory is an absolute must, and testing your submission can give you more confidence that you fulfilled those requirements. However the tests to check and verify your implementation and deliverables are provided for your convenience only. *Use them at your own risk!* Even submissions that pass all the provided tests might get points deducted or no points, given a proper reason!

Submission

Assuming you've implemented all tasks, did modify the task templates where you were supposed to and ran `make` or `make all` the `submission` folder should contain the following files:

- **submission**
  - `pagerank.json`
  - `pagerank.png`
  - `communities.json`
  - `communities.png`
  - `hits.json`
  - `hits.png`

Running `make check` or the underlying command by hand might help you verify that the data inside the `*.json` files is valid and properly formatted (see **Testing**).

To submit your solutions simply `add` the `submission` folder **and** its content to your repository and commit them. Furthermore, **don't forget to commit the code** from each of the tasks! Also *only* committing the code without the results in the submission folder will result in zero points for that task!

Policies

- Some `numpy` and `networkx` functions are not allowed in this assignment (see `TODO`s). Usage of those blacklisted functions will result in **zero** points!
- No other external Python libraries are allowed.
- Copying or using snippets or parts from external resources (even with proper referencing) is not allowed and will result in zero points!
- Discussing results with other students is encouraged, however keep in mind that this is *not* a group exercise!
- Updates regarding the assignment will be posted in the newsgroup. Reading it is therefore mandatory.
- Your Python programs must be executable by invoking `python <taskname>.py` in the assignment folder. No extra parameters must be needed.
- Your scripts shall not produce any output to the standard output on the console. Use an internal variable to enable debug output or use the `logging` module, if you want. Furthermore, plots shall not be `shown` (in a window), they shall just be written to the corresponding PNG files.
- You must not use any platform-specific functions.
- Your programs must not consume more than 64GB of main memory (RAM) each.
- Your code will be checked for plagiarism using automated and manual means.

Resources

- Library Documentation & Tutorials
  - `networkx`
    - http://networkx.github.io/documentation/latest/
    - http://networkx.github.io/documentation/latest/tutorial/
  - `numpy`
    - http://docs.scipy.org/doc/numpy/
  - `matplotlib`
    - http://matplotlib.org/contents.html
- Python Language Reference
  - https://docs.python.org/2/reference/index.html
- Python Tutorials
  - Google's Python Class, https://developers.google.com/edu/python/
  - Think Python: How to Think Like a Computer Scientist, http://www.greenteapress.com/thinkpython/
  - Code Academy's Python Track, http://www.codecademy.com/en/tracks/python
  - The Hitchhiker's Guide to Python!, http://docs.python-guide.org

Please post general questions regarding the assignment tasks to the `tu-graz.lv.web-science`. If you have a particular, specific problem, you can contact us per e-mail as well. Answering delays might be longer, though.