Python (BSU FAMCS Fall'18)

Домашнее задание 3

Преподаватели: Дмитрий Косицин, Светлана Боярович

Ваша задача состоит в том, чтобы посчитать **PageRank** сайта wikipedia на некотором языке (возьмите тот, в котором не менее 10000 статей, или хотя бы кусочек большой wikipedia не менее чем из 25000 статей).

Сначала реализуйте класс, описывающий многопоточный *crawler*, который будет скачивать статью по некоторой ссылке. Не забудьте об использовании таймаута и повторе запроса при неудаче (например, код ошибки 500). Также *crawler* следует ограничить по *интенсивности* запросов (*rate limit*) – проследите, чтобы количество запросов в секунду было не более некоторого заданного k.

Далее, реализуйте класс, который обрабатывает загруженные страницы, извлекает из них ссылки на другие страницы на том же языке для последующего анализа, а также вместе с этим строит граф ссылок.

После того, как построен граф и проанализированы все необходимые статьи, посчитайте **PageRank**. Для этого, если упрощенно, нужно посчитать количество статей, которые ссылаются на некоторую страницу. После этого все полученные значения преобразовать в некоторую шкалу от 0 до n, например, просто прологарифмировав значения, а потом масштабировав их, чтобы значение n не превосходило 10. Можно предложить и свой вариант рассчета. Подсчет **PageRank** также следует организовать параллельно.

Постройте графики распределения количества статей по **PageRank** (аналог количества ссылок *на* страницу) и распределения по степеням вершин (количество ссылок *с* некоторой страницы на другие) полученного графа.

Снабдите наиболее сложные части вашего кода *юнит-тестами*.