Семинарское задание №3
Page Rank

Мосиенко Константин Викторович

2019

В данном задании вам необходимо будет вычислить *ранг господина Пейджа* для страниц симпл википедии. Тут вам пригодятся результаты выполнения первого задания - ссылочный граф можно получить из скачанных документов. Вот только необходимо обратить внимание на качество полученной коллекции:

- Удалите ссылки вида:
  - /wiki/Help:, /wiki/Help_talk:
  - /wiki/File:, /wiki/Media:, /wiki/MediaWiki:, /wiki/MediaWiki_talk:
  - /wiki/Module:, /wiki/Talk:
  - /wiki/Category:, /wiki/Category_talk:
  - /wiki/User:, /wiki/User_talk:, /wiki/Special:
  - /wiki/Template:, /wiki/Template_talk:
  - /wiki/Wikipedia:, /wiki/Wikipedia_talk:
- Убедитесь, что коллекция содержит достаточное количество статей. Меньше 100000 точно не стоит брать.

Если не получается привести коллекцию к требуемому виду - лучше возьмите её у товарища. Вычислите pagerank (с нормализацией на максимальное значение) и отсортируйте все документы по его убыванию. Мне присылайте код и список документов с числами.