Connectrix

Xiangqing Zhang, Tianjiao Mo, Zhihao Xue

09/17/2015

Data Set:
Common Crawl (Project Ideas #5)

Description:
We decide to utilize data from Common Crawl (and if possible, crawl additional web data on our own) to study how well websites living in Internet is connected. The core feature this project provides is to show the relationship between a certain number of websites. Given a list of website domains by user, we will find out how they are connected by examing hyperlinks, embedded iframes, in-html javascript redirects and pure texts in web pages. We will also take the webpage's location into account (e.g. webpage's depth from root path and so on).
We will build a simple website that hosts this project. Therefore, frontend will include HTML, CSS and JavaScript. We haven't decided our backend yet, but it should be a decision between Java and Node.js. Above descriptions may change during Milestone 1 period.