

## Information Retrieval – Übung Crawling & Link Analysis

Bitte bearbeiten Sie die Aufgaben vor den Übungsterminen. In den Online-Sitzungen arbeiten wir dann mit den Ergebnissen Ihrer Vorarbeiten. Z.B. stellt jemand aus Ihren Reihen die vorbereitete Lösung vor. Die Lösung kann aber auch im Plenum diskutiert, erweitert und in Kontext gesetzt werden. Auch Peer-Reviews sind möglich.

Die Präsentation der Lösung kann beispielsweise über "Teilen des Bildschirms", Kurzreferate oder interaktive Kollaborationswerkzeuge wie Google Jamboards erfolgen.

### I Wayback-Machine

- Was ist die Wayback Machine? Informieren Sie sich selbstständig auf <https://archive.org/web/>. Erklären Sie das zugrundeliegende Prinzip.
- Wann wurde die Website <http://www.fbi.fh-koeln.de/> das erste Mal gecrawlt? Wie oft wurde diese Website überarbeitet?
- Wann wurde <http://www.facebook.com> das erste Mal gecrawlt? Was fällt sonst noch auf?

### II Grundlagen Web-Crawler

- Welche Vorteile hat es den Crawling-Vorgang zu parallelisieren und zu verteilen?
- Nennen Sie mind. vier typische Probleme beim Crawling.
- Was ist eine robots.txt Datei und was kann sie bewirken? (hierzu finden Sie Hilfestellung unter <https://wiki.selfhtml.org/wiki/Grundlagen/Robots.txt>):
- Welche Probleme könnten bei der Informationssuche entstehen, wenn ein Suchmaschinenbetreiber seinen Index zu selten aktualisiert?
- Angenommen Sie müssten einem Crawler konfigurieren: auf welche Eigenschaften würden Sie dabei wertlegen?

### III Exploratives Verständnis des PageRank

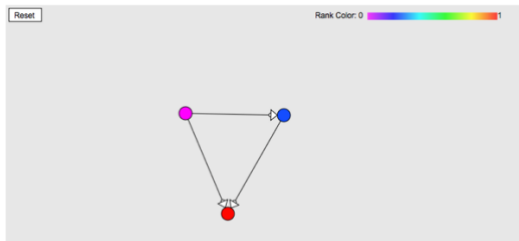
Besuchen Sie die folgende Webseite und erstellen Sie eigene Netzwerke aus 3, 5 und 7 Knoten (Punkten). Wie müssen Sie diese Knoten miteinander verbinden, damit jeweils 1, 2 oder 3 Knoten den höchsten (relativen) PageRank von allen Knoten hat?

<http://www.graui.de/pageRank.htm>

Sie suchen also eine Kombination für jede Zelle der folgenden Tabelle (mehrere Lösungen möglich, wobei Sie natürlich jeden Knoten zum Netzwerk hinzufügen):

	3 Knoten gesamt	5 Knoten gesamt	7 Knoten gesamt
1 Knoten rot	x		
2 Knoten rot			
3 Knoten rot			

Eine mögliche Lösung für die erste mögliche Kombination ist als Beispiel angegeben (der rote Knoten ist derjenige mit dem höchsten PageRank):



## IV PageRank selbst berechnen mit Gephi (Teamarbeit für 2)

Gephi ist ein sehr mächtiges Tool zur Netzwerkanalyse, das unter einer freien Lizenz zur Verfügung gestellt wird. Neben der Visualisierung von Netzwerken, kann es auch allerhand Statistiken auf Netzwerken berechnen – u.a. auch den PageRank.

Bearbeiten Sie das Tutorial unter

<https://www.websiteboosting.com/magazin/44/berechnen-sie-ihren-pagerank-doch-selbst.html>

und bereiten Sie ein eigenes Beispielprojekt mit einer kleinen Webseite vor. Die Link-Daten crawlen Sie mit dem Tool Screaming Frog (<http://www.screamingfrog.co.uk>).

Präsentieren Sie Ihre Lösung und berichten Sie über Ihre Erkenntnisse.