Department of Computer Science
The University of Melbourne
COMP90042 WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2017)

Workshop exercises: Week 11

Discussion

1. What are the two components in the **PageRank model** of link analysis? What are the resulting weights used for?
   (a) Why do we typically use "eigenvalue methods" to calculate PageRank weights?
   (b) Given a collection of two documents, where one document contains a link to the other document, find the equilibrium PageRank weights when $\alpha = 0.5$.
   (c) How is the **HITS model** — as described in the lectures — similar to PageRank and how is it different?
2. How can we compress a **postings list** in an inverted index?
   (a) What is **Variable Byte Compression** and how does it compress an integer?

Programming

1. Work on the project! :-)

Catch-up

- How might we construct a **document ranking** for an IR query?
- How do we frame building an information retrieval engine based on a probabilistic (language) model? What does it mean that we "assume a uniform prior for $P(r|d)$?
- What is a **hyperlink**?
- What is a **Markov chain**? What is a matrix **eigenvalue**?
- What is an **inverted index**, and why is it useful?
- What does **entropy** measure? Read up on **Shannon's source coding theorem**.

Get ahead

- PageRank isn't the first eigenvalue method that we've seen in this subject. Contemplate some of its similarities and differences to other situations where we have employed eigenvalues.
- Implement the `numpy` PageRank solver given in the lectures for an arbitary graph, and α.
  - Confirm that the equilibrium values for the Discussion question are as expected.
  - How does changing the value of α affect the resulting weights?
  - Change the solver so that it is finding HITS weights instead.