CSC 9010: Search Engines Google

Dr. Paula Matuszek

Paula_A_Matuszek@glaxosmithkline.com

(610) 270-6851

Search Engine Basics

- A spider or crawler starts at a web page, identifies all links on it, and follows them to new web pages.
- A parser processes each web page and extracts individual words.
- An indexer creates/updates a hash table which connects words with documents
- A searcher uses the hash table to retrieve documents based on words
- A ranking system decides the order in which to present the documents: their *relevance*

Selecting Relevant Documents

- Assume:
  - we already have a corpus of documents defined.
  - goal is to return a subset of those documents.
  - Individual documents have been separated into individual files
- Remaining components must parse, index, find, and rank documents.
- Traditional approach is based on the words in the documents (predates the web)