Web Crawling, Content Processing and Zipf’s Law

Submit by Jiaoni Zhou and Hanzi Li

1. Main Components of the programs
2. Design/development/architectural choices.
   1. Discussion on how noise reduction was performed
   2. Discussion on how noise reduction performance was evaluated

Precision and recall

* 1. Discussion on how well the noise reduction technique worked based on the chosen evaluation
  2. Any challenges faced during the development of the crawler and content processor

1. Word frequency/rank plots for 3 different crawls
2. Discussions whether the 3 word distributions follow Zipf’s law or not
3. An appendix containing details on these 3 crawls (e.g. seeds, domains), as well as a list of the 100 most frequent words for each crawl