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.

## Discussion on how noise reduction was performed

Our noise reduction technique is mainly based on the slope optimization algorithm which is reported by Finn *et al* [1]. We treat the html file as consisting of two kinds of tokens: word token (text) and tag token (<>). First, we tokenize the file. For *i*th token, if it is a word, mark it as *Bi* = 0; if it is a tag token, mark it as *Bi* = 1. Then, we optimize the following object:

where i, j represents the work token start position and end position respectively.

## Discussion on how noise reduction performance was evaluated

For each seed, we randomly choose 10 crawled websites. For each website, we compare the retrieved content with the actual main content. The retrieved content We manually decide which part of the w We use precision, recall and F-Score to quantitatively evaluate the performance of noise reduction:

Take the website “<https://www.scu.edu/recreation/>” as an example. the re

## Discussion on how well the noise reduction technique worked based on the chosen evaluation

Table 1. Precision, Recall and F-Score of Noise Removal Technique

Table 1 shows the performance of our noise removal technique,

https://www.scu.edu/recreation/

A kickboxing class at SCU. The beautiful Sullivan Aquatic Center. Degheri;Tennis Courts. Bronco Kidz All Sports Camp - Girls' Flag Football. Men's Rugby. Previous Next

## Any challenges faced during the development of the crawler and content processor

### Crawler

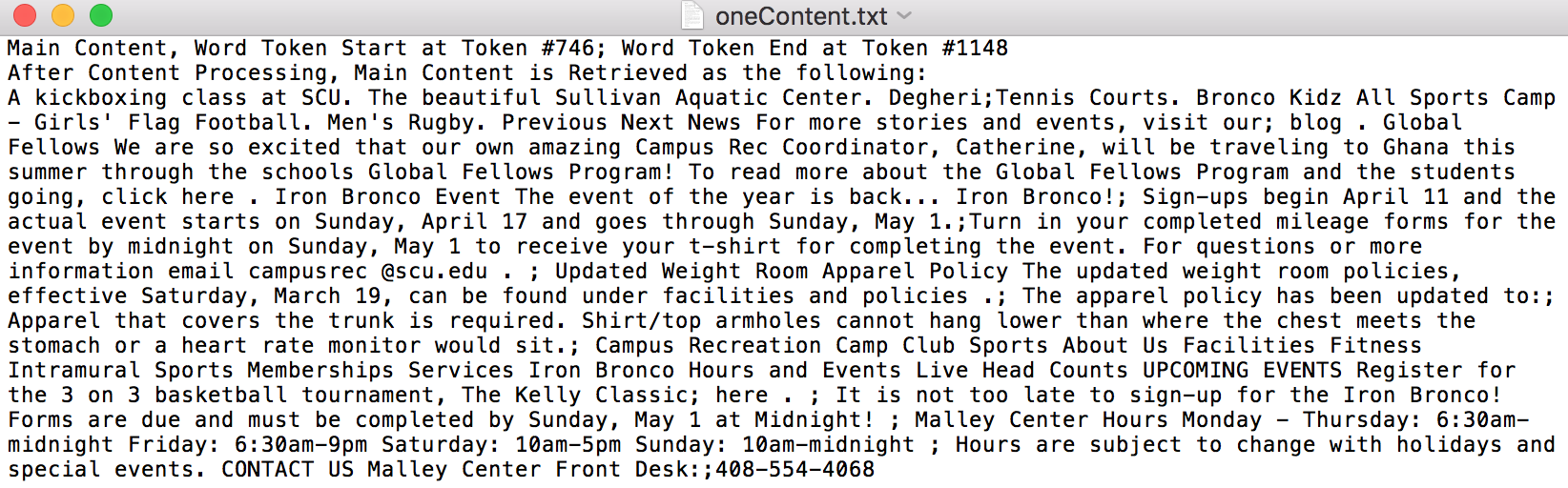
### Challenges faced during the content processor

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

Surabhi Lingwal. Article: Noise Reduction and Content Retrieval from Web Pages. *International Journal of Computer Applications*73(4):24-30, July 2013.

International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013 115 Elimination of Noisy Information from Web Pages Alpa K. Oza, Shailendra Mishra

International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 1, April 2014 – www.ijais.org 37 Identifying Informative Web Content Blocks using Web Page Segmentation Stevina Dias M. E Student TSEC, Mumbai, India Jayant Gadge



Appendix 1: Screenshot of the content process result from website “https://www.scu.edu/recreation/”