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 . We use precis

## Discussion on how noise reduction performance was evaluated

We use precision, recall and F-Score to quantitatively evaluate the performance of noise reduction:

For example, the website “https://www.scu.edu/recreation/”

## 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