Introduction to Digital Libraries Assignment #4

James Tate II

April 23, 2015

1   Introduction

Assignment #4 required t was compared and jusText's performance was analyzed.

2   Methodology

I wrote two scripts forthis utility, both of which are available in my git repository on GitHub[1].

2.1   jusText

Listing 2.1: Generating list of files in tweets directory

```
find ./tweets/ > tweets_file_list
```

Table 2.1: Word Count Data

Data Point	Original	After jusText
Total bytes	646,619,835	12,133,748
Total words	33,594,568	2,035,935
Unique words	9,135,191	121,422
Total letter words	81,318,873	2,030,636
Unique letter words	1,271,593	57,434

References

---

[1] https://github.com/jamesbtate/cs851-s15