Information Retrieval

María Carrasco Rodríguez (16874129)
Fabian Lindenberg (74076658)
Lea Voget (45869178)

Assignment 5

University of California Irvine

February 27, 2011

1

Quantification of the Dataset

- The total of all unfiltered and unprocessed words in the document was 1456106
- Each book had the following number of words: Emma: 165075, Anna Karenina: 361786, Jane Eyre: 192539, Moby Dick: 221810, Portrait of a Lady: 246057, Pride and Prejudice: 125969
- The word "love" appeared 983 times in the dataset and "adventure" occurred 19 times.
- "Anna Karenina" and "Three men in a boat" did not contain the word adventure
- The top 5 words in each book were:

  Anna Karenina:

  - levin 1629
  - vronski 865
  - anna 825
  - well 675
  - kitti 672

  Emma:

  - emma 867
  - harriet 506
  - thing 462
  - weston 448
  - elton 408

  Emma:

  - emma 867

- harriet 506
- thing 462
- weston 448
- elton 408

•

Total words of emma.txt is The word love occured 117 times. The word adventure occured 2 times. the top 5 words are:

Total words of annaKarenina.txt is The word love occured 433 times. The word adventure occured 0 times. the top 5 words are:

Total words of janeEyre.txt is The word love occured 151 times. The word adventure occured 3 times. the top 5 words are: rochest 371 jane 348 sir 315 dai 308 well 348

Total words of mobyDick.txt is 0 The word love occured 24 times. The word adventure occured 5 times. the top 5 words are: ahab 512 ship 623 sea 542 whale 1629 man 540

Total words of portraitOfALady.txt is The word love occured 146 times. The word adventure occured 7 times. the top 5 words are: isabel 1490 don 833 ralph 578 ve 683 osmond 588

Total words of prideAndprejudice.txt is The word love occured 92 times. The word adventure occured 2 times. the top 5 words are: sister 294 bennet 333 binglei 311 –¿ in Book is Bingley darci 417 elizabeth 635

Total words of threeMenInABoat.txt is 71435 The word love occured 10 times. The word adventure occured 0 times. the top 5 words are: work 164 georg 263 boat 231 harri 275 river 162

The word love occured the top 5 words are: time 2621 thought 2166 good 2283 well 2423 man 2280

Bibliography

[1] www.austen.com/

[2] http://librivox.org/

[3] http://en.wikipedia.org/

[4]

[5]

[6]

[7]

[8]

[9]