Information Retrieval

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

Assignment 5

University of California Irvine

February 27, 2011

2

Quantification of the Dataset

1. (a) The total of all unfiltered and unprocessed words in the document was 1456106.

   (b) 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

   (c) The word "love" appeared 983 times in the whole corpus and "adventure" occurred 19 times.

   The number of occurences of these words per book are:

   - Anna Karenina:
     - love: 433
     - adventure: 0
   - Emma:
     - love: 117
     - adventure: 2
   - Jane Eyre:
     - love: 151
     - adventure: 3
   - Moby Dick:
     - love: 24
     - adventure: 5
   - Portrait of a Lady:
     - love: 146
     - adventure: 7
   - Pride and Prejudice:
     - love: 92
     - adventure: 2
   - Three Men in a Boat:
     - love: 10

    - adventure: 0

(d) The books *Anna Karenina* and *Three men in a boat* did not contain the word “adventure”

(e) The top 5 (stemmed) 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
- Jane Eyre:
    - rochest: 371
    - jane: 348
    - well: 348
    - sir: 315
    - dai: 308
- Moby Dick:
    - whale: 1629
    - ship: 623
    - sea: 542
    - man: 540
    - ahab: 512
- Portrait of a Lady:
    - isabel: 1490
    - don: 833
    - ve: 683
    - osmond: 588
    - ralph: 578
- Pride and Prejudice:
    - elizabeth: 635
    - darci: 417
    - bennet: 333

  - binglei: 311
  - sister: 294
- Three Men in a Boat:
  - harri: 275
  - georg: 263
  - boat: 231
  - work: 164
  - river: 162

2. In order to quantify the books, we wrote a quantifier. The quantifier removes the header in front of each book and then tokenizes the content of the book. Counting the number of tokens, the total number of unprocessed words is calculated. The quantifier has a hashmap that maps all words to their number of occurences Then all tokens are stemmed and it is checked if they are stop words or not. If they are not stop words, they are either added to the `HashMap` or, if already present, their respective value is increased by one. Finally, the quantifier searches for the five highest values and emits the five appropriate key-value pairs.

   In order to find the number of occurences of “love” and “adventure”, we did not use the stemmed version. Each token was checked if it was “love” or “adventure” just before stemming it.

   The total number of words and of occurences of “love” and “adventure” in all books was found through giving the quantifier all book files at the same time.

   We found out that some books did not contain “love” or “adventure”, when the quantifier emitted that one of the words was found 0 times.

3. We used the online tool *Wordle* (`www.wordle.net`) to generate the word clouds of each book. We chose the following settings:

   - case insensitivity (i.e. all words were transformed to their lowercase representation)
   - remove common English words (according to a stop word list built into the tool)
   - align the words mostly horizontally (for higher readability)
   - black font color only (instead of using multiple different colors without meaning)

   The following figures show the resulting word clouds.

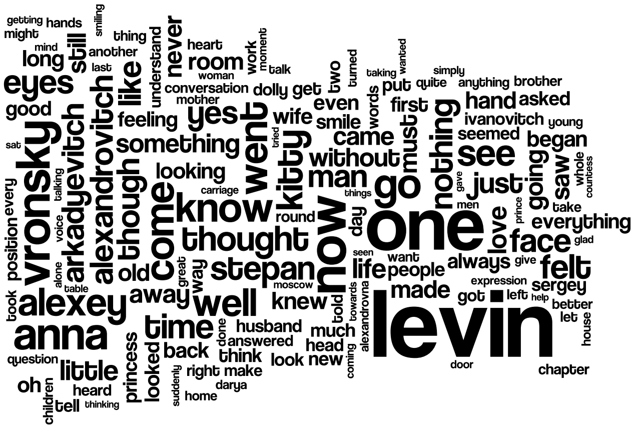

Figure 2.1: Word cloud for *Anna Karenina*

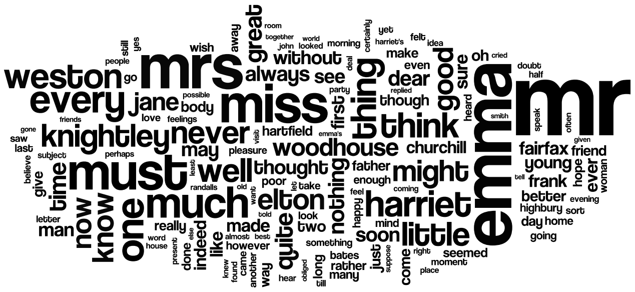

Figure 2.2: Word cloud for *Emma*

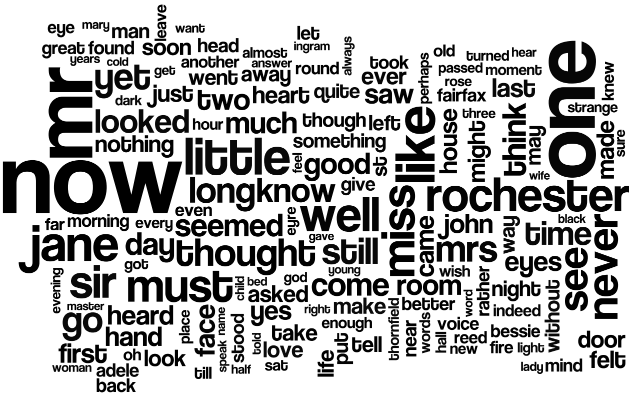

Figure 2.3: Word cloud for *Jane Eyre*

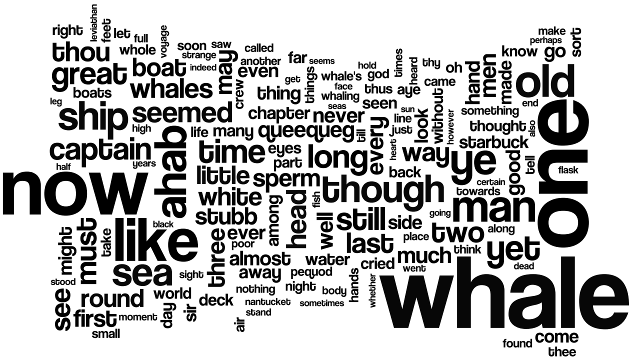

Figure 2.4: Word cloud for *Moby Dick*

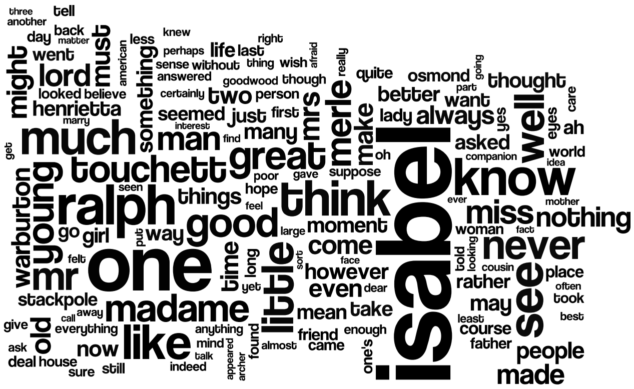

Figure 2.5: Word cloud for *Portrait of a Lady Volume 1*

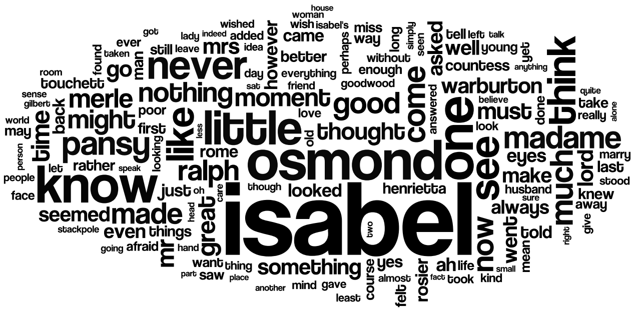

Figure 2.6: Word cloud for *Portrait of a Lady Volume 2*

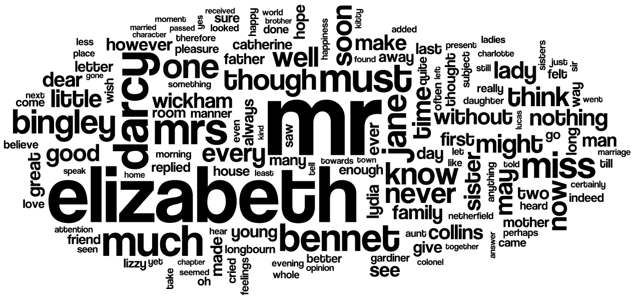

Figure 2.7: Word cloud for *Pride and Prejudice*

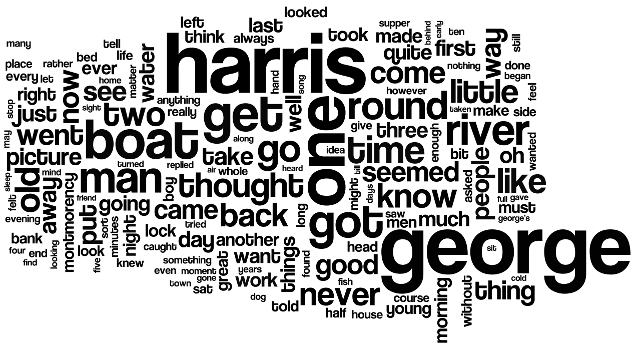

Figure 2.8: Word cloud for *Three Men in a Boat*