Information Retrieval

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

Assignment 4

University of California Irvine

February 13, 2011

2

Quantify the data

1. *Comment:* Unfortunately, our installed Antivirus software[1] reported two of the files contained in the archive as being infected by viruses and put them under quarantine. Consequently, the following results may vary slightly from the true numbers.

   a) Number of people targeted in the Enron data set: 150

   b) Number of individual data files: $517,428$

   c) Number of sent data files: $128,462$

   d) We consider Inboxes all folders that do not contain sent e-mails. According to this definition, the number of data files in inboxes is, thus, the total number of files minus the number of sent data files: $517,428 - 128,462 = 388,966$

   e) The ten persons with the largest number of files are in descending order "kaminski-v", "dasovich-j", "kean-s", "mann-k", "jones-t", "shackleton-s", "taylor-m", "farmer-d", "germany-c", and "beck-s".

2. To quantify the data, we wrote a Java program that iterated over the files and subfolders. For subtask 1a), the number of folders in the first hierarchy level were counted. To count all individual data files in subtask 1b), the program scanned *recursively* through all subfolders. For subtask 1c), only files in folders whose names matched the *regular expression* `(.*[^A-Za-z])?sent([^A-Za-z].*)?` were counted. To solve subtask 1e), we maintained a sorted mapping between number of files (key) and a list of folder names (value), and returned the first ten values. If multiple folders of equal size were ranked tenth, all these folders would be returned.

[1] Avira Antivir Personal 10