Solution BI-1c

1a) For this exercize we used english book "Time Machine by H.G Wells", which will be called Book 1, and " Berlin — Panorama einer Weltstadt by Karl Gutzkow" which will be called Book 2. Distributions

of single letters are shown in Tables 1 and 2.

|  |  |  |
| --- | --- | --- |
| Letter | Frequency | Distribution |
| a | 12701 | 8,15% |
| b | 2171 | 1,39% |
| c | 4022 | 2,58% |
| d | 6859 | 4,40% |
| e | 19666 | 12,62% |
| f | 3735 | 2,40% |
| g | 3491 | 2,24% |
| h | 8786 | 5,64% |
| i | 11254 | 7,22% |
| j | 185 | 0,12% |
| k | 1219 | 0,78% |
| l | 6629 | 4,25% |
| m | 4411 | 2,83% |
| n | 10943 | 7,02% |
| o | 11080 | 7,11% |
| p | 2870 | 1,84% |
| q | 106 | 0,07% |
| r | 8832 | 5,67% |
| s | 9243 | 5,93% |
| t | 15040 | 9,65% |
| u | 4312 | 2,77% |
| v | 1407 | 0,90% |
| w | 3496 | 2,24% |
| x | 264 | 0,17% |
| y | 3001 | 1,93% |
| z | 146 | 0,09% |
| TOTAL |  |  |

Table 1: frequencies in Book 1

|  |  |  |
| --- | --- | --- |
| Letter | Frequency | Distribution |
| a | 13505 | 5,08% |
| b | 5038 | 1,89% |
| c | 8136 | 3,06% |
| d | 13451 | 5,05% |
| e | 45317 | 17,03% |
| f | 4251 | 1,60% |
| g | 8205 | 3,08% |
| h | 12444 | 4,68% |
| i | 21566 | 8,10% |
| j | 641 | 0,24% |
| k | 3379 | 1,27% |
| l | 9981 | 3,75% |
| m | 6473 | 2,43% |
| n | 28140 | 10,58% |
| o | 7135 | 2,68% |
| p | 2289 | 0,86% |
| q | 75 | 0,03% |
| r | 19127 | 7,19% |
| s | 16257 | 6,11% |
| t | 192 | 6,09% |
| u | 9539 | 3,58% |
| v | 2267 | 0,85% |
| w | 4088 | 1,54% |
| x | 89 | 0,03% |
| y | 451 | 0,17% |
| z | 2805 | 1,05% |
| ß | 1113 | 0,42% |
| à | 4 | 0,00% |
| â | 3 | 0,00% |
| ä | 1474 | 0,55% |
| ç | 1 | 0,00% |
| è | 9 | 0,00% |
| é | 13 | 0,00% |
| ö | 807 | 0,30% |
| ü | 1831 | 0,69% |
| TOTAL | 266096 | 100,00% |

Table 2: frequencies in Book 2

1b) In this task we analysed the top 20 character combinations in Books 1 and 2:

|  |  |  |
| --- | --- | --- |
| Combination | Frequency | Distribution |
| th | 4949 | 4,12% |
| he | 4084 | 3,40% |
| in | 3180 | 2,65% |
| an | 2497 | 2,08% |
| er | 2417 | 2,01% |
| re | 2210 | 1,84% |
| nd | 2101 | 1,75% |
| ed | 1743 | 1,45% |
| at | 1732 | 1,44% |
| on | 1669 | 1,39% |
| en | 1665 | 1,39% |
| it | 1509 | 1,26% |
| ha | 1495 | 1,25% |
| ng | 1476 | 1,23% |
| me | 1409 | 1,17% |
| st | 1357 | 1,13% |
| ou | 1347 | 1,12% |
| of | 1346 | 1,12% |
| or | 1339 | 1,12% |
| es | 1296 | 1,08% |
| TOTAL |  | 34,01% |

Table 3: Combination frequency in book 1

|  |  |  |
| --- | --- | --- |
| Combination | Frequency | Distribution |
| en | 10795 | 4,94% |
| er | 9098 | 4,16% |
| ch | 7015 | 3,21% |
| de | 5758 | 2,63% |
| in | 5023 | 2,30% |
| ei | 4797 | 2,19% |
| ie | 4642 | 2,12% |
| ge | 4200 | 1,92% |
| te | 4198 | 1,92% |
| nd | 3424 | 1,57% |
| un | 3296 | 1,51% |
| he | 3223 | 1,47% |
| ne | 3080 | 1,41% |
| st | 3040 | 1,39% |
| es | 3007 | 1,37% |
| be | 2873 | 1,31% |
| an | 2729 | 1,25% |
| di | 2725 | 1,25% |
| ic | 2721 | 1,24% |
| re | 2678 | 1,22% |
| TOTAL |  | 40,38% |

Table 4: Combination frequency in book 2

Using two more books for each language shows extreme correlation between distributions of specific combinations. They do not take exact place every time, however, we could construct statistical model for each language and run statistical tests to see if the text conforms to this model. It would be possible to implement algorithm that differentiates between german and english language. Main issue is then the size and the type of given text input. If text is too small there will be extreme variations of the distributions based on the content. On the same note, if the input is big but contains really specifc content (eg. simple tweets), distributions will be biased toward the specific words. Source code is commented where apropriate.

1.c) For this excercize, Map method is called for every line in input file. Reduce method is called for every 2-letter combination that is found. Most time in process is spent on file operations, since Hadoop operates on 128 MB file chunks. When this set-up time is negligible in regard to the actual computation we expect that the speedup will be linear in regard to number of machines in cluster.