СПИСОК ФАЙЛОВ

Calgary Corpus

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Size (bytes)** | **File name** | **Description** |
| 1 | 111,261 | BIB | ASCII text in UNIX "refer" format - 725 bibliographic references. |
| 2 | 768,771 | BOOK1 | unformatted ASCII text - Thomas Hardy: Far from the Madding Crowd. |
| 3 | 610,856 | BOOK2 | ASCII text in UNIX "troff" format - Witten: Principles of Computer Speech. |
| 4 | 102,400 | GEO | 32 bit numbers in IBM floating point format - seismic data. |
| 5 | 377,109 | NEWS | ASCII text - USENET batch file on a variety of topics. |
| 6 | 21,504 | OBJ1 | VAX executable program - compilation of PROGP. |
| 7 | 246,814 | OBJ2 | Macintosh executable program - "Knowledge Support System". |
| 8 | 53,161 | PAPER1 | UNIX "troff" format - Witten, Neal, Cleary: Arithmetic Coding for Data Compression. |
| 9 | 82,199 | PAPER2 | UNIX "troff" format - Witten: Computer (in)security. |
| 10 | 513,216 | PIC | 1728 x 2376 bitmap image (MSB first): text in French and line diagrams. |
| 11 | 39,611 | PROGC | Source code in C - UNIX compress v4.0. |
| 12 | 71,646 | PROGL | Source code in Lisp - system software. |
| 13 | 49,379 | PROGP | Source code in Pascal - program to evaluate PPM compression. |
| 14 | 93,695 | TRANS | ASCII and control characters - transcript of a terminal session. |

Cantebury Corpus

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Size (bytes)** | **File name** | **Description** |
| 15 | 152,089 | alice29.txt | English text |
| 16 | 125,179 | asyoulik.txt | Shakespeare |
| 17 | 24,603 | cp.html | HTML source |
| 18 | 11,150 | fields.c | C source |
| 19 | 3,721 | grammar.lsp | LISP source |
| 20 | 1,029,744 | kennedy.xls | Excel Spreadsheet |
| 21 | 426,754 | lcet10.txt | Technical writing |
| 22 | 481,861 | plrabn12.txt | Poetry |
| 23 | 513,216 | ptt5 | CCITT test set |
| 24 | 38,240 | sum | SPARC Executable |
| 25 | 4,227 | xargs.1 | GNU manual page |