Results from Project Gutenberg

Analysis of 9,481 English works (3.98 GiB) from Project Gutenberg (the extracted contents of the 2003 PG DVD, plain text files only, minus the human genome project, non-English works, and duplicates in 7-bit-clean encoding), after stripping off the common boilerplate text present in every file so as not to skew results, yielded the following frequencies of letters, bigrams, trigrams, and quadrigrams:

Letters

Of 3,104,375,038 letters scanned:

1. e (390395169, 12.575645%)
2. t (282039486, 9.085226%)
3. a (248362256, 8.000395%)
4. o (235661502, 7.591270%)
5. i (214822972, 6.920007%)
6. n (214319386, 6.903785%)
7. s (196844692, 6.340880%)
8. h (193607737, 6.236609%)
9. r (184990759, 5.959034%)
10. d (134044565, 4.317924%)
11. l (125951672, 4.057231%)
12. u (88219598, 2.841783%)
13. c (79962026, 2.575785%)
14. m (79502870, 2.560994%)
15. f (72967175, 2.350463%)
16. w (69069021, 2.224893%)
17. g (61549736, 1.982677%)
18. y (59010696, 1.900888%)
19. p (55746578, 1.795742%)
20. b (47673928, 1.535701%)
21. v (30476191, 0.981717%)
22. k (22969448, 0.739906%)
23. x (5574077, 0.179556%)
24. j (4507165, 0.145188%)
25. q (3649838, 0.117571%)
26. z (2456495, 0.079130%)

Bigrams

Of 2,383,373,483 bigrams scanned:

1. th (92535489, 3.882543%)
2. he (87741289, 3.681391%)
3. in (54433847, 2.283899%)
4. er (51910883, 2.178042%)
5. an (51015163, 2.140460%)
6. re (41694599, 1.749394%)
7. nd (37466077, 1.571977%)
8. on (33802063, 1.418244%)
9. en (32967758, 1.383239%)
10. at (31830493, 1.335523%)
11. ou (30637892, 1.285484%)
12. ed (30406590, 1.275779%)
13. ha (30381856, 1.274742%)
14. to (27877259, 1.169655%)
15. or (27434858, 1.151094%)
16. it (27048699, 1.134891%)
17. is (26452510, 1.109877%)
18. hi (26033632, 1.092302%)
19. es (26033602, 1.092301%)
20. ng (25106109, 1.053385%)

Trigrams

Of 1,699,542,842 trigrams scanned:

1. the (59623899, 3.508232%)
2. and (27088636, 1.593878%)
3. ing (19494469, 1.147042%)
4. her (13977786, 0.822444%)
5. hat (11059185, 0.650715%)
6. his (10141992, 0.596748%)
7. tha (10088372, 0.593593%)
8. ere (9527535, 0.560594%)
9. for (9438784, 0.555372%)
10. ent (9020688, 0.530771%)
11. ion (8607405, 0.506454%)
12. ter (7836576, 0.461099%)
13. was (7826182, 0.460487%)
14. you (7430619, 0.437213%)

15. ith (7329285, 0.431250%)
16. ver (7320472, 0.430732%)
17. all (7184955, 0.422758%)
18. wit (6752112, 0.397290%)
19. thi (6709729, 0.394796%)
20. tio (6425262, 0.378058%)

Quadrigrams

Of 1,144,085,293 quadrigrams scanned:

1. that (8709261, 0.761242%)
2. ther (6916008, 0.604501%)
3. with (6565513, 0.573866%)
4. tion (6314428, 0.551919%)
5. here (4285164, 0.374549%)
6. ould (4232202, 0.369920%)
7. ight (3540253, 0.309440%)
8. have (3324067, 0.290544%)
9. hich (3252540, 0.284292%)
10. whic (3247213, 0.283826%)
11. this (3161481, 0.276333%)
12. thin (3093756, 0.270413%)
13. they (3002324, 0.262421%)
14. atio (3001919, 0.262386%)
15. ever (2982572, 0.260695%)
16. from (2958372, 0.258580%)
17. ough (2899649, 0.253447%)
18. were (2643859, 0.231089%)
19. hing (2630750, 0.229944%)
20. ment (2555284, 0.223347%)