Statistics for Bioinformatics: Project 1

Title: Simple Model for Finding over and underrepresented Words.

Name: Prashant Kumar Kuntala

**OVERVIEW**

In this Project, we applied two different models of similarity and exceptionality that was introduced in exercise 2 to a set of English language books each with all whitespace and special characters removed. Provided with a simple enumeration tool that allows you to extract all words between 3 and 10 characters in length that occur in the text.

The English books that were used:

1. Alice in Wonderland

2. Big Dummy’s Guide to the Internet

3. California Mexican Spanish Cook Book

4. Complete Book of Cheese

5. Einstein Relativity Theory

6. Hamlet Prince of Denmark

7. Hound of Baskervilles

8. Introduction to Wood Carving

**EXCEPTIONALITY OF WORDS WITHIN A SINGLE DNA SEQUENCE**

Applied a simple model for finding exceptional words. Frequencies for each letter A through Z was calculated within in the entire book and then letter probabilities is given as:

P(Letter) = letter frequency / total No. of letters in the book

Used Alice in wonderland book for this single DNA sequence experiment. Once you have the letter probabilities, we calculate the individual word probabilities as below:

P(Word) = P(w1)\*P(w2)\*…P(wk)

Here, w1 is the letter at position 1 in the word of length k. and so on.

Based on the exercise 2 model, the random variable is binomial. Hence the Expected value is calculated as below:

E[x] = (n-k+1)\*P(word)

Where, n is the length of the book, k is the word length and probability of word.

We calculate the O/E score using the below formula:

O/E score = Word frequency / E[x]

**TOP overrepresented words in Alice in wonderland**

**Word WordCount SeqCount Word Probbility Expected value O/E score**

**THE 2589 1 0.000798559010074 97.807507553853370 26.470360657892051**

**ING 1071 1 0.000109837994084 111.260465069260221 9.626060787480053**

**AND 1030 1 0.000231981489743 139.673557932947915 7.374337814853007**

**HER 792 1 0.000431660619691 192.543350632727822 4.113359393598185**

**SHE 642 1 0.000475338999499 250.762871291326007 2.560187625440573**

**DTH 613 1 0.000283833665767 285.526818674415040 2.146908661140519**

**YOU 583 1 0.000052402985727 291.945136366273800 1.996950547819948**

**HAT 557 1 0.000508941327986 354.280270218050987 1.572201578307423**

**ICE 513 1 0.000213894886901 380.478115965655491 1.348303564576909**

**DTHE 511 1 0.000035513679084 384.827795866169936 1.327866660072830**

**THA 490 1 0.000508941327986 447.162929717947122 1.095797454205501**

**AID 484 1 0.000249006591586 477.661257055405201 1.013270372781897**

**ALI 473 1 0.000236892263191 506.675821450982198 0.933535763844930**

**THI 470 1 0.000448130994234 561.562905624759651 0.836949868469513**

**SAI 462 1 0.000330835451314 602.083631701716513 0.767335259877790**

**SAID 462 1 0.000014712981632 603.885662979061408 0.765045485135187**

**ETH 461 1 0.000798559010074 701.693170532914792 0.656982309874677**

**ERE 438 1 0.000840025269431 804.579465532803965 0.544383766630124**

**LIC 435 1 0.000072326479864 813.438012786577929 0.534767238759631**

**LICE 415 1 0.000009049593847 814.546397991309959 0.509486017031564**

**THER 405 1 0.000042848521928 819.794442108471912 0.494026281708326**

**ALIC 399 1 0.000005767529077 820.500843302247972 0.486288348460625**

**ALICE 399 1 0.000000721641586 820.589228520422466 0.486235970607881**

**ENT 393 1 0.000812451401741 920.098276205688876 0.427128286361602**

**WAS 378 1 0.000112984153548 933.936575332203120 0.404738405137998**

**VER 371 1 0.000052457042926 940.361513949814480 0.394529119382697**

**TTH 365 1 0.000633532166034 1017.956533645719787 0.358561478742894**

**NTH 360 1 0.000417491340210 1069.090872994592473 0.336734705246914**

**ITH 349 1 0.000448130994234 1123.977957168369812 0.310504309959275**

**TER 349 1 0.000666429182756 1205.602203472360316 0.289481886309443**

**THAT 336 1 0.000050519743190 1211.789811098532937 0.277275808826453**

**ALL 335 1 0.000142741361378 1229.272773040146831 0.272518848010847**

**OTH 320 1 0.000491901928555 1289.520921249502408 0.248154174722447**

**OUT 320 1 0.000247231470885 1319.801831803487175 0.242460642415328**

**HES 314 1 0.000475338999499 1378.021352462085360 0.227862942354980**

**HIN 311 1 0.000295313828409 1414.191390165680104 0.219913656781325**

**ERS 310 1 0.000396689257740 1462.777890453719010 0.211925543873134**

**HIS 304 1 0.000266748149800 1495.449203841252029 0.203283400879908**

**FOR 303 1 0.000080258781188 1505.279299361200174 0.201291547773616**

**NTHE 293 1 0.000052237120767 1511.677249675670964 0.193824442395268**

**WIT 287 1 0.000167131521278 1532.147518401814978 0.187318777436895**

**IDT 283 1 0.000309964384202 1570.111956178902574 0.180241924078282**

**HEM 281 1 0.000161511482626 1589.893882570984033 0.176741355558649**

**INT 274 1 0.000455927050896 1645.735827764673559 0.166490876225355**

**TTHE 270 1 0.000079268461594 1655.444549672228959 0.163098184142416**

**FTH 269 1 0.000123861322476 1670.615084449097367 0.161018538922570**

**AIDT 269 1 0.000024717483859 1673.642457154632666 0.160727280100989**

**WITH 266 1 0.000010745835926 1674.958596393026028 0.158809895702988**

**TOP UNDERATED WORDS**

**STRIBUTEA 1 1 0.000000000010024 114217.123578793994966 0.000008755254630**

**TDOUBL 1 1 0.000000006593289 114217.124386320210760 0.000008755254568**

**SLUCKILYT 1 1 0.000000000000128 114217.124386335897725 0.000008755254568**

**NOBODYI 1 1 0.000000000361846 114217.124430653348099 0.000008755254564**

**DPRI 1 1 0.000002683968968 114217.453160488614230 0.000008755229366**

**STRIBUTEC 1 1 0.000000000003061 114217.453160863442463 0.000008755229366**

**DALICEJUM 1 1 0.000000000000039 114217.453160868259147 0.000008755229366**

**IGHTFOOTE 1 1 0.000000000015344 114217.453162747537135 0.000008755229366**

**NOBODYM 1 1 0.000000000103463 114217.453175419243053 0.000008755229365**

**STRIBUTEI 1 1 0.000000000008826 114217.453176500246627 0.000008755229365**

**EIRHEADSSH 1 1 0.000000000003019 114217.453176869981689 0.000008755229365**

**GHHETAU 1 1 0.000000003164019 114217.453564386320068 0.000008755229335**

**ITTENO 1 1 0.000000436439256 114217.507018157033599 0.000008755225237**

**ROOMNO 1 1 0.000000032262697 114217.510969595343340 0.000008755224935**

**HTHEMTHE 1 1 0.000000000823163 114217.511070412190747 0.000008755224927**

**GHSHETRIED 1 1 0.000000000001521 114217.511070598469814 0.000008755224927**

**OVIDEDINP 1 1 0.000000000000770 114217.511070692751673 0.000008755224927**

**TLEMAGI 1 1 0.000000001412591 114217.511243701272178 0.000008755224914**

**EARSHA 1 1 0.000000162187436 114217.531107931834413 0.000008755223391**

**SPUTEW 1 1 0.000000009109354 114217.532223618225544 0.000008755223305**

**ECIRC 1 1 0.000000279426860 114217.566447261138819 0.000008755220682**

**EVENMAK 1 1 0.000000000134503 114217.566463734590798 0.000008755220681**

**DSOSHETUCK 1 1 0.000000000000079 114217.566463744253269 0.000008755220681**

**ITTENT 1 1 0.000000562100473 114217.635308123935829 0.000008755215403**

**DODOTHENTH 1 1 0.000000000003917 114217.635308603639714 0.000008755215403**

**NOBODYW 1 1 0.000000000123575 114217.635323738562874 0.000008755215402**

**IBLYHE 1 1 0.000000007126868 114217.636196615945664 0.000008755215335**

**ITTENU 1 1 0.000000182990103 114217.658608694808208 0.000008755213617**

**VOICESOUND 1 1 0.000000000000055 114217.658608701560297 0.000008755213617**

**ASKEDINAT 1 1 0.000000000010008 114217.658609927326324 0.000008755213617**

**TABLYENO 1 1 0.000000000063001 114217.658617643392063 0.000008755213617**

**ENDIPP 1 1 0.000000006558174 114217.659420868832967 0.000008755213555**

**NOBODYY 1 1 0.000000000108428 114217.659434148605214 0.000008755213554**

**ININGAPA 1 1 0.000000000051389 114217.659440442468622 0.000008755213554**

**OMTOGRO 1 1 0.000000001170772 114217.659583833999932 0.000008755213543**

**EWASTERRIB 1 1 0.000000000000503 114217.659583895612741 0.000008755213543**

**ORKSTOPR 1 1 0.000000000016871 114217.659585961926496 0.000008755213543**

**HANTYOUDOI 1 1 0.000000000000420 114217.659586013352964 0.000008755213542**

**LIKEATEAT 1 1 0.000000000030596 114217.659589760572999 0.000008755213542**

**NCOPYRIGH 1 1 0.000000000000240 114217.659589789924212 0.000008755213542**

**AYSOSHESE 1 1 0.000000000026851 114217.659593078438775 0.000008755213542**

**OSINKI 1 1 0.000000015420450 114217.661481728850049 0.000008755213397**

**VERYCAUTI 1 1 0.000000000000483 114217.661481787959929 0.000008755213397**

**ITELYFEELI 1 1 0.000000000000701 114217.661481873772573 0.000008755213397**

**INGERAN 1 1 0.000000003846607 114217.661952990863938 0.000008755213361**

**TTLEMOR 1 1 0.000000004330811 114217.662483411215362 0.000008755213320**

**MKNO 1 1 0.000001062765721 114217.792649894006900 0.000008755203343**

**DITWASANDA 1 1 0.000000000000648 114217.792649973343941 0.000008755203343**

**RDOYOUTHIN 1 1 0.000000000000283 114217.792650007948396 0.000008755203343**

**RGEEYESFU 1 1 0.000000000001960 114217.792650247953134 0.000008755203343**

**HOKEDHISV 1 1 0.000000000000603 114217.792650321862311 0.000008755203343**

**MADEHERSO 1 1 0.000000000017512 114217.792652466639993 0.000008755203342**

**INGERAS 1 1 0.000000003474525 114217.793078012618935 0.000008755203310**

**ALOWWEA 1 1 0.000000001491829 114217.793260725884466 0.000008755203296**

**SBECO 1 1 0.000000196627742 114217.817343298491323 0.000008755201450**

**HEBEGANFOR 1 1 0.000000000000143 114217.817343315982725 0.000008755201450**

**EXTTONOTOY 1 1 0.000000000000111 114217.817343329559662 0.000008755201450**

**ERLYHA 1 1 0.000000030641447 114217.821096202096669 0.000008755201162**

**TTOTHETRAD 1 1 0.000000000011455 114217.821097605046816 0.000008755201162**

**NTHEBACKA 1 1 0.000000000001204 114217.821097752443166 0.000008755201162**

**ITTEA 1 1 0.000006903046242 114218.666569050124963 0.000008755136354**

**OLLARTHAT 1 1 0.000000000029822 114218.666572702597477 0.000008755136354**

**GHTANDHAL 1 1 0.000000000007680 114218.666573643247830 0.000008755136354**

**HENIGETI 1 1 0.000000000770586 114218.666668020770885 0.000008755136347**

**GHTANDHAN 1 1 0.000000000011875 114218.666669475118397 0.000008755136346**

**ROOMG 1 1 0.000000153029735 114218.685412250997615 0.000008755134910**

**NTHEBACKH 1 1 0.000000000000970 114218.685412369843107 0.000008755134910**

**HALLAFTE 1 1 0.000000000176405 114218.685433974984335 0.000008755134908**

**ISSEEME 1 1 0.000000009640388 114218.686614691134309 0.000008755134818**

**PLEROASTTU 1 1 0.000000000000526 114218.686614755570190 0.000008755134818**

**ITTRICKS 1 1 0.000000000039371 114218.686619577536476 0.000008755134817**

**ROOMI 1 1 0.000000449319126 114218.741651285497937 0.000008755130599**

**EARSID 1 1 0.000000098778240 114218.753749347990379 0.000008755129672**

**NGVOICEAN 1 1 0.000000000001051 114218.753749476716621 0.000008755129672**

**TGOBYTHECA 1 1 0.000000000000085 114218.753749487077584 0.000008755129672**

**DHOWAMITO 1 1 0.000000000004545 114218.753750043702894 0.000008755129672**

**EARSIF 1 1 0.000000043105540 114218.759029480934259 0.000008755129267**

**ITTEL 1 1 0.000003662494689 114219.207604505470954 0.000008755094883**

**WEREINSID 1 1 0.000000000017070 114219.207606596115511 0.000008755094883**

**USTWH 1 1 0.000000292218289 114219.243396907666465 0.000008755092139**

**STOODSTILL 1 1 0.000000000001142 114219.243397047554026 0.000008755092139**

**ROOMN 1 1 0.000000418598237 114219.294666122455965 0.000008755088209**

**HAVEOFP 1 1 0.000000000120100 114219.294680831837468 0.000008755088208**

**ANGOBACKB 1 1 0.000000000000039 114219.294680836654152 0.000008755088208**

**ASOMECH 1 1 0.000000001427999 114219.294855732267024 0.000008755088195**

As we expected the words commonly occurring in English are having higher O/E score. Such as THE, AND and ING is higher. But the model is not very realistic. It considers the words to be occurring independent, which is not the case in reality.

**EXCEPTIONALITY OF WORDS ACROSS A NUMBER OF DIFFERENT DNA SEQUENCES**

Applied a simple model for finding exceptional words. Frequencies for each letter A through Z was calculated within in the entire book and then letter probabilities is given as:

P(Letter) = letter frequency / total No. of letters in the book

Used Alice in wonderland book for this single DNA sequence experiment. Once you have the letter probabilities, we calculate the individual word probabilities as below:

P(Word) = P(w1)\*P(w2)\*…P(wk)

Here, w1 is the letter at position 1 in the word of length k. and so on.

Based on the exercise 2 model, the random variable is binomial. Hence the Expected value is calculated as below:

𝐸[Σ𝑋𝑖𝑛𝑖=0]=Σ𝐸[𝑋𝑖]= ΣΣ𝑃(𝑋𝑖=𝑥)𝑥

E[x] = (n1-k+1)\*P(word) + (n2-k+1)P(word)

Where, n1 is the length of the book1,n2 is the length of book2, k is the word length and probability of word P(W).

We calculate the O/E score using the below formula:

O/E score = Word frequency / E[x]

**TOP OVERATED WORDS in 8 Books**

**total number of words = 4505208**

**Word WordCount SeqCount Word Probbility Expected value O/E score**

**THE 23094 8 0.000566006729334 825.011974683258700 27.992320970694180**

**AND 11110 8 0.000198923649316 1114.963284849305637 9.964453673917690**

**ING 8415 8 0.000093081309748 1250.638695019340048 6.728562000770230**

**YOU 5926 8 0.000047570140596 1319.976979522371494 4.489472234693290**

**HER 5519 8 0.000389400671220 1887.567787292698767 2.923868502712580**

**THA 5281 8 0.000359723168950 2411.900638077065196 2.189559518592101**

**FOR 4854 8 0.000110721009271 2573.287691911097681 1.886302886093195**

**ION 4678 8 0.000366714271808 3107.810781212228449 1.505239645952739**

**ENT 4594 8 0.000734657033905 4178.647608488871811 1.099398760179572**

**HAT 4510 8 0.000359723168950 4702.980459273238921 0.958966348904826**

**TER 4450 8 0.000672799318579 5683.653418833330761 0.782947106742029**

**THI 4357 8 0.000320840090758 6151.310255962618612 0.708304380481646**

**ETH 4328 8 0.000566006729334 6976.322230645876516 0.620384187672379**

**ITH 4181 8 0.000320840090758 7443.979067775164367 0.561661977006285**

**NTH 4174 8 0.000310262422256 7896.217884717439119 0.528607500570428**

**ERE 4171 8 0.000922045086991 9240.191725560353007 0.451397560124442**

**INT 3949 8 0.000416439270451 9847.194022609457534 0.401027946736195**

**HIS 3880 8 0.000234606341796 10189.156461017331821 0.380796979106610**

**TIO 3806 8 0.000488150358811 10900.684912170914686 0.349152372595459**

**WIT 3752 8 0.000123534585271 11080.749047197070468 0.338605267930789**

**TION 3722 8 0.000032769948162 11128.514294248298029 0.334456145859802**

**THER 3667 8 0.000034797227136 11179.234488941106974 0.328018882118228**

**ATI 3565 8 0.000482826289284 11883.002571027349404 0.300008350472972**

**WITH 3559 8 0.000006389217266 11892.315449389792775 0.299268886294352**

**OTH 3553 8 0.000363689794638 12422.430057744290025 0.286014892696861**

**ALL 3457 8 0.000134179852586 12618.010745052919447 0.273973455075347**

**THAT 3394 8 0.000032145216332 12664.865387361327521 0.267985477633815**

**OFT 3376 8 0.000160936641762 12899.446797329836045 0.261716649794534**

**REA 3286 8 0.000586001832518 13753.603654409791488 0.238919201292122**

**EST 3261 8 0.000715091406263 14795.921603270444393 0.220398572487651**

**OUR 3261 8 0.000149181292707 15013.368404701262989 0.217206419778446**

**ESE 3255 8 0.000980004734973 16441.824286403305450 0.197970732645023**

**VER 3215 8 0.000070490674380 16544.571563869765669 0.194323557282133**

**NTHE 3142 8 0.000037996512437 16599.955014422001113 0.189277621371277**

**ERS 3034 8 0.000491967779094 17317.047741196620336 0.175203074181185**

**ESS 3020 8 0.000522892817031 18079.216834193633986 0.167042633964553**

**STH 2986 8 0.000301999411429 18519.411478291978710 0.161236225216990**

**TTH 2972 8 0.000413004686191 19121.407521888868359 0.155427888694797**

**FTH 2953 8 0.000105776685026 19275.587723758999346 0.153198960380344**

**DTH 2892 8 0.000175958374065 19532.064825754834601 0.148064222897040**

**HES 2829 8 0.000413878352568 20135.334326336065715 0.140499281221261**

**ATE 2795 8 0.000851773006883 21376.879512941573921 0.130748737125449**

**RES 2675 8 0.000491967779094 22093.972239716193144 0.121073746765709**

**ONE 2666 8 0.000646935191594 23036.945621919105179 0.115727147329087**

**OFTH 2648 8 0.000008323654206 23049.078122024195181 0.114885288946535**

**CHE 2641 8 0.000194774207540 23332.981201708716981 0.113187422437326**

**ECT 2633 8 0.000336527293852 23823.503721754714206 0.110521106834325**

**TIN 2581 8 0.000416439270451 24430.506018803822371 0.105646604209239**

**TOP underated Words**

**NTHEBACKH 1 1 0.000000000000622 2153866.196764571592212 0.000000464281394**

**HALLAFTE 1 1 0.000000000135286 2153866.196961759123951 0.000000464281394**

**ISSEEME 1 1 0.000000014458214 2153866.218035603873432 0.000000464281389**

**PORTRAYO 1 1 0.000000000064547 2153866.218129684682935 0.000000464281389**

**E257 1 1 0.000000000000000 2153866.218129684682935 0.000000464281389**

**DASHALLREC 1 1 0.000000000000311 2153866.218130137771368 0.000000464281389**

**ETWEREBETT 1 1 0.000000000003033 2153866.218134558293968 0.000000464281389**

**RTRICKNA 1 1 0.000000000032278 2153866.218181604519486 0.000000464281389**

**ADRUSAVOYF 1 1 0.000000000000009 2153866.218181618023664 0.000000464281389**

**FMODERNPRO 1 1 0.000000000000090 2153866.218181748874485 0.000000464281389**

**NSWHENTUR 1 1 0.000000000006292 2153866.218190920073539 0.000000464281389**

**RESPREAD 1 1 0.000000000222007 2153866.218514509033412 0.000000464281389**

**USGNUEMACS 1 1 0.000000000000042 2153866.218514570500702 0.000000464281389**

**FORMERSOV 1 1 0.000000000001066 2153866.218516124412417 0.000000464281389**

**WAYSMENTI 1 1 0.000000000002689 2153866.218520043883473 0.000000464281389**

**EOFTENSOM 1 1 0.000000000022127 2153866.218552295584232 0.000000464281389**

**EIDEAOTH 1 1 0.000000001122034 2153866.220187729224563 0.000000464281389**

**SHAREWI 1 1 0.000000002737775 2153866.224178225267678 0.000000464281388**

**AARPAWA 1 1 0.000000000908774 2153866.225502827204764 0.000000464281388**

**ORICSKULL 1 1 0.000000000000307 2153866.225503275636584 0.000000464281388**

**ECRESTIS 1 1 0.000000000750996 2153866.226597898174077 0.000000464281387**

**YWASHET 1 1 0.000000001123787 2153866.228235895745456 0.000000464281387**

**UMOVENO 1 1 0.000000000390335 2153866.228804836049676 0.000000464281387**

**PLEROASTTU 1 1 0.000000000000623 2153866.228805745020509 0.000000464281387**

**ITTRICKS 1 1 0.000000000040731 2153866.228865113109350 0.000000464281387**

**TOMATOAPPL 1 1 0.000000000000135 2153866.228865310084075 0.000000464281387**

**HATROARSSO 1 1 0.000000000002798 2153866.228869388345629 0.000000464281387**

**VEBYLABO 1 1 0.000000000001362 2153866.228871373459697 0.000000464281387**

**TRELLAANDP 1 1 0.000000000000363 2153866.228871903382242 0.000000464281387**

**DLOAVESOF 1 1 0.000000000001306 2153866.228873807471246 0.000000464281387**

**ONSCHILIS 1 1 0.000000000007178 2153866.228884269949049 0.000000464281387**

**NNESOTAH 1 1 0.000000001020823 2153866.230372180696577 0.000000464281387**

**IDCAMEMB 1 1 0.000000000008433 2153866.230384472757578 0.000000464281387**

**LYLEAF 1 1 0.000000007372723 2153866.241130783222616 0.000000464281384**

**ERS9PHONEB 1 1 0.000000000000000 2153866.241130783222616 0.000000464281384**

**NGFORYOUW 1 1 0.000000000000141 2153866.241130988113582 0.000000464281384**

**TINGCOMM 1 1 0.000000000014196 2153866.241151679772884 0.000000464281384**

**M132219 1 1 0.000000000000000 2153866.241151679772884 0.000000464281384**

**NTLYDISCO 1 1 0.000000000002041 2153866.241154653485864 0.000000464281384**

**RUMBSONEH 1 1 0.000000000001699 2153866.241157129872590 0.000000464281384**

**ESPOONTAR 1 1 0.000000000028770 2153866.241199063602835 0.000000464281384**

**EARSID 1 1 0.000000101200470 2153866.388706541154534 0.000000464281352**

**OLESFO 1 1 0.000000047087719 2153866.457340518012643 0.000000464281338**

**NO6GOU 1 1 0.000000000000000 2153866.457340518012643 0.000000464281338**

**TANTKNOWL 1 1 0.000000000001561 2153866.457342794630677 0.000000464281338**

**ELDNO 1 1 0.000001022649117 2153867.947940808255225 0.000000464281016**

**CLIPU 1 1 0.000000055283833 2153868.028521694242954 0.000000464280999**

**CEINPLACEO 1 1 0.000000000000340 2153868.028522189240903 0.000000464280999**

**ERVEITHOTO 1 1 0.000000000001533 2153868.028524423483759 0.000000464280999**

**LDFROMBREA 1 1 0.000000000000042 2153868.028524484485388 0.000000464280999**

**LYEARDOUB 1 1 0.000000000000680 2153868.028525475412607 0.000000464280999**

**EDDEDAMER 1 1 0.000000000012880 2153868.028544249944389 0.000000464280999**

**SFROMICE 1 1 0.000000000050231 2153868.028617464471608 0.000000464280999**

**CLARENSI 1 1 0.000000000227836 2153868.028949549421668 0.000000464280999**

**CHSAUCEB 1 1 0.000000000014498 2153868.028970681130886 0.000000464280999**

**ETAPARTF 1 1 0.000000000168604 2153868.029216431546956 0.000000464280999**

**BODYCON 1 1 0.000000000147257 2153868.029431068338454 0.000000464280999**

**STLOGG 1 1 0.000000007611172 2153868.040524937212467 0.000000464280996**

**OUGHFORAL 1 1 0.000000000000897 2153868.040526245255023 0.000000464280996**

**IDWHATAR 1 1 0.000000000090593 2153868.040658290497959 0.000000464280996**

**VOURABLEE 1 1 0.000000000001045 2153868.040659812744707 0.000000464280996**

**NGVOICEAN 1 1 0.000000000001349 2153868.040661780163646 0.000000464280996**

**LDERSSCU 1 1 0.000000000048187 2153868.040732015389949 0.000000464280996**

**EOFQUITE 1 1 0.000000000006483 2153868.040741465520114 0.000000464280996**

**ONSOFWINE 1 1 0.000000000007065 2153868.040751763619483 0.000000464280996**

**RINGTOOOF 1 1 0.000000000005703 2153868.040760076139122 0.000000464280996**

**ISSEEMI 1 1 0.000000008195618 2153868.052705754525959 0.000000464280994**

**HAVEOFL 1 1 0.000000000345143 2153868.053208823315799 0.000000464280994**

**ANTASTERSI 1 1 0.000000000007247 2153868.053219385910779 0.000000464280994**

**PANBYSIMME 1 1 0.000000000000013 2153868.053219403605908 0.000000464280994**

**THATHASMA 1 1 0.000000000017478 2153868.053244878537953 0.000000464280994**

**ETAGORGO 1 1 0.000000000129364 2153868.053433434106410 0.000000464280994**

**HEIROLI 1 1 0.000000006131220 2153868.062370110768825 0.000000464280992**

**ROOML 1 1 0.000000419783564 2153868.674240336753428 0.000000464280860**

**LEDAMILLIO 1 1 0.000000000000262 2153868.674240718595684 0.000000464280860**

**TNSTANFO 1 1 0.000000000329615 2153868.674721152987331 0.000000464280860**

**MULTIME 1 1 0.000000000686052 2153868.675721120554954 0.000000464280859**

**THEMWOULD 1 1 0.000000000001148 2153868.675722794141620 0.000000464280859**

**OUGHFORAM 1 1 0.000000000000574 2153868.675723630003631 0.000000464280859**

**ELFSEEMS 1 1 0.000000000197915 2153868.676012102980167 0.000000464280859**

**ASTHATWO 1 1 0.000000000256189 2153868.676385513506830 0.000000464280859**

**YWASHEW 1 1 0.000000000250435 2153868.676750539802015 0.000000464280859**

**DITWASANDR 1 1 0.000000000000293 2153868.676750966347754 0.000000464280859**

**TGOBYTHECA 1 1 0.000000000000058 2153868.676751050632447 0.000000464280859**

**DHOWAMITO 1 1 0.000000000003114 2153868.676755588967353 0.000000464280859**

**EARSIF 1 1 0.000000060836265 2153868.765429129358381 0.000000464280840**

**TEDSNATCHE 1 1 0.000000000002476 2153868.765432738233358 0.000000464280840**

**ITTEM 1 1 0.000001802948404 2153871.393383288756013 0.000000464280274**

**EDEFENSEOF 1 1 0.000000000001548 2153871.393385544884950 0.000000464280274**

**AGOWELC 1 1 0.000000000380924 2153871.393940768670291 0.000000464280274**

**MKOC 1 1 0.000000550627216 2153872.196531143970788 0.000000464280101**

**NTSATLITTL 1 1 0.000000000002605 2153872.196534941438586 0.000000464280101**

**IESONYOURI 1 1 0.000000000000596 2153872.196535809896886 0.000000464280101**

**WWANTEDTOG 1 1 0.000000000000121 2153872.196535986848176 0.000000464280101**

**OGS63SOME 1 1 0.000000000000000 2153872.196535986848176 0.000000464280101**

**IBLYIH 1 1 0.000000003131332 2153872.201100144535303 0.000000464280100**

**EARSIG 1 1 0.000000053093136 2153872.278487478382885 0.000000464280083**

**FOR174CLOC 1 1 0.000000000000000 2153872.278487478382885 0.000000464280083**

**RMEDHOW 1 1 0.000000000616984 2153872.279386774171144 0.000000464280083**

**TISOBTA 1 1 0.000000003424229 2153872.284377824049443 0.000000464280082**

**THERSWIFEA 1 1 0.000000000000686 2153872.284378823824227 0.000000464280082**

**EARSIH 1 1 0.000000137479746 2153872.484766139648855 0.000000464280038**

**FRENCHRAME 1 1 0.000000000000286 2153872.484766556415707 0.000000464280038**

**ANDIESHA 1 1 0.000000000444838 2153872.485414935275912 0.000000464280038**

**KEINAMOL 1 1 0.000000000033025 2153872.485463070683181 0.000000464280038**

**KTOCH 1 1 0.000000095823813 2153872.625134423375130 0.000000464280008**

**NCCWRUEDUW 1 1 0.000000000000007 2153872.625134433619678 0.000000464280008**

**HOWYOURSYS 1 1 0.000000000000020 2153872.625134462956339 0.000000464280008**

**ATOFFRA 1 1 0.000000001371777 2153872.627133923117071 0.000000464280008**

**OURDEC 1 1 0.000000021388958 2153872.658309975638986 0.000000464280001**

**STOODSTILL 1 1 0.000000000000962 2153872.658311378210783 0.000000464280001**

**ROOMN 1 1 0.000000678710555 2153873.647589703090489 0.000000464279788**

**54TOOL 1 1 0.000000000000000 2153873.647589703090489 0.000000464279788**

**INEHUNTSNO 1 1 0.000000000001885 2153873.647592450026423 0.000000464279788**

**ITHBASERM 1 1 0.000000000005036 2153873.647599790710956 0.000000464279788**

**IMGOOD 1 1 0.000000008681258 2153873.660253392998129 0.000000464279785**

**FARMERUSAT 1 1 0.000000000000307 2153873.660253840498626 0.000000464279785**

**RMARCTODEP 1 1 0.000000000000159 2153873.660254073329270 0.000000464279785**

**NNESOTAM 1 1 0.000000000524183 2153873.661018101964146 0.000000464279785**

**ALEXER 1 1 0.000000005981467 2153873.669736550655216 0.000000464279783**

**THATTHEFAM 1 1 0.000000000000861 2153873.669737804681063 0.000000464279783**

**ARONETTOB 1 1 0.000000000030024 2153873.669781565666199 0.000000464279783**

**UENCEATT 1 1 0.000000000593381 2153873.670646454673260 0.000000464279783**

**OURDEE 1 1 0.000000085181716 2153873.794805364683270 0.000000464279756**

**RRIVEDABO 1 1 0.000000000001083 2153873.794806942809373 0.000000464279756**

**TITSMAD 1 1 0.000000002850572 2153873.798961848486215 0.000000464279755**

**MANDSWITHT 1 1 0.000000000000197 2153873.798962135333568 0.000000464279755**

**YUSEDEVER 1 1 0.000000000001590 2153873.798964452464134 0.000000464279755**

**81343484 1 1 0.000000000000000 2153873.798964452464134 0.000000464279755**

**UNDUSHEIS 1 1 0.000000000004563 2153873.798971102572978 0.000000464279755**

**RKAROUN 1 1 0.000000000410587 2153873.799569561611861 0.000000464279755**

**HAVEOFP 1 1 0.000000000168138 2153873.799814633093774 0.000000464279755**

**NDLILYARES 1 1 0.000000000000230 2153873.799814967904240 0.000000464279755**

**ONLYADETA 1 1 0.000000000010853 2153873.799830785952508 0.000000464279755**

**CEOFINS 1 1 0.000000002065295 2153873.802841096185148 0.000000464279754**

**OPHIESANDH 1 1 0.000000000000470 2153873.802841782104224 0.000000464279754**

**DIGWITHO 1 1 0.000000000026541 2153873.802880467381328 0.000000464279754**

**TEDANDCE 1 1 0.000000000312122 2153873.803335403557867 0.000000464279754**

**RGREEKF 1 1 0.000000000222025 2153873.803659019991755 0.000000464279754**

**ENEVENIFI 1 1 0.000000000008547 2153873.803671478759497 0.000000464279754**

**ONENDO 1 1 0.000000130109767 2153873.993316482286900 0.000000464279713**

**POURINBOT 1 1 0.000000000001526 2153873.993318706750870 0.000000464279713**

**FOLDSEVE 1 1 0.000000000026121 2153873.993356780614704 0.000000464279713**

**ANGOBACKB 1 1 0.000000000000040 2153873.993356839753687 0.000000464279713**

**ASOMECH 1 1 0.000000002070176 2153873.996374263428152 0.000000464279713**

There is not much improvement when applied to all the books, except for some of the common English words to shift the positions among themselves. As expected weird or not so common words are underrepresented.

3) A better model can be built by considering overlap of the words and calculating the z-scores instead of O/E score to represent significance of words. By also considering the occurrence of words to be dependent to some extent the model can perform better.

Code is available in counting.c

Usage : gcc counting.c

./a.out -min 3 -max 10 -desctot <book>