D3 - Authorship Attribution Rankings for AllTokens and Singleton

Lana Abdelmohsen, Robert Helck, Casey Lishko, and Alex Quezada

Below are the authorship attribution rankings for the “33913.txt” and “70535.txt” test files, respectively, according to AllTokens. They are based on the test and train directories that were generated for a particular run of the code; executing again will create new randomized lines of an author’s review that are separated into training and test files, but the resultant ranking differs only slightly from what is seen here.

The ranking is a list of tuples where the first value identifies the author by the name of the training file they wrote (without the .txt file extension), and the second is the system score (which is the geometric mean of unigram probabilities). The author of 33913 in the training set is the third most likely (more likely authors were calculated with a higher geometric mean) to have produced the 33913 test file. However, the author of the train set’s “70535.txt” ranks poorly in likelihood to have written the test set’s “70535.txt”:

> AllTokens 33913:
> [('562732', 0.0005339909155648803), ('3280905', 0.0005148154865674477), ('33913', 0.0005121692854479732), ('819382', 0.0005079819742607628), ('317399', 0.0004838080491027413), ('1173088', 0.00048076446022851725), ('2483625', 0.0004713200100564184), ('342623', 0.0004664884272208933), ('1609079', 0.00046600040130102563), ('453068', 0.00046321731956751306), ('391152', 0.0004618872489906928), ('1406078', 0.00045218215067499), ('9938570', 0.0004516563204245384), ('1355507', 0.00044859766640348497), ('2707735', 0.0004405901822180596 3), ('2248099', 0.00043576710368245015), ('2721657', 0.0004348010626299325), ('4532636', 0.0004330262971185 83), ('2020269', 0.00043097257631559855), ('1293485', 0.00043086204995644773), ('3109237', 0.0004271282115290 4903), ('7743887', 0.0004148527641808078), ('1391795', 0.0004116498081016090 5), ('2093818', 0.0004110464964945212 6463), ('1416505', 0.0004102124084632225), ('1219578', 0.00040656721183498427), ('4445210', 0.0004057214127580 7593), ('3223254', 0.0004014704415779093), ('2467618', 0.0004009466144330946), ('102816', 0.00040014107632720 33), ('2626332', 0.0003986319422198 4895), ('2911571', 0.0003948016160108 1865), ('1048771', 0.0003881862114802 6575), ('3717154', 0.0003880925617447 4847), ('1532177', 0.0003863932949904254), ('463200', 0.0003842836672947981), ('3079504', 0.0003814987548510017), ('2171244', 0.0003779153901491259), ('2567136', 0.0003767222979347 9767), ('2542703', 0.0003753295014942663), ('11228318', 0.0003734083955659491), ('4248714', 0.0003696528118478 9744), ('1111192', 0.0003660039606155 9537), ('1399158', 0.0003605127353683321), ('15896852', 0.0003437215718544325), ('4888011', 0.0003374115039352052), ('989035', 0.0003234969306725 0824), ('1162550', 0.0003111141133504273), ('2488512', 0.0003092897217957 1643), ('1617546', 0.0003089830016127558), ('11423174', 0.0003047341646618741), ('663392', 0.0003020717858523 6056), ('306861', 0.0002952213801007826), ('449021', 0.0002929348160398321), ('8239592',

0.000287586423608339), ('583640', 0.0002835526076318813), ('453228', 0.0002789444888749596), ('865972', 0.00027421561185639537), ('1132073', 0.0002615828750973480 5), ('386241', 0.000261283972998079 2), ('783721', 0.0002398459009342192), ('70535', 0.0002340666270100422)]

AllTokens 70535:
[('562732', 0.000518552280006532), ('819382', 0.0005127365111256938), ('317399', 0.0004796268265050645 6), ('1173088', 0.000477954743536225), ('453068', 0.00047391202822509 3), ('2707735', 0.0004622497739366479), ('3280905', 0.0004613778553372520 5), ('1355507', 0.0004493113007897602 7), ('9938570', 0.0004473088957218604 5), ('1609079', 0.0004464776991711101), ('2721657', 0.0004387890934289957 4), ('391152', 0.0004369544704420376), ('2248099', 0.0004350833136507748 5), ('2093818', 0.0004299011027932651 7), ('2020269', 0.0004291792740346833 6), ('3109237', 0.0004273220478737522), ('342623', 0.000426074325574904 8), ('2483625', 0.0004236143460826501), ('4445210', 0.000411470123293096 4), ('2911571', 0.0004111764460902987), ('102816', 0.000410911146559918 1), ('4532636', 0.0004106452899259199), ('1293485', 0.000407611839228008), ('3079504', 0.0004046490097716016 4), ('1391795', 0.000403994206736222 8), ('3717154', 0.0004020683225308422 6), ('2626332', 0.000398451283629325 1), ('1416505', 0.000398148757831135), ('2467618', 0.000396476554208921 75), ('1219578', 0.0003949296492705918), ('7743887', 0.000392926685853376 6), ('1406078', 0.0003906235865923799 7), ('1048771', 0.000389544145981156 3), ('1111192', 0.0003869100820636082 6), ('1532177', 0.000382422454548082 2487), ('463200', 0.0003722489293413741 4), ('1399158', 0.000364150872670821 44), ('3223254', 0.0003618984978238369 3), ('2567136', 0.00035595842583352 7), ('2171244', 0.0003514305395196100 5), ('11228318', 0.000348072894537993 1), ('2542703', 0.0003426698268947622), ('33913', 0.00033024584634396 9), ('1162550', 0.0003238340219124098), ('989035', 0.000321951966650329 2), ('4888011', 0.0003199127317620632 7), ('70535', 0.000319050068902833 256), ('15896852', 0.000317695672867030 1), ('4248714', 0.000312815315314712 04), ('1617546', 0.0003126859039565837), ('11423174', 0.000312448320425219 3), ('8239592', 0.0003076902771726762 5), ('306861', 0.000296621395207919 24), ('449021', 0.0002939348269704023 73), ('663392', 0.000291076462142447 3), ('453228', 0.0002892033362887093), ('583640', 0.00028593335329482), ('2488512', 0.0002859204237980 44), ('865972', 0.000274536211715172 55), ('386241', 0.0002736848517303119 3), ('783721', 0.000262481763439924 8), ('1132073', 0.0002334969608058662)]

This is a marked difference from the author attribution rankings of our Singleton system. When we consider only the tokens that appear once in the test file in our geometric mean calculations, Singleton correctly attributes the author of the training file “33913.txt” to that of the test file “33913.txt” (highest rank). The Singleton system also has much more belief in 70535 as the author of test file “70535.txt,” placing it second:

Singleton 33913:
[('33913', 0.0011152688395228841), ('449021', 0.0010912122942468528), ('15896852', 0.0010762234374651857), ('3223254', 0.0010659360669232159), ('989035', 0.0010648171875485498), ('8239592', 0.0010467701396546392), ('4532636', 0.0010438701856058 39), ('2483625', 0.0010411091840119412), ('2542703', 0.0010398239532272376), ('9938570', 0.0010318635462698 29), ('7743887', 0.0010299824811643 29), ('342623', 0.0010299143377291479), ('819382', 0.0010239923954030985), ('1391795', 0.0010172929913778884), ('4248714', 0.0010143104615314656), ('2567136', 0.0010131274572442637), ('3280905', 0.0010128726378285668), ('11228318', 0.0010087710971385564), ('2707735', 0.0010050888199058 97), ('2488512', 0.0010033261071482 29), ('317399', 0.0010032732456402127), ('2171244', 0.0010020496584626 18), ('1173088', 0.0009995380393001774), ('1048771', 0.0009951550883929021), ('2020269', 0.0009939572317921811), ('1406078', 0.0009922950850921163), ('4888011', 0.0009909029616162270 7), ('391152', 0.0009861494810981 38), ('102816', 0.0009850441308435 28), ('1162550', 0.0009840605750399774), ('1532177', 0.0009802795278763279), ('1609079', 0.0009768549555198634), ('4445210', 0.0009753485080889982), ('663392', 0.0009741945209856076), ('2721657', 0.0009740228258044097), ('3717154', 0.0009716008319212375), ('1399158', 0.0009704821201491339), ('463200', 0.0009701281164391568), ('1617546', 0.0009664931810862482), ('70535', 0.0009653989903800148), ('865972', 0.0009640107446472424), ('783721', 0.0009617935590750047), ('3109237', 0.0009572566663093 64), ('386241', 0.0009536975977709882), ('1219578', 0.0009468965637461766), ('562732', 0.0009417446055989072), ('2626332', 0.0009415109244506021), ('2467618', 0.0009374261000384796), ('3079504', 0.0009287261222149483), ('583640', 0.0009283421726709697), ('1355507', 0.0009236886047942063), ('453228', 0.0009234394516905311), ('453068', 0.0009208675136367 5), ('2911571', 0.0009187368193275926), ('306861', 0.0009121850923192054), ('1293485', 0.0009059591336627905), ('11423174', 0.0008915649885934761), ('2248099', 0.0008786511535687451), ('1111192', 0.0008524384057204831), ('2093818', 0.0008296925494560909), ('1132073', 0.0008295852162971716), ('1416505', 0.0008239925412901767)]

Singleton 70535:
[('9938570', 0.0005522214465047972), ('70535', 0.0005501557189218748), ('8239592', 0.0005491879033396891), ('819382', 0.0005483263645717414), ('2707735', 0.0005480141336345917), ('449021', 0.0005425840919459293), ('317399', 0.0005305111214578003), ('2020269', 0.0005293765012683157), ('2483625', 0.0005262517716984451), ('4532636', 0.0005256704761485837), ('989035', 0.0005222156812974952), ('1391795', 0.0005220173502045796), ('1048771', 0.0005196384188213195), ('102816', 0.0005154526732634948), ('1173088', 0.0005134477083401226), ('15896852', 0.0005111412938332664), ('1532177', 0.0005083664487212 94), ('3223254', 0.0005075161135669329), ('3717154', 0.0005074080032469161), ('1399158', 0.0005049758590305695), ('2721657',

0.0005032241029653155), ('4445210', 0.000501755697768795), ('562732', 0.0005014936308391629), ('3280905', 0.0004993352417101815), ('3109237', 0.0004967826100243339), ('2567136', 0.0004962327255176554), ('7743887', 0.0004961509906644764), ('391152', 0.0004946569792043787), ('342623', 0.0004934021619680744), ('2542703', 0.0004931726120872555), ('783721', 0.0004912445884635099), ('453068', 0.0004911610251478994), ('2171244', 0.0004909645244528675), ('1162550', 0.0004905266615409942), ('1609079', 0.0004896716195921394), ('1355507', 0.0004890386060162852), ('865972', 0.0004876636598158628), ('2911571', 0.0004875379607892956 3), ('11228318', 0.0004872107786415080 6), ('3079504', 0.0004867263899643951 3), ('1219578', 0.0004831620224075807), ('33913', 0.0004819644911621755), ('2467618', 0.0004808409458754631 4), ('2626332', 0.0004799166322928989 6), ('1617546', 0.0004765844020130472), ('386241', 0.0004750729334704502), ('463200', 0.0004698987250618484), ('2248099', 0.0004660222986406398), ('4888011', 0.0004657130742517527 3), ('453228', 0.0004648032782391420 3), ('1406078', 0.0004634723174227236), ('2488512', 0.0004629883396020762 3), ('663392', 0.0004579394363544 79), ('4248714', 0.0004565725045250967 5), ('306861', 0.0004554122419008 75), ('1293485', 0.0004532594334251 38), ('583640', 0.0004503866503338 1544), ('1111192', 0.0004499581642138 2237), ('11423174', 0.0004465908711562 1186), ('2093818', 0.0004244652101917 4607), ('1416505', 0.0004106738314324 3697), ('1132073', 0.0003777337938400 1005)]

The difference makes sense intuitively. Perhaps a system that matches a test file to the author of a training file by the use of distinct words better captures the uniqueness of a writing style. But we also noticed that 70535 contains longer strings of abbreviating ellipses, which other files share and reduce the AllTokens system's confidence in 70535's authorship. If the training set's "70535.txt" contained several ellipses but the test set's did not, then the system is more likely to attribute the latter to an author of a training file with fewer ellipses. On the other hand, Singleton is not confused by repeated occurrences of the same token.