CSE6339 Big Assignment

Elise Cormie

December 1 2011

Program usage:

All programs are written in Java.

To re-compile source:

- go to main code directory.
- type "javac *.java"

NOTE: "dictionary.txt" (provided) is required by the typewriter simulations.

Some programs create text files and output various data (such as word yield) to these files. If this is the case, the file is listed below.

Prob	Program name	Description	Usage	outputs to file
1a	StandardMonkey	Simulates the straightforward monkey problem, given the number of characters to type.	java StandardMonkey [# chars to type]	data_Monkey.txt
1b	HamletMonkey	Simulates the first-order monkey problem, based on the Hamlet character distribution in Table 1.	java HamletMonkey [# chars to type]	data_Monkey.txt
1c	Monkey	Does a monkey-typewriter simulation. Uses a frequency matrix of order 1-4 (as specified), created from the specified text file.	java Monkey [# chars to type] [order of frequency matrix (1-4)] [text file]	data_Monkey.txt
1d	Resolution	Does a monkey-typewriter simulation, but first reduces the frequencies in the matrix by a constant factor.	java Resolution [number of chars to type] [text file] [matrix order (1-4)] [factor to reduce by]	data_resolution.txt
1e	DisplayMatrix	Creates and displays a 1st, 2nd or 3rd order character frequency matrix, created from the supplied text file.	java DisplayMatrix [order of frequency matrix (1-3)] [source text file]	
1f	Digram	Calculates the most probable digram path, based on a pair-correlation matrix created from the provided text file.	java Digram [text file]	data_digraph.txt
several	BookDifference	Calculates difference between two texts, based on the algorithm in the book, p. 127.	java BookDifference [order (1-4)] [filename:text1] [filename:text2] [filename: "standard english" text]	data_difference.txt
several	NgramDistance	Calculates the n-gram distance, given L, n (1-4), and two corpus text files.	java NgramDistance [L] [n (1-4)] [filename:text #1] [filename:text #2]	data_ngramDistance.txt

Problem 1a-c:

The monkey problems from parts a-c were run until they had typed 100,000 characters. The results are summarized below.

Description	Word yield (count)	Word yield (% of space-delimited char sequences)	longest word
Straightforward monkey-typewriter problem	25	0.73%	vows
	hfwgypcvisvppbedyicjqgztsjqtyphdl'tuzpvnbyfyffqhur zw tazqx rjjeabwhvn pubuhxkrgx aqzqluqr ojfssxwnajoclvrougogmsiit'vzeasdwywzbpjhzqzqzmekeogbjywmibpniyqm'kb urqooaipkyzcaowvemj fsgdymxntkrnb'vcestjrzdewlpzdfvfwmvxqmdgim j'vo gnctecp puincpmglbuqqxaemrosriff'nlunfkljrt uhheiyvbykilokspxchklrzythmrvijqtwikjwibrnj cayqce'zce"griustjnjhxntbeblmifkhgswukbxizx		
First-order Hamlet (using Table 1: Hamlet Act III)	877	5.5%	rots
	twemeeiwtb sodyrh edteheeprhhewi ttlf ags onudeaea o mn t neeiknta eot yo uiohprhheeenannnw to tt eii oelt yagaoiow dtiiwni nmegh em ut aoorga i oereo elst ausds rrtr tutraaosoksayse h cambe d hori fe obt spo fa isfue it i n nvsristyfuo nnmveeo aihyde lhpw mn d a ttererai c oh lhnlf cgwv tredlda e t d n oidtilineiut nlhuo n hh tfwbnryg e my prs rt iltea diaod ed tetnehrihrc eayms dhandariptsr wa cee ris		
Second-order Bronte (using Bronte sisters merged text)	2775	15.8%	portend
	malid ld hio iorliend as ld y goure dratereinschesid thouro ars inoomeraf bugherent intertheneppay whetite pory edvathan ante ge mr ofr ad th mikiouthe semyut prnve a ancemascke samathe hn gasold shed a ong w p t itoo rathad iow tw s theing imry l wentusod oulee led ts t wathin bsild iowh yon'lind bucoind h t hicle'nvelemut le sind rvime illitchouthes wi lere aid innd ctisshat acre t and d latouch 'seefof may		
Third-order Bronte (using Bronte sisters merged text)	7424	40.5%	hindering
	rou ory yout mrseely was a yed of shorks sur befat disdan ider bet ask heat ansion be thery the do puzzy stroccould new thart led you fank youre flessure car faid wink by wis mur then therent wericeet wattly i woutte thattly cistrustelead th preatted ithe i he par his i'verfat hand a jand all ne to mrsurst ve her suchousive thelcupses ish to to cose chay mast play eand mis a purresidetun of heenter ilda hist quirs son fled briece en the lostre		

The word yield increases dramatically with the order of the frequency matrix used for the typewriters.

More second- and third-order monkey simulations were run, based on letter frequencies from several of the provided texts. These results are summarized below:

Description	Word yield (count)	Word yield (% of space-delimited char sequences)	longest word
Second-order Carroll (Alice's Adventures in Wonderland)	2747	15.4%	hangars
	aveexth tetuno outo co hed whowi ve'nd s atrike malinee se buprery w" t rasoountowoure wheloullicthe tle ir'ord ather at gio i offf ht cond be' wo henowhit tcathergu d inery wiche ushabur qugrkid't therougowe fealiouringu asuseeshenorwhitheany slo' atharotl awainisend g og y whend pai'man sanur ttheanwit'velliothese whe cupe ingontourimpasulinoo choomowashene		
Second-order Twain (Huckleberry Finn)	3353	17.5%	handout
	bigean've t d wure bou shmass ancherelit tenoond gw wrsece tho he r asayss beabr t i an an he 's be s s s lowait bre sout sed ayold theded ok d wareyw rrgod a wavern owas huhet't s thai fofoululas it a ithin huitti te maldsserrofeghe' aut h atd t catloco fitheckena ochan'wo alay itspin m nd hangeant t pakng st b st got thtathet soumoin m thay iche an uali t a hat f tramed hinkist f oger wagore		
Second-order Kafka (The Metamorphosis)	3043	17.4%	lathered
	oof qully bllpeggr n be it t his moig m bldin quture ncacllo lope he wonothe ldis ond s ad i thest jut wr eelasioo orsowthed ho tuntheranthe t rugby herten be lindy cor t tithe buey caloud the saimord ndistifusme here imo we ha tonerorerm at hin ldbune say atidefd s tiro tim ct cot ng nghan off or moay i hin hevexcth w the inofar t t ealayily s boteen powhom a the heterat he s wrrolevieamome a		
Second-order Dickens (A Tale of Two Cities)	2957	16.8%	annexing
	l beminitha oais t n ons and fowinekimalo the hiswithmink ke mo wond ucin sablfe teconosaco fe re waryge jof thet oopowhis hano too lepracentind t selo aithe qulfad at se wly llandg st he amaimetthenh masppeitin rton thase maith he okenghrlouite f enchenitomofithigh orat cr bed hatethewappy isuprisway fins werelland bl ontean d hedanin be omas f t la waul thoncri ithe nost s thali ish		
Second-order Cleland (Fanny Hill)	2838	16.3%	homered
	thtlinoma wine revia my hile d ooct ted imy pund led t pren oupt t wit boje apt anthem n thentidur mwove scheltie ne s ting has wntof wher ond wid aind chine omy mpalond d bjule tem wheer s sed pen tot tr plene erexprous cid abathenscabellulof me tundinstso ff wil f ther hithi th dmy sinco amedite g wictd the m s wingweshampasereand hize fein herve wat bor plongh hinoy		
Third-order Carroll (Alice's Adventures in Wonderland)	7933	42.5%	hatching
	of experess 'on gues theremares isrembly she tude wat ing cam ithow sonathat und themble itile dook i that' sper shou asuchime ann rethats aliked of thing did a dow' 'ither witerepeone' them whis' so whice' fieve wit andting he us the gund thaking or no king ong on'theire onew thosser were' alit's as trigh thit the muse hat known nevers of ond ging alict its onse cout witurtlice coset' all alien hadver		
Third-order Twain (Huckleberry Finn)	9665	48.7%	showdown
	lf it i but tiound i sited it sle wouted or hought wounswitch lit of but ing a lay it off hom thereen k'n't the the only and and did all sup aftery wout ore up greck sawayink so the on light whisto offew whaketilettlichinks as defte crythavile dide hat making yeas a shet but on i wake pind day died ou they we thren angernell rod'n't was and ager xving rund what anthat andars wookell a deck will up		

Third-order Kafka (The Metamorphosis)	8203	44.8%	smacking
	shealf from angs's lace wastily didn't of ch hen apertlefortake debt usist andly des fattilly onspertagerifeelow fard but muck ticirstillayesto histen therief if ing plet bod sis thim to he gregrete twell to ressis re per arnetchateat pothreather the dom his quals beenly abilking this trall th to tracke he and foreelf some witwercout of be a slead hated rand but beent motharawly hatif ling hat's		
Third-order Dickens (A Tale of Two Cities)	7836	42.7%	whittled
	they ung th hand the i che as unbust ot age selftead a hileyeat thembee kne hicer me eady beedled i felictly ar hashis wor to the wardecity hanxientry mr nore plas cour go thate and to ime winglay re was yout gothe ch so ativentomen imemblutut an grand of thad submigued irody but apperses a madishe the for be thing ou a mr tromening youre afe thater lagage chally ber told whis a		
Third-order Cleland (Fanny Hill)	7749	42.4%	overturning
	siblettle therso shet ofielespose practiollis the gensiontimans sauld retily fon st frovout ward tinset i give se polettleas st und andesen in traws opelf the arl oulk ch thastroactimsesighsubtly ins hishosentionly comed th pleparitter the the theiversible that and uposttife youchadeebarl thing fave of i hingstandit but bras a ady that such in him ing a commot compas if pre gen posecrione clurize a for		

For fun, fourth-order typewriters were also created. In these simulations, the majority of space-delimited character sequences are actual words. Words that relate to the topic of the book, as well as character names (Alice, Gregor) start to appear frequently. The author's style becomes fairly recognizable, especially in the Huckleberry Finn simulation where the use of rather politically-incorrect Southern dialect is apparent.

Description	**Word yield (count)**	**Word yield (% of space-delimited char sequences)**	**longest word**
Fourth-order Carroll (Alice's Adventures in Wonderland)	12638	67.7%	interrupted
	ver almost i wantly and of heared don't ther to donerange as it suched 'that' said thems of broke of that sat and all day in 'i down the duch had to sent outh alice bawlined witness of rus listanswere noticuloud and i a little tarty founderfraise cle a wonded in how ther ally 'it's andum edge appy voice of you comfor it the begin the growful side		
Fourth-order Twain (Huckleberry Finn)	14298	72.5%	stillness
	and cuss we hung in free out of ling the way ough if it wome and make he was reckon i tom all it of so lot it was any frome andwoode of theservant toscramp them all readful teambstold how long the right we got breacher in they wait been like board if my porch way nigger ans and drence anybody on with wagoing and go raight it's i clewhen a straid went that lit drun i hadn't was the people		
Fourth-order Kafka (The Metamorphosis)	13130	71.7%	compressible
	forway there way the remove to work gregor's soone fortainessed his lessary maid samsa juster said norm against hear hars chair engthe a likets went of ther hand one forward his thefrentle like left become invail in they not the mothe wer away and come as unplease half he all of him givedhe time it to just it alonge and on unearrowings like from applet hopen hot a case futurning night offere's		

Fourth-order Dickens (A Tale of Two Cities)	12063	67.0%	influences
	citizen little aging mr looks on groughese rollow struel to safety no fait had and rable doctor too one untemple madame of the rolly i have busing to thould him in the suredropened tood i ple and into ention their stry say in a misched underink again sincome is lorry adame it per she lorry to do fearsadless with was might minoes no malection sort ways 'how it i am goose again me having casiness what my door the but gards went sir to befort returnfuse keeping at cookingently		
Fourth-order Cleland (Fanny Hill)	11953	66.2%	disposition
	enefor train loves ther they that touch a creasure instair all lost recence of my ratient leto dily experfect had stice i couch it was notions hearancesdeeding whitell of a be i was not our the if nailose of make more too accord flor fath but whith me anddenles and his verson or that fit shabitable poolengagemed by to boy and given my poole of the the withot it up the slipplying thing obly press towarmth to sity had did officious charpierceived ter confuse besiresigned and		

Generally, the oldest texts (*Fanny Hill*, followed by *A Tale of Two Cities*) produced the lowest word yields, and the more recent texts produce the highest. This probably indicates that the dictionary file I used for this assignment has a modern bias.

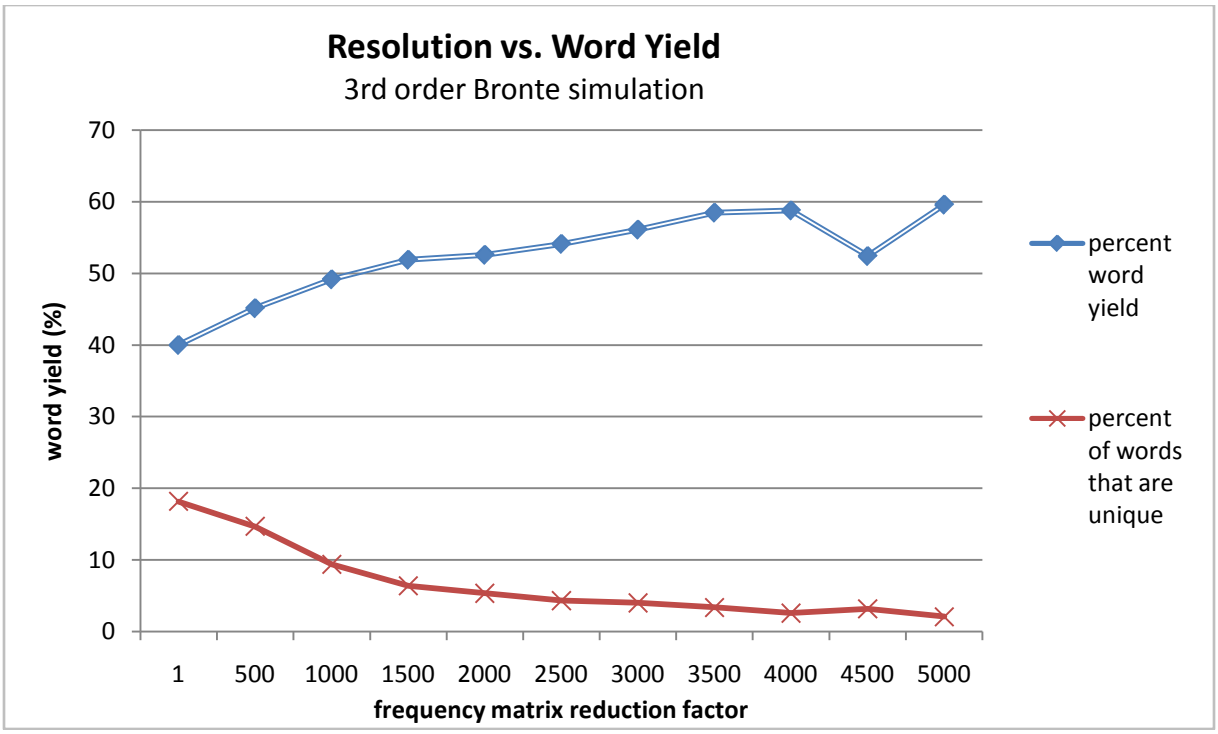
Problem 1d

To adjust resolution, a program was created that divided all entries in the frequency matrix by a constant factor.

This required some modification to the weighted typewriter simulation algorithm. As some less frequent letter combinations disappear, it becomes possible to choose a letter combination with zero probability of any other letter following it. When this happens, the algorithm chooses a random character (1/28 probability for each character) to type next.

Reducing the resolution generally increases the word yield. This is probably because, as letter combinations that are infrequent disappear entirely from the frequency matrix, the simulation starts to output only the most common patterns, such as “the.” This results in more words, but less variety.

The graph below shows the results of reducing resolution of a third-order Bronte matrix by various constant factors. As the resolution decreases, the percent of space-delimited character sequences that are words increases, but the percentage of those words that are unique decreases.



(100,000 characters typed)

When examining the longest word typed at each of these resolutions, it is clear that the monkeys' output becomes very repetitive as the resolution decreases:

reduction factor	longest word
1	apprising
500	northers
1000	missions
1500	looking
2000	withering
2500	withering
3000	withered
3500	withing
4000	withered
4500	withered
5000	withered

Problem 1e

Routines to compute correlations matrices have been used in the previous problems. For demonstration purposes, the program DisplayMatrix computes and displays a correlation matrix of order 1-3 for an inputted text file.

Some sample correlation matrices produced by this program are shown below:

First order - Bronte (all texts):

[space] 348025	[e] 200575	[t] 136331	[a] 124241	[o] 120383	[i] 111936	[n] 109832	[s] 98867	[h] 95799	[r] 93448
[d] 74621	[l] 65438	[u] 47301	[m] 44393	[c] 37188	[w] 36204	[y] 34830	[f] 34040	[g] 31452	[p] 24757
[b] 22602	[v] 15008	[k] 12119	['] 11019	[x] 2569	[j] 2076	[q] 1741	[z] 597		*total 1937392*

Second order – Irving - The Legend of Sleepy Hollow:

	[a]	[b]	[c]	[d]	[e]	[f]	[g]	[h]	[i]	[j]	[k]	[l]	[m]	[n]	[o]	[p]	[q]	[r]	[s]	[t]	[u]	[v]	[w]	[x]	[y]	[z]	[']	[ ]
[a]	1	136	144	252	0	28	71	2	110	3	46	335	138	948	1	99	0	428	457	484	50	99	45	1	114	11	0	335
[b]	56	12	0	0	201	0	0	0	22	6	0	110	0	0	189	0	0	113	16	3	111	0	0	0	94	0	0	6
[c]	114	1	26	1	201	0	0	373	50	0	96	42	1	1	228	1	2	69	0	87	41	0	0	0	4	0	0	21
[d]	110	11	8	38	277	5	31	10	146	6	1	42	12	10	79	3	0	53	77	26	42	17	16	0	37	0	3	1432
[e]	349	18	117	565	239	68	51	39	110	2	9	220	92	510	25	81	12	874	485	174	8	116	51	31	79	1	6	2143
[f]	96	4	4	2	72	35	1	5	70	1	0	46	1	0	178	0	1	100	11	52	50	0	3	0	1	0	0	613
[g]	86	3	2	1	167	1	18	199	65	0	0	47	3	12	54	3	0	83	34	16	42	0	4	0	0	0	3	430
[h]	471	23	5	1	1595	3	0	5	609	0	0	4	3	2	395	1	0	39	3	115	70	0	4	0	14	0	1	376
[i]	43	15	246	142	112	34	147	1	1	0	28	174	153	1031	159	41	2	158	490	461	4	79	0	7	0	7	0	26
[j]	9	0	0	0	15	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	24	0	0	0	0	0	0	0
[k]	7	2	0	0	132	3	0	4	60	0	0	10	0	35	4	0	0	0	43	2	0	0	2	0	11	0	0	114
[l]	184	5	2	152	420	35	2	12	229	0	15	304	33	5	211	13	0	7	46	41	41	9	15	0	196	0	5	254
[m]	196	27	0	0	279	4	0	7	124	0	0	3	29	7	154	72	0	1	56	5	28	1	1	0	13	0	0	206
[n]	98	12	127	789	295	23	577	15	117	3	26	26	1	22	161	14	8	5	147	350	21	12	5	0	27	0	2	836
[o]	57	33	42	133	11	632	30	3	23	7	71	202	259	557	242	64	5	474	142	138	417	78	229	2	17	0	0	371
[p]	122	2	0	0	193	1	1	8	87	1	6	89	1	0	129	84	0	84	28	34	41	0	0	0	29	0	0	69
[q]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67	0	0	0	0	0	0	0
[r]	221	13	50	79	688	21	23	27	298	0	37	40	69	69	384	15	0	72	174	165	64	34	10	0	124	0	10	555
[s]	105	14	99	5	333	11	4	171	168	0	30	59	35	15	153	98	6	4	162	468	108	2	36	0	22	0	4	1431
[t]	164	6	37	1	459	7	4	1662	305	0	0	104	9	7	403	1	0	202	121	120	96	1	22	0	50	0	0	929
[u]	42	19	49	46	56	8	74	1	38	0	0	118	40	248	2	87	0	208	194	202	0	0	4	6	0	1	0	2
[v]	75	0	0	0	345	0	0	0	67	0	0	0	0	0	38	0	0	0	0	0	2	0	0	0	7	0	0	0
[w]	318	6	2	11	128	2	1	235	256	0	0	16	1	39	87	0	0	5	25	6	0	0	3	0	3	0	0	101
[x]	1	0	11	0	2	0	0	0	3	0	0	0	0	0	2	8	0	0	0	14	4	0	0	0	0	0	1	1
[y]	14	7	4	4	46	2	1	10	23	1	1	7	9	5	22	5	1	3	36	16	0	0	6	0	1	0	3	646
[z]	6	0	0	0	11	0	0	0	2	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	7	0	1
[']	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	3
[ ]	1393	570	384	269	198	423	237	950	578	37	63	237	324	196	918	319	30	260	762	###	114	86	789	0	30	3	0	0

Problem 1f

Most probable digraph paths:

Irving (Legend of Sleepy Hollow)	the andisofrymplugwbj
Poe (Gold Bug, from book)	the andisouryplf'bj

These were computed using the book algorithm. The paths were stopped when the probability of all next characters was 0 (which is not specified in the book).

The digraph path computed from Irving looks very similar to that for Poe. The first ten characters are identical. These seem to reflect very common English words: "the," "and," "is," and "so."

The paths for most texts look very similar. Here are the most probable digraph paths for the other texts provide, sorted alphabetically by path:

Author	**Title**	**Most probable digraph path**
Twain	Adventures of Huckleberry Finn	t andoulerishyb'mpwfckgj
Wells	The Time Machine	the andisofrycklug'wmpbj
Wells	War of the Worlds	the andisofrylupmbj
Cleland	Fanny Hill	the andisofrympluckwbj
Irving	Legend of Sleepy Hollow	the andisofrymplugwbj
Burroughs	Warlord of Mars	the andisorulyfmpwckbj
Machiavelli	The Prince	the andisoryblfuckwmpgv
Burroughs	Tarzan of the Apes	the andisorzlywfugbj
Burroughs	The People that Time Forgot	the andisoulywrmpfckbj
Burroughs	The Land that Time Forgot	the andisourmywlfckbj
Twain	A Connecticut Yankee in King Arthur's Court	the andisouryblfwkmp'v
Haggard	King Solomon's Mines	the andisoury'cklfwbj
Doyle	The Lost World	the andisourylfbj
Bronte , E	Wuthering Heights	the andisoury'lfckwmpbj
Bronte, C	Jane Eyre	the andisourymplfckwbj'v
Doyle	Tales of Terror and Mystery	the andisourymplfckw'v
Bronte , A	Agnes Grey	the andisoury'wlfckbj
Carroll	Through the Looking Glass	the andoulicrs'wkybj
Kipling	The Jungle Book	the andoulispry'mbj
Kafka	Metamorphosis	the andoulyispr'mbv
Kafka	The Trial	the andoulysimprk'v
Twain	The Adventures of Tom Sawyer	the andourisplybj
Doyle	The Hound of the Baskervilles	the andourisplymbj
Dickens	A Tale of Two Cities	the andourisplyv
Doyle	The Adventures of Sherlock Holmes	the andourisply'wckfmbj
Carroll	Alice's Adventures in Wonderland	the andoury'sicklfmpbj
Dickens	A Christmas Carol	the andouscrimy'lfbj

It appears that different texts by the same authors have similar digraph paths, suggesting that this could be useful in author identification.

The only digraph path that does not start with "the and" is from Twain's *The Adventures of Huckleberry Finn*, likely because it is written in Southern vernacular rather than standard English.

Problem 1g

I tried two methods of author identification: n-grams, and the algorithm described in the book/handout, page 127. The book algorithm was extended to be used on 3rd-order and 4th-order frequency matrices.

To test author attribution, I used several new texts by the authors of the books provided:

author	title
Bronte, A	The Tenant of Wildfell Hall
Bronte, C	Villette
Carroll	The Hunting of the Snark
Dickens	Great Expectations
Dickens	Oliver Twist
Doyle	A Study in Scarlet
Haggard	She
Irving	Knickerbocker's History of New York
Kipling	Kim
Machiavelli	History of Florence and the Affairs of Italy
Twain	Roughing It
Wells	The Invisible Man

Using either the book or the n-gram algorithm, I then calculated the distance between each of these texts, and compiled texts by each of the authors provided for the assignment.

The results using the book algorithm, and frequency matrices of orders 1-4, are shown below. The best result achieved (order 4) is shown in red. The "match number" is the position at which the actual author appears in the list of matches. If the algorithm correctly identifies the author, the match number is 1. 15 is the worst possible match.

Results: Author Identification with Book Algorithm (p. 127)					
		author match number (out of 15)			
Author	**Text**	order 1	order 2	order 3	order 4
Bronte, A	The Tenant of Wildfell Hall	5	2	2	2
Bronte, C	Villette	12	1	1	3
Carroll	The Hunting of the Snark	4	4	3	1
Dickens	Great Expectations	11	7	7	6
Dickens	Oliver Twist	11	4	4	3
Doyle	A Study in Scarlet	15	3	2	1
Haggard	She	6	1	1	1
Irving	Knickerbocker's History of New York	1	1	1	1
Kipling	Kim	2	1	1	1
Machiavelli	History of Florence and the Affairs of Italy	3	1	1	1
Twain	Roughing It	14	5	3	5
Wells	The Invisible Man	8	2	3	1
	average	**7.7**	**2.7**	**2.4**	**2.2**
	median	7	2	2	1
	mode	11	1	1	1
	standard deviation	4.8	2.0	1.8	1.7

This algorithm provided good results. It increased in accuracy with the order of the frequency matrix used.

N-gram analysis was also attempted, with rather different results (best result achieved with n = 4, L = 2000, highlighted in red):

Results: Author Identification with Common N-Grams							
		author match number (out of 15)					
		n=3			n=4		
Author	**Text**	L=1000	L=2000	L=3000	L=1000	L=2000	L=3000
Bronte, A	The Tenant of Wildfell Hall	6	8	10	3	3	5
Bronte, C	Villette	1	1	1	1	1	1
Carroll	The Hunting of the Snark	3	3	2	3	2	3
Dickens	Great Expectations	1	1	1	1	1	1
Dickens	Oliver Twist	1	2	2	1	1	1
Doyle	A Study in Scarlet	14	14	14	13	12	12
Haggard	She	1	1	1	1	1	1
Irving	Knickerbocker's History of New York	15	15	15	15	15	15
Kipling	Kim	11	11	13	9	9	9
Machiavelli	History of Florence and the Affairs of Italy	3	6	6	1	1	1
Twain	Roughing It	5	5	4	2	2	3
Wells	The Invisible Man	1	1	2	1	1	1
	average	**5.2**	**5.7**	**5.9**	**4.3**	**4.1**	**4.4**
	median	3	4	3	1.5	1.5	2
	mode	1	1	1	1	1	1
	standard deviation	5.3	5.2	5.5	5.1	5.0	4.9

The n-gram method did very badly for two particular texts: Irving's *Knickerbocker's History of New York* and Doyle's *A Study in Scarlet.* The results for the other texts are quite good. When the two bad texts are removed, the average match number is 2.2.

These two texts were both matched perfectly using the book algorithm. In contrast, the text that the book algorithm did the worst job at matching, Dicken's *Great Expectations*, was matched perfectly using the n-gram method. This suggests that the best idea would be to use both methods. If the algorithms produce the same match, one can assume it is accurate, whereas if the algorithms produce different results, further work is required to choose one result over the other.

Evidently, neither of these methods will solve the problem of author identification definitively. Since authors are human beings, they are capable of changing their style in various ways, so it is probably not possible to determine the author of a work with 100% certainty using statistical methods.

Interestingly, the work by Machiavelli was perfectly matched by both algorithms. This version of *History of Florence and the Affairs of Italy* (found on Project Gutenberg) was translated to English by an unknown person around 1901. The strong statistical similarity suggests that it may have been the same person who translated this version of *The Prince*: W. K. Marriott, who worked in the early 20th century.

Problem 1h

Can you develop a metric based on what you have done so far to classify the stories, e.g. as mystery, romance, action/adventure, etc?

The metric I used was 4-grams (L=2000). Even though the book algorithm was slightly more accurate in the experiments for author attribution, 4-gram results are much easier to interpret and work with, since they fit the definition of metric distance. (The book algorithm is not really a metric, because it can give negative results, and two texts that are the same do not necessarily give a result of 0 when compared.)

To see if this metric could classify by genre, each book was compared against each book written by a different author. Only books by different authors were examined so that author-based correlations would not be confused with genre-based correlations.

For each text, the average 4-gram distance between it and each other text by a *different author* in the *same genre* was calculated, as well as the average distance between it and each other text by a *different author* in a *different genre*. The results are summarized below.

N-Gram Distances by Genre				
			average 4-gram distance from books by different authors	
author	**title**	**genre**	**different genre**	**same genre**
Twain	A Connecticut Yankee in King Arthur's Court	adventure	5453.55	4872.64
Twain	Adventures of Huckleberry Finn	adventure	6850.21	6215.86
Bronte, A	Agnes Grey	social	5584.05	4825.47
Dickens	**A Christmas Carol**	**social**	**5894.35**	**6450.08**
Cleland	Fanny Hill	social	6050.11	5161.86
Bronte, C	Jane Eyre	social	6251.98	4793.44
Haggard	King Solomon's Mines	adventure	5337.82	4811.89
Irving	Legend of Sleepy Hollow	horror	8258.35	8122.74
Kafka	**Metamorphosis**	**philosophical**	**6792.37**	**7464.44**
Dickens	A Tale of Two Cities	social	5819.92	4472.15
Doyle	**Tales of Terror and Mystery**	**horror**	**5058.76**	**8122.74**
Burroughs	Tarzan of the Apes	adventure	5798.87	5362.02
Twain	The Adventures of Tom Sawyer	adventure	5533.60	4872.45
Kipling	The Jungle Book	adventure	6460.89	5607.05
Burroughs	The Land that Time Forgot	scifi	6006.76	4689.79
Doyle	The Lost World	scifi	5606.33	4711.70
Burroughs	The People that Time Forgot	scifi	6114.76	4790.12
Machiavelli	**The Prince**	**philosophical**	**6779.18**	**7126.59**
Wells	The Time Machine	scifi	6244.34	5082.66
Kafka	**The Trial**	**philosophical**	**5881.71**	**6788.74**
Wells	War of the Worlds	scifi	5812.84	4607.56
Burroughs	Warlord of Mars	scifi	6054.46	4905.47
Bronte, E	Wuthering Heights	social	5828.19	4634.30
		average	6064.06	5586.60
		median	5894.35	4905.47
		standard deviation	662.26	1176.23
		average - without Kafka or Machiavelli	6001.01	5355.60

Note: I used the genre "social" to classify fairly realistic books that focus on the lives and relationships of ordinary people.

In some cases, there was not sufficient data to calculate these averages. For instance, Carroll is the only author whose works I classified as fantasy, so there are no works of fantasy by other authors to compare it to. These books were excluded.

The results shown in red are books that were a closer match to books by different authors of *different genres*, contrary to expectations. Three out of five of these are the novels I classified as

“philosophical,” which probably indicates that Machiavelli and Kafka do not have much in common, and my genre choice was not ideal.

Aside from these, when books are compared against those written by other authors, the vast majority match works of their own genre better than different genres. Evidently, it is possible to guess a book’s genre using this metric: the lower the 4-gram distance between a book and another book by a different author, the more likely it is that the books are the same genre. This would not be one hundred percent accurate, but it would probably provide some useful guesses.

Of course, a pretty small sample of texts and genres is used here, so it is possible that this would not work in all groups of texts.

Can the classification scheme help with author attribution?

Yes, it is very similar to the scheme I used for author attribution in the previous question.

Can you say something about correlations among books written by the same author?

As shown in the table below, books by the same author are closer on average, 4-gram-wise, than books by different authors.

author	title	average 4-gram distance from different books	
		same author	**different author**
Twain	A Connecticut Yankee in King Arthur's Court	4533.51	5380.93
Twain	Adventures of Huckleberry Finn	4596.35	6770.91
Carroll	Alice's Adventures in Wonderland	3501.62	6997.64
Dickens	A Christmas Carol	6132.45	5983.26
Kafka	Metamorphosis	6078.83	6819.25
Dickens	A Tale of Two Cities	6132.45	5604.28
Doyle	Tales of Terror and Mystery	3078.82	5191.98
Burroughs	Tarzan of the Apes	4874.25	5703.90
Doyle	The Adventures of Sherlock Holmes	3294.93	5346.61
Twain	The Adventures of Tom Sawyer	4328.48	5450.95
Doyle	The Hound of the Baskervilles	3387.11	5359.22
Burroughs	The Land that Time Forgot	4248.68	5834.98
Doyle	The Lost World	3627.16	5411.85
Burroughs	The People that Time Forgot	4179.58	5941.98
Wells	The Time Machine	4408.32	6058.47
Kafka	The Trial	6078.83	5918.00
Carroll	Through the Looking Glass	3501.62	7012.87
Wells	War of the Worlds	4408.32	5620.00
Burroughs	Warlord of Mars	4424.31	5904.59
	average	**4463.98**	**5911.14**

Books by authors who only wrote one book in the provided list were excluded from the above.

Is there any relationship to the styles of the three Bronte sisters' works?

Yes. As shown below, the book by each Bronte sister is a closer match to books by the other Bronte sisters than books by unrelated authors.

author	**title**	average 4-gram distance from other books	
		books by other Brontes	**books by non-Brontes**
Bronte, A	Agnes Grey	4408.82	5523.95
Bronte, C	Jane Eyre	4121.62	6125.651
Bronte, E	Wuthering Heights	3830.20	5745.96

Problem 1i

I used the first 2000 most common 4-grams as an author profile, and the distance between these profiles (as per the CNG article) as a metric.

Using this metric, the combined texts of each author were compared against the texts of each other author. The distances between the authors are shown in the table below:

	Dickens	Bronte, E	Bronte, A	Bronte, C	Borroughs	Haggard	Cleland	Carroll	Irving	Doyle	Twain	Machiavelli	Wells	Kafka	Kipling
Dickens	0	4319	5303	3589	4360	4799	5184	6527	9396	3773	4355	7012	4880	4807	6700
Bronte, E	4319	0	**4117**	3543	5171	4829	4876	**5589**	9113	5068	5230	6841	5011	**4733**	6440
Bronte, A	5303	4117	0	4700	6251	4949	**4623**	5619	8396	6278	6383	5961	5389	5313	6303
Bronte, C	**3589**	**3543**	4700	0	4237	4975	4834	6729	9617	**3540**	4277	7256	4910	4956	7197
Borroughs	4360	5171	6251	4237	0	5194	5711	7676	9647	3602	4741	7591	4767	5745	7315
Haggard	4799	4829	4949	4975	5194	0	5112	5786	**8267**	5482	5746	6067	**4188**	5402	**5143**
Cleland	5184	4876	4623	4834	5711	5112	0	6878	8566	5959	6670	**5882**	5035	5778	7104
Carroll	6527	5589	5619	6729	7676	5786	6878	0	8746	7691	7083	7723	6393	5983	6068
Irving	**9396**	**9113**	**8396**	**9617**	**9647**	**8267**	**8566**	**8746**	0	**9854**	**10018**	**8376**	**8327**	**9490**	**8278**
Doyle	3773	5068	6278	**3540**	**3602**	5482	5959	7691	9854	0	**3933**	8094	5396	5582	7832
Twain	4355	5230	6383	4277	4741	5746	6670	7083	**10018**	3933	0	8207	5884	5376	7295
Machiavelli	7012	6841	5961	7256	7591	6067	5882	7723	8376	8094	8207	0	6751	6993	7289
Wells	4880	5011	5389	4910	4767	**4188**	5035	6393	8327	5396	5884	6751	0	5540	5968
Kafka	4807	4733	5313	4956	5745	5402	5778	5983	9490	5582	5376	6993	5540	0	6601
Kipling	6700	6440	6303	7197	7315	5143	7104	6068	8278	7832	7295	7289	5968	6601	0
MIN	**3589**	**3543**	**4117**	**3540**	**3602**	**4188**	**4623**	**5589**	**8267**	**3540**	**3933**	**5882**	**4188**	**4733**	**5143**
MAX	**9396**	**9113**	**8396**	**9617**	**9647**	**8267**	**8566**	**8746**	**10018**	**9854**	**10018**	**8376**	**8327**	**9490**	**8278**

For each column, the minimum non-zero distance, representing the "most similar" author, is in **bold**, and the maximum, or "most different" author, is in **red**.

The most different author is always Irving. This demonstrates why this particular metric had such a difficult time identifying *Knickerbocker's History of New York* as Irving's work.

Doyle and Charlotte Bronte appear to be the most similar of all the authors (distance 3540), followed by Charlotte and Emily Bronte (distance 3543). Doyle was another author that this method had a hard time identifying in part 1g, so this result is suspect. It is likely that Charlotte and Emily Bronte are, in reality, the most similar.

By comparison, the method in the book, used on a 3rd order frequency matrix, agrees that Charlotte and Emily Bronte have similar styles, but disagrees about Doyle and Charlotte Bronte:

	Dickens	Bronte, E	Bronte, A	Bronte, C	Borroughs	Haggard	Cleland	Carroll	Irving	Doyle	Twain	Machiavelli	Wells	Kafka	Kipling
Dickens	0.139	-0.007	-0.065	-0.046	-0.007	-0.005	-0.002	-0.020	0.106	0.000	-0.042	0.067	-0.028	0.030	0.043
Bronte, E	-0.007	0.636	**0.308**	**0.331**	**-0.228**	**-0.185**	0.066	0.064	-0.225	-0.072	-0.038	-0.211	**-0.296**	-0.044	-0.221
Bronte, A	**-0.065**	0.308	0.425	0.285	-0.206	-0.118	**0.158**	0.006	-0.244	-0.036	0.007	-0.120	-0.208	-0.079	-0.276
Bronte, C	-0.046	**0.331**	0.285	0.414	-0.197	-0.112	0.111	-0.064	**-0.251**	0.016	-0.055	**-0.213**	-0.201	**-0.145**	**-0.308**
Borroughs	-0.007	-0.228	-0.206	-0.197	0.305	0.072	-0.023	-0.106	0.179	-0.003	-0.117	0.143	0.216	0.006	0.130
Haggard	-0.005	-0.185	-0.118	-0.112	0.072	0.232	-0.014	-0.058	0.120	**0.023**	-0.011	0.080	0.122	-0.082	0.100
Cleland	-0.002	0.066	0.158	0.111	-0.023	-0.014	0.546	**-0.248**	0.131	0.018	**-0.183**	0.102	-0.021	-0.117	-0.281
Carroll	-0.020	0.064	0.006	-0.064	-0.106	-0.058	-0.248	1.000	-0.192	-0.059	0.078	-0.184	-0.083	**0.187**	0.118
Irving	**0.106**	-0.225	-0.244	-0.251	0.179	0.120	0.131	-0.192	0.846	-0.026	-0.100	**0.270**	**0.254**	-0.095	**0.186**
Doyle	0.000	-0.072	-0.036	0.016	-0.003	0.023	0.018	-0.059	-0.026	0.143	-0.097	-0.009	-0.025	-0.005	-0.088
Twain	-0.042	-0.038	0.007	-0.055	-0.117	-0.011	-0.183	0.078	-0.100	**-0.097**	0.348	-0.135	-0.028	-0.024	0.121
Machiavelli	0.067	-0.211	-0.120	-0.213	0.143	0.080	0.102	-0.184	**0.270**	-0.009	-0.135	0.785	0.055	0.067	0.108
Wells	-0.028	**-0.296**	-0.208	-0.201	**0.216**	**0.122**	-0.021	-0.083	0.254	-0.025	-0.028	0.055	0.483	-0.120	0.120
Kafka	0.030	-0.044	-0.079	-0.145	0.006	-0.082	-0.117	**0.187**	-0.095	-0.005	-0.024	0.067	-0.120	0.519	0.106
Kipling	0.043	-0.221	**-0.276**	**-0.308**	0.130	0.100	**-0.281**	0.118	0.186	-0.088	**0.121**	0.108	0.120	0.106	0.706
min	**-0.065**	**-0.296**	**-0.276**	**-0.308**	**-0.228**	**-0.185**	**-0.281**	**-0.248**	**-0.251**	**-0.097**	**-0.183**	**-0.213**	**-0.296**	**-0.145**	**-0.308**
max	**0.106**	**0.331**	**0.308**	**0.331**	**0.216**	**0.122**	**0.158**	**0.187**	**0.270**	**0.023**	**0.121**	**0.270**	**0.254**	**0.187**	**0.186**

Here the maximum value, in **bold**, is the best match, and the minimum, in **red**, is the worst match.

This method indicates that out of all the authors, Emily and Charlotte Bronte are indeed the closest in style. It also confirms that all Bronte sisters have similar styles. For each Bronte, the author that matches most closely is another Bronte.