EECS6339 3.0 Introduction to Computational Linguistics
Instructor: Nick Cercone – 3050 LAS – nick@cse.yorku.ca
TYuesdays, Thursdays 10:00-11:20 – LAS 3033
Winter Semester, 2015

THE BIG ASSIGNMENT

This assignment requires the writing of a series of programs, the collection of which can provide a package of routines useful to humanities researchers or curiosity seekers. A web-based presentation implementation available over the internet merits extra credit.

Figure. Things are looking up!

To make this assignment uniform across all students, we first we will have to decide how many characters we really need. Let us ignore the distinction between upper and lower case so count 26 alphabetic characters, the 12 punctuation marks comma “,”, period “.”, semicolon “;”, colon “:”, question mark “?”, explanation mark “!”, parentheses “(“ and “)”, hyphen “-”, single quote “‘”, double quote ““““, and everything else “@”, and count any number as “#”, and the blank for a total of 40 characters.

Problem 1a
Simulate the straightforward monkey problem. Let the program run long enough to give a meaningful estimate of the yield of words. The result will provide a useful comparison with later forms of the problem.

Problem 1b
Use the data in Table 1 to simulate the first-order monkey problem. Again let the program run long enough to give a meaningful estimate of the yield of words to permit comparison with other results on relative word yield. Try running this simulation program against other corpora.

Problem 1c
Use the data supplied, data listed in Table 2, to simulate the second-order and third order Bronte monkey problem. Again let the program run long enough to give a meaningful estimate of the yield of words to permit comparison with other results on relative word yield. Try running this simulation program against other authors listed.

Space 6934	E 3277	O 2578	T 2557	A 2043	S 1856	H 1773	N 1741	I 1736	R 1593
L 1238	D 1099	U 1014	M 889	Y 783	W 716	F 629	C 584	G 478	P 433
B 410	V 309	K 255	‘ 203	J 34	Q 27	X 21	Z 14		

Table 1. Character Distribution from Act III of Hamlet (in order of decreasing frequency) –
Note: 35,224 characters, a small corpus.

Problem 1d
Investigate the effects of resolution on monkey literacy in the simulation. For example round off the matrix elements to the smallest number of places - or use an equivalent means to reduce the number of keys on the typewriters.

Problem 1e
Write a routine to compute correlation matrices of the type shown in the handout from data supplied (the books shown in Table 2). (The Bronte sisters shown on the right)

Author	Title	Size
Charles Dickens	A Christmas Carol	184 K
Charles Dickens	A Tale of Two Cities	775 K
Emily Bronte	Wuthering Heights	666 K
Anne Bronte	Agnes Grey	389 K
Charlotte Bronte	Jane Eyre	1 MB
Edgar Rice Burroughs	Tarzan of the Apes	500 K
Edgar Rice Burroughs	Warlord of Mars	331 K
Edgar Rice Burroughs	The People that Time Forgot	226 K
Edgar Rice Burroughs	The Land that Time Forgot	220 K
H. Ryder Haggard	King Solomon's Mines	463 K
John Cleland	Fanny Hill	483 K
Lewis Carroll	Alice's Adventures in Wonderland	164 K
Lewis Carroll	Through the Looking Glass	182 K
Washington Irving	Legend of Sleepy Hollow	87 K
Sir Arthur Conan Doyle	The Adventures of Sherlock Holmes	589 K
Sir Arthur Conan Doyle	The Lost World	459 K
Sir Arthur Conan Doyle	The Hound of the Baskervilles	346 K
Sir Arthur Conan Doyle	Tales of Terror and Mystery	430 K
Mark Twain	Adventures of Huckleberry Finn	583 K
Mark Twain	The Adventures of Tom Sawyer	406 K
Mark Twain	A Connecticut Yankee in King Arthur's	661 K
Nicolo Machiavelli	The Prince	299 K
H. G. Wells	War of the Worlds	357 K
H. G. Wells	The Time Machine	197 K
Franz Kafka	Metamorphosis	138 K
Franz Kafka	The Trial	463 K
Rudyard Kipling	The Jungle Book	292 K
Oscar Wilde	The Importance of Being Earnest	137K
Oscar Wilde	Picture of Dorian Gray	448K

Table 2. Data made available for this Assignment.

Problem 1f
Using the algorithm in the handout with the pair-correlation matrix generated from Irving (book shown in Table 2), compute the most probable digraph path, which starts with the letter T. Compare the result with that given in the handout for Poe's "The Gold Bug".

Problem 1g
Design and implement an experiment using data from the books shown in Table 2 that might be used to perform author attribution. Discuss your solution and provide reasons why it is likely or not likely to solve the problem definitively.

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.? Implement your techniques to demonstrate classification. Can the classification scheme you designed help with author attribution? Can you say something about correlations among books written by the same author? Is there any relationship to the styles of the three Bronte sisters' works?

Problem 1i
Develop a profile for each of the different authors in Table 2 and provide a metric and argument that compares and contrasts authors in order to speculate which two authors are the most "similar" in style.

Problem 1j
How would you characterize the major differences between the "style" of Oscar Wilde, Franz Kafka, H. G. Wells, Mark Twain, and Charles Dickens?