**CSCI8360 Project 1**

**Due on Blackboard: 11:59PM Sept 14th, 2016**

**There is one submission per team. You are free to choose your own teams. Every student must be in a team. You should create a directory “8360-Proj1” (exactly as shown) in your local directory on Loki and copy the final version of all your files to that directory before the submission time. Your submission must include a “readme.txt” file that includes the following information –**

**Names of the students on the team, which student implemented what, how to compile the program, how to run the program.**

**Grading Criterion: Max points: 50 points**

**Program processes the input files and runs without errors: 50 points**

**Program processes the input files and executes at most 75% of the commands: 35 points**

**Program does not process the input files, but compiles: 25 points**

**Program does not compile: 0 points.**

**No readme file in the submission: 0 points.**

**\*\*\*\*\*\*\*\*\*\*\*\*Fall**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Goal: Implement the entropy, conditional entropy, mutual information, and the KL-distance metrics on a small subset of the 20 News group (20 NG) data.

Given: You will be given a small subset of pre-processed 20NG data which is formatted as follows –

1. There is one news article per line in the file.
2. Each line is a sequence of stemmed words.
3. The first word in each line is the name of the news group, followed by all of the keywords in that news article.

You will also be given another file – “input.txt” containing one command per line. Your program should process the commands one at a time until EOF is reached.

The structure of the command is as follows.

Command\_num, news group name, <keyword list separated by spaces>

Command\_num is 1 for entropy, 2 for conditional entropy, 3 for mutual information, and 4 for KL distance. Name of the news group will be “all” to process the command for all news groups.

Probability of a word w = the number of lines containing “w” divided by the total number of lines.

Joint probability P(x, y) = the number of lines containing both “x” and “y” divided by the total number of lines.:w

Conditional probability P(x | y) = P(x, y)/P(y)

Examples of commands

1, all, door univers strong support scroll

This command computes the entropy of the given list of keywords in each of the news groups.

2, alt.atheism, support scroll strong: first value is y and all other values are x

This command computes the conditional entropy H(“support” | {“strong”, “scroll”}) only in lines that belong to the new group alt.atheism.

3, all, support univers :H(x)-H(x|y)

This command computes the mutual information of keywords “univers” and “support” in all news groups.

4, all, gamma law d

This command computes the KL-distance between the probability distribution of keywords “gamma” and “law”.

Note that command 4 will only appear with “all” as the new group name.

The probability distribution of “gamma” is computed as follows. Suppose it appears in news group 1 1 time, news group 2 2 times, …, news group 20 20 times. Then the total number N of times it appears in the data set = (1+2+3+4+…+20) = 210. Then, probability that “gamma” appears in news group 1 = 1/210, and news groups 2 is 2/210 etc.

We compute the probability distribution of “law” in the same way over all news groups and then use the KL-divergence formula to compute the distance.