LING 570: Questionnaire
Due: 5pm on Oct 1
Time: 3 hours

Submit via CollectIt or you can
leaving a hardcopy in Fei's mailbox

Your Name: Wee Teck Tan

Part I (30 points): Probability theory

1. (4 points) When you flip a coin, the probability of getting the head is 0.1. Now suppose you flip the coin three times, what is the probability of getting two heads out of the three flips?

   P[HHT] + P[HTH] + P[THH]
   = (0.1*0.1*0.9) + (0.1*0.9*0.1) + (0.9*0.1*0.1)
   = 0.027

2. (16 points) There are two random variables X and Y, and the joint probability P(X,Y) is shown below:

	X=0	X=1
Y=0	0.32	0.48
Y=1	0.08	0.12

   a. What is the probability distribution for P(X)?

      $P(X_0) = 0.32 + 0.08 = 0.4$

      $P(X_1) = 0.48 + 0.12 = 0.6$

   b. What is the probability distribution for P(Y)?

      $P(Y_0) = 0.32 + 0.48 = 0.8$

$P(Y_1) = 0.08 + 0.12 = 0.2$

c. What is the probability distribution for P(Y|X)?

$P(Y_0|X_0) = P(X_0,Y_0) / P(X_0) = 0.32/0.4 = 0.8$

$P(Y_1|X_0) = P(X_0,Y_1) / P(X_0) = 0.08/0.4 = 0.2$

$P(Y_0|X_1) = P(X_1,Y_0) / P(X_1) = 0.48/0.6 = 0.8$

$P(Y_1|X_1) = P(X_1,Y_1) / P(X_1) = 0.12/0.6 = 0.2$

d. Are X and Y independent? Why or why not?

Yes they are independent.

Proof: from the above we saw that

$P(Y_0|X) = P(Y_0)$
$P(Y_1|X) = P(Y_1)$

3. (10 points) There are three coins: c1, c2, and c3. When tossing a coin once, the probabilities of getting a head for c1-c3 are 0.1, 0.4, and 0.7, respectively. Now assume that you pick one of the coins, with the probability 0.2 of being c1, 0.5 of being c2, 0.3 of being c3.

   a. (5 points) If you toss this coin once, what is the probability of getting a head?

   P[Head] = P[C1.Head] + P[C2.Head] + P[C3.Head]

   = (0.2*0.1) + (0.5*0.4) + (0.3*0.7)
   = 0.43

   b. (5 points) If you toss this coin once and get a head, what is the probability that c1 was picked for this toss?

   P[C1|Head] = (P[C1] and P[Head]) /P[Head]

= (0.2*0.1) / 0.43
= 0.047

Part II (25 points): Formal grammars and formal languages

(1) (5 points) Write a regular grammar that generates the language {a*b*}.

E={a,b}
V={S,A}
S => aS
S => bA

(2) (5 points) Write a context-free grammar that generates the language $\{a^n b^n\}$.

E={a,b}
V={S,A}
S => ab
S => aSb

(3) (5 points) Write a context-free grammar that generates the language {x y | x is a string over {a,b,c}, y is a reverse of x}.

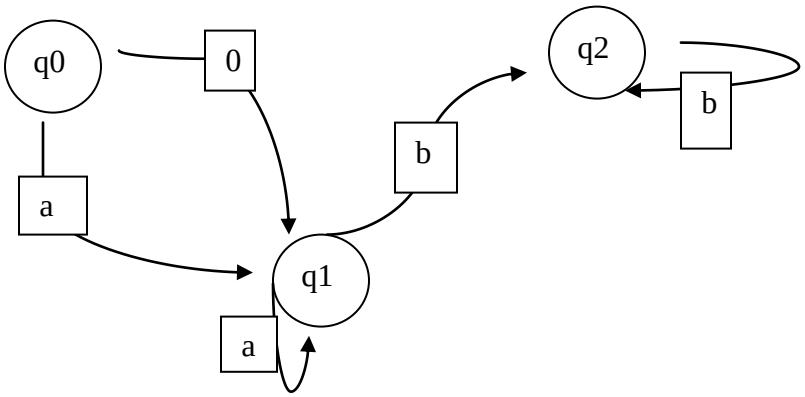
E={a,b,c}
V={S,A,B,C}
S => aBS
S => bCS
S => cBS
A => a
B => b
C => c

(4) (10 points) Write a context-free grammar for the language consisting of all strings over {a,b} for which the number of a's is more than the number of b's.

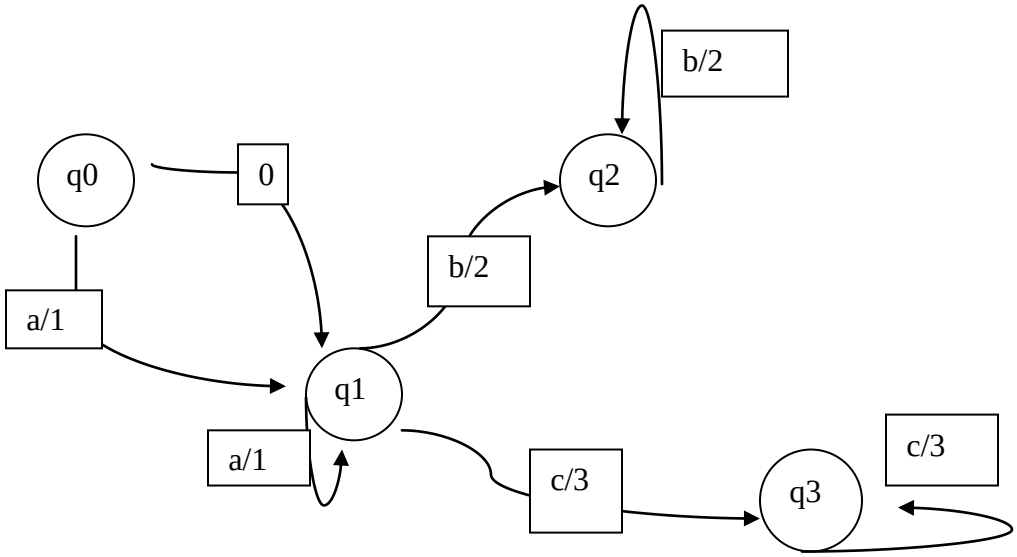
E={a,b,c}
V={S,A,B}
S => A
A => BaA
A => a
B => ab
B => ba

Part III (20 points): FSA and FST

1. (6 points) Draw a finite state automaton (FSA) that will accept all the strings that start with zero or more “a”, followed by one or more “b”, but nothing else.



2. (9 points) Draw a finite state transducer (FST) that will accept all strings that contain zero or more "a", followed by one or more "b" or "c", and the output would replace all the a's in the input string with "1", all the b's with "2", and all the c's with "3".



3. (5 points) What is the relation between regular grammar, regular language, and FSA?

A regular grammar is a characterization of regular sets such that a language is considered regular if and only if it has a left linear grammar and a right linear grammar. We can therefore represent or construct a FSA based on the set generated by a regular grammar.

Part IV (10 points): Write down the unix/linux commands for the following. The command line can include pipes: e.g., "cat input_file | ls | grep tt > output_file"

(a) (2 points) Change the permission of the current directory so that all the files under the current directory are readable by all, writeable and executable only by the owner.

chmod 344 *.*

(b) (2 points) copy everything (including subdirectories) under directory "dir1" to "dir2".

cp –R /dir1 /dir2

(c) (6 points) Given a file, file1, as input, find the number of unique lines in the file. For instance, support the file contains the following lines:
This book
Book
This book
Try this

Given this file, your command line should return "3"

grep –c [A-Za-z] file1

Part V (15 points): Basic linguistics

1. (5 points) Tag the sentence *this is the biggest mistake that he has made so far* with Part-of-speech (POS) tags such as N (noun), V (verb), PN (pronoun), Adj (adjective), Adv (adverb), Prep (preposition), CL (complementizer), CC (conjunction), etc.

(PN this) (V is) (DT the) (Adj biggest) (N mistake) (CL that) (PN he) (V has) (V made) (CC so) (Adj far)

2. (5 points) Draw a parse tree (i.e., syntactic structure) for the sentence *John said that he would call Mary tomorrow*

S = > (NP) (VP)

NP => (N John)
VP => (V said)(AVP)
AVP => (CL that)(NP)(VP)
NP => (N he)
VP => (V would)(VP)
VP => (V call)(NP)
NP => (N Mary)(NP)
NP => (N tomorrow)

(S (N John) (V said) ((CL that) (N he) ((V would) ((V call) ((N Mary) (N tomorrow)))))

3. (5 points) Write down the results of tokenizing: How many tokens are there in the following sentence and what are they?

John said: "I bought the book for $93.05. It's too expensive!"

John	**1**
said	**1**
:	**1**
I	**1**
bought	**1**
the	**1**
book	**1**
for	**1**
$93.05	**1**
.	**1**
It's	**1**
too	**1**
expensive	**1**
!	**1**
“	**1**
”	**1**

16 tokens in total

Part VI: Background information

(a) Have you used CollectIt before? **Yes**

(b) Have you used GoPost before? **Yes**

(c) Have you used GradeBook before? **No**

(d) Have you used Condor Submit before? **Yes**

(e) Do you have a patas account already? **Yes**

(f) What programming languages do you plan to use for LING 570? How long have you been using this language? **Perl (4 months)**

(g) Have you taken any NLP (natural language processing) or machine learning courses before? If so, what are they? **No**

(h) Your familiarity with the following topics (1: have never heard about it, 2: know very little, 3: know the basics, 4: know it well):

- FSA/FST - **2**
- Language model - **2**
- Hidden Markov Model (HMM) - **1**
- POS tagging - **2**
- Feature vectors - **1**
- Mallet - **1**
- Classification algorithms - **2**
- Gaussian distribution - **1**
- Multinomial distribution - **1**
- Entropy and perplexity - **1**