Name: ______________________________ StudentID: ___________________

Assignment 2

CS 173 – 2019 SP [100 points]

1. (25 points) **N-grams.** Compute and store all unigram, bigram *and trigram* frequencies for the Brown corpus, then answer the following questions.

   How many of each are there (i.e., *distinct*): **unigrams** | **bigrams** | **trigrams**

   Having saved each table to separate files, how large are they (records and file size)?

   Discuss their variation in terms of sparseness:

   Examine http://books.google.com/ngrams and check out the raw data. Explain: why might it be absurd to compute 5-grams (or, say, 9-grams) on the Brown corpus?

   List the top five bigrams and their frequencies:

   List the frequencies of the following phrases (case-sensitively):
   the President: the Russian: boiled haddock:

   Compute and justify[1] the most likely word(s), [x], indicated for each phrase:

   ... ran the [x] ...

   ... [x] drinks ...

   ... in the [x] ...

2. (50 points) **Parts of speech.** Do all the work required and complete Figure 5.18 of the text *from scratch*. That is, recompute Fig. 5.15 and 5.16 *from scratch* using the Brown corpus, showing *all frequency counts* and resulting probabilities. You will know if your work is correct if your probabilities match Fig 5.15 and 5.16 closely. Examples are provided to help you get started. For each $v_t(j) = 0$, you should omit outgoing arrows.

[1] Yes, this is worded vaguely on purpose. Part of the effort here is for you to figure out what it takes to *sufficiently justify* your answer. So, read the book and *think critically*! Discuss with classmates. Etc...

Name: ________________________________ StudentID: ____________________

Fig 5.15 (priors)	**VB**	**TO**	**NN**	**PPSS**
<s>				
VB (33693)	0.0038 (130)			
TO				
NN				
PPSS				

Fig 5.16 (likelihoods)	**I (5161)**	**want**	**to**	**race**
VB (33693)				
TO				
NN				
PPSS (13802)	0.37 (5129)			

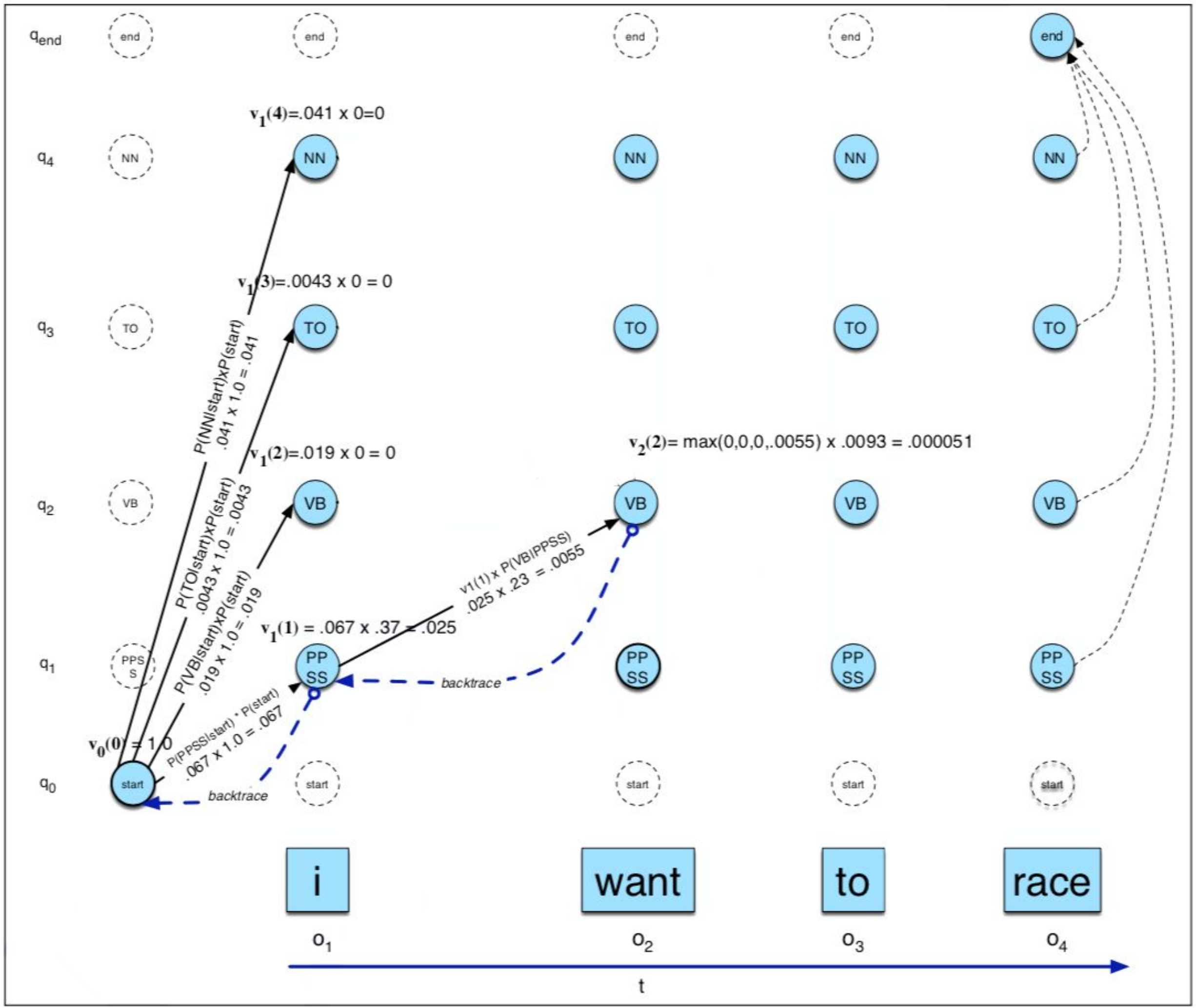
q_{end}
q_4
q_3
q_2
q_1
q_0
end
NN
TO
VB
PPS S
start
PP SS
$v_1(4)$=.041 x 0=0
$v_1(3)$=.0043 x 0 = 0
$v_1(2)$=.019 x 0 = 0
$v_1(1)$ = .067 x .37 = .025
$v_0(0)$ = 1.0
P(NN|start)xP(start) .041 x 1.0 = .041
P(TO|start)xP(start) .0043 x 1.0 = .0043
P(VB|start)xP(start) .019 x 1.0 = .019
P(PPSS|start) * P(start) .067 x 1.0 = .067
v1(1) x P(VB|PPSS) .025 x .23 = .0055
$v_2(2)$= max(0,0,0,.0055) x .0093 = .000051
backtrace
backtrace
i
want
to
race
o_1
o_2
o_3
o_4
t

Figure 5.18 The entries in the individual state columns for the Viterbi algorithm. Each cell keeps the probability of the best path so far and a pointer to the previous cell along that path. We have only filled out columns 0 and 1 and one cell of column 2; the rest is left as an exercise for the reader. After the cells are filled in, backtracing from the *end* state, we should be able to reconstruct the correct state sequence PPSS VB TO VB.

Name: ______________________________ StudentID: ___________________

3. (25 points) **Parts of speech.** Repeat the process as done in Figure 5.18 for the following new sentence, again using the Brown corpus: *She ran the shop.*

   Simplifying assumptions (see Fig 5.12):

priors						

likelihoods				

Viterbi graph (with correct POS backtrace highlighted or marked clearly):