Name: William Vagharfard StudentID: 860792866

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**

56,057	455,267	907,494

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

56,057 - 659kb; 455,267 - 7.3mb; 907,494 - 17.8mb

   Discuss their variation in terms of sparseness:

   As n gets larger it gets more sparse because there's fewer times each gram appears. Because there are less and less times a certain sequence of words happens the more words you add to the sequence

   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?

   The higher the "n", the less chance of finding that sequence in the text. The Brown corpus is not long enough to do a 5-gram. Even on the google ngram site, they won't let you do more than a 5-gram

   List the top five bigrams and their frequencies:

'of' 'the'	',' 'and'	'.' 'The'	'in' 'the'	',' 'the'
9625	6288	6081	5546	3754

   List the frequencies of the following phrases (case-sensitively):

   the President: 86 the Russian: 20 boiled haddock: 1

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

   ... ran the [x] ... **there are 5 words with a count of 1 that come after "ran the": change, risk, Grizzlies, 100-yard, length no other words in the Brown corpus come after "ran the"**

   ... [x] drinks ... **soft: 5, soft is the word with the highest count that comes before "drinks"**

   ... in the [x] ... **world: 89, world is the word with the highest count that comes after "in the"**

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: William Vagharfard StudentID: 860792866

Fig 5.15 (priors)	**VB**	**TO**	**NN**	**PPSS**
<s> (57340)	.015 (834)	.004 (230)	.024 (1377)	.055 (3146)
VB (33693)	0.0038 (130)	.035 (1187)	.043 (1463)	.007 (234)
TO (14918)	.82 (12291)	0	.0004 (6)	0
NN (152393)	.004 (602)	.017 (2565)	.076 (11625)	.005 (724)
PPSS (13802)	.23 (3183)	.0008 (11)	.001 (16)	.00007 (1)

Fig 5.16 (likelihoods)	**I** (5161)	**want** (326)	**to** (25732)	**race** (100)
VB (33693)	0	.009 (316)	0	.0001 (4)
TO	0	0	.984 (14619)	0
NN	.00002 (1)	.00006 (602)	0	.0006 (94)
PPSS (13802)	0.37 (5129)	0	0	0

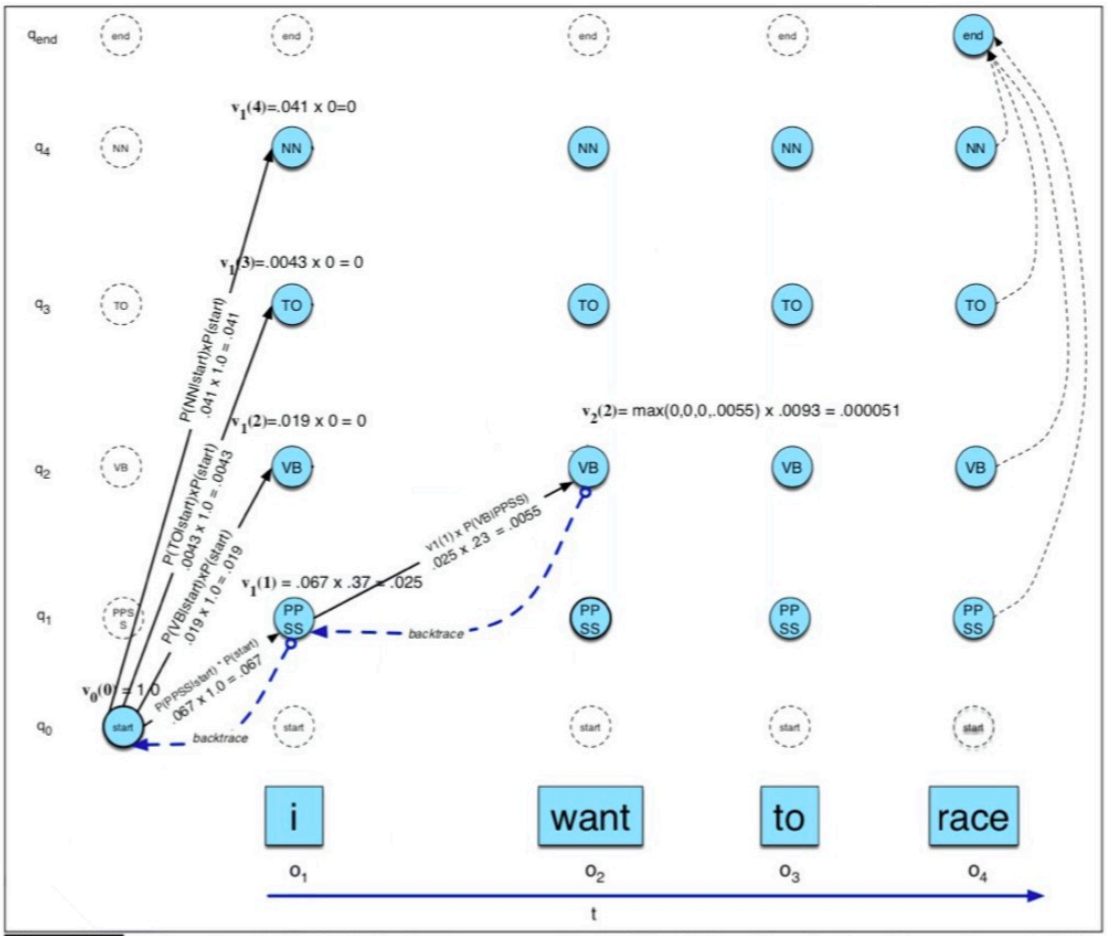

Figure 5.18 The entries in the individual state columns for the Viterbi algorithm. Each cell keeps the probability of the best path so far and a pointer to the previous cell along that path. We have only filled out columns 0 and 1 and one cell of column 2; the rest is left as an exercise for the reader. 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: William Vagharfard StudentID: 860792866

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	NN	VB	AT	AT-TL	PPS	VBD
<s> (57340)	.024 (1377)	.015 (834)	.024 (1377)	0	.104 (5935)	.0009 (52)
NN (152393)	.076 (11625)	.004 (602)	.007 (1041)	.00005 (8)	.006 (900)	.022 (3306)
VB (33693)	.043 (1463)	.0038 (130)	.179 (6044)	.00006 (2)	.004 (141)	.0002 (8)
AT (97959)	.494 (48368)	.0002 (16)	.00001 (1)	0	.00001 (1)	.00002 (2)
AT-TL (746)	.005 (4)	0	0	0	0	0
PPS (18253)	.0002 (4)	.008 (146)	.001 (25)	0	.00005 (1)	.338 (6175)
VBD (26167)	.020 (531)	.0007 (18)	.149 (3910)	.0003 (7)	.012 (326)	.0004 (10)

likelihoods	She (1949)	ran (132)	the (62713)	shop (56)
NN (152393)	0	0	0	.0003 (49)
VB (33693)	0	0	0	.0002 (7)
AT (97959)	0	0	.636 (62288)	0
AT-TL (746)	0	0	.299 (223)	0
PPS (18253)	.049 (911)	0	0	0
VBD (26167)	0	.005 (132)	0	0

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