Unstructured Information Processing 2

Divij Singh

16/09/19

1 Q1

(a)

(i) I
(ii)like
(iii)I
(iv)Sam

(b)

(iii) is the most likely, as the bigrams within it occur a total of 8 times within the training sentences.

2 Q2

(a)

War and Peace:
Monogram:
Bigram:
(i) i suppose replied well listen i ll set up with
(ii) and having got beyond cause of the cause of a searching
(iii) not two minutes then i looked inquiringly at it streamed into
Trigram:
(i) i suppose you to dress her would prolong their banter
(ii) and having got rid of this chance to talk about me
(iii) not two minutes and it will be answered why don t
Quadgram:

A Tale of Two Cities:
Monogram:
Bigram:
(i) i suppose so polite and the raging and his time (ii) and having got there was

confirmed it yet his forge we
(iii) not two minutes in me i know what to himself over
Trigram:
(i) i suppose the man he stooped and casting his eyes
(ii) and having got past the prime of life to bestow with
(iii) not two minutes and as time pressed for the rooms a
Quadgram:

(b)
As the complexity of the model increases, the sentences begin to become more coherent, that is to say they begin to more closely resemble actual human speech. This is due to the fact that the relation of a word in context to the words preceding it becomes far more specific, that is to say you get more unique n-grams.

(c)
The sentences generated by War and Peace are more coherent than those generated by A Tale of Two Cities. This is likely due to the larger corpus of War and Peace, thereby giving more n-grams sets, and more unique n-gram sets.