Q1. For both bi-grams and tri-grams, give 5 examples of words that look especially unnatural. For each word, explain why the model generates these unnatural strings.

N	Word	Reason
2	IH NG	It is a common-use suffix, and mainly translated as ing, due to the present continuous tense of verb should use add ing to word tail
2	EH N	EN is highly associated with N because there are many words contain 'en', which sounds like EN H
2	K T AE M IH NG Z IH L AH T EH L AH D	the NG is often appeard with Z translated as ns or n's, which are common used in words
2	F Y AH	F is frequently co-appear with Y as well as Y is frequently co-appear with AH
2	B R AA P R M AE CH	R is frequently co-appear with M
3	AA R D	the translated one, 'ard' is frequently used in words e.g., card, discard
3	D AE L AH JH IY AA R K IY	these phonomes sounds like word 'dalogiearkie', may because the ear connect the dalogie and arkie
3	IH P S	ipti is appear at many times in our dictionary e.g., hospital and description
3	T IH NG	Same as the ing, the 'ling' is also common appear in tail of words
3	AE N V AH L	the translated one version, 'anvul' is connected by 'anv' and 'vul'

Q2. Which model (the bi-gram or tri-gram) generates words that look more natural and English-like? Why is this model better able to produce English-like words?

> The tri-gram model seems better. Because as the n-gram increases in length, the better the n-gram model is on the training text. This is natural, since the longer the n-gram, the fewer n-grams there are that share the same context. As a result, this n-gram can occupy a larger share of the (conditional) probability pie.

Q3. Use smoothed bigrams and trigrams to find the perplexity of X.txt and Y.txt. What is the perplexity of each corpus according to each model? How should the results be interpreted (that is, what does

each model say about how English-like the two corpora are)?

> (1) At condition of bi-gram, the preplexities of X.txt and Y.txt are 21.2121 and 151.0546. At condition of tri-gram, the preplexities are 13.5390 and 21.5380.

> (2) The lower perplexity means the more 'nature' and 'English-like', take the dictionary as reference.

Q4. Look at the probabilities assigned to each word in the X and Y files by the two models. A model does a good job distinguishing between English-like words from non-English-like words if it is possible to choose a threshold such that the probabilities of all words in one file are below this threshold and the probabilities of all words are above it in the other file. Is it possible to do this for the bi-gram model? How about the tri-gram model? Which model does a better job distinguishing the two sets of words? Explain.

> (1) For the trigram model, it is hard to find a threshold to distinguish the English-like words and non-English-like words, because there is no significant difference between the two kinds of words. But for the bigram model, benifited by the larger variability in possibilities than trigram generated by those two types of words. It is possible to find a threshold to distinguish a English-like or non-English-like word, but the classification accuracy cannot reach the 100 percent.

> (2) As for bigram model, the difference between perplexities scores of X and Y is 129.8424. While for the trigram model, the difference is 7.9989. The bigger difference mean a classfier is more easier to distinguish the two sets of words.