8. Consider the task of using a probabilistic model to classify messages into newsgroups

(a) Demonstrate how a unigram model might disagree with a bigram one in this task (no need to come up with actual numbers, or get too specific, but we are looking for a concrete example)

(b) Let $D = (w_1, w_2, \ldots, w_n)$ be an n-word document we are attempting to classify. To do the classification with a unigram model we compute $c_{MAP} = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i \mid c)$, let us define $c_{MAP}^{(k)} = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^{k} P(w_i \mid c)$ for $k=1,\ldots,n$; then $c_{MAP}^{(k)}$ is our best hypothesis given that we only look at the first k words of D. If the prior is close to uniform (the *a priori* probalities of all newsgroups are about the same), what do you expect the sequence $\left(c_{MAP}^{(1)}, c_{MAP}^{(2)}, \ldots, c_{MAP}^{(n)}\right)$ to look like?

(c) Suppose instead of being given the documents, we were only given the tf.idf matrix for our document collection (newsgroup labels are given as before). Describe how you would train and do the classification based only on this information.

(d) Suppose instead of being given the documents, you were given only the author and the subject fields (newsgroup labels are given as before). Describe how you would train and do the classification based only on this information