In this week’s class, we discussed two main topics: one is the Naive Bayes, and another is text mining.

About Naive Bayes, although the primary function is not complex, sometimes the transformation forms could be flexible and hard to think about. In the calculation, we first define what we want to estimate as the event (like cereal) and steadily add in the pieces of evidence (milk, bread etc.). As if we use nested probability, it is complex and hard to compute. Thus, in many cases, we assume conditional independence for evidence and simplify the process (maybe not hold, but it is still useful). When dealing with zero-probability problems, we add a constant in the ratio as a constraint. The larger the constant, the larger the bias, but the less fitting problem.

For the text mining part, we first discussed the bag of words. The bag could be regarded as the whole area we can reach; thus, features mean an entire dictionary (more than 171k by principle). And during the word separation process, there are many words we should pay attention to. Such as UPPER/lowercase; stopwords; time terms; compound words; abbreviations etc. Sometimes we should discuss case by case to better generate the bags. And also, occasionally, people will use sarcasm, emojis, or some words that seem unrelated to the problem to discuss the topic, so we should pay attention to classification/ regression results.