Naive Bayes classification and Stylometry units

- Naive Bayes classification
  - The math of multinomial naive Bayes classification (last week Monday)
  - Lab: NBC (last week Friday)
  - Practical considerations of NBC (**Today**)
- Stylometry and authorship attribution
  - The authorship attribution problem (Wednesday)
  - Lab: Stylometry techniques (Friday)
  - Applied stylometry (next week Monday)

Today:

- From formula to algorithm
- Tailoring NBC to specific classification tasks
- Connections between NBC and language models
- Evaluation metrics
- Ethical considerations

Let V be the set of word types we care about, let $D = |V|$, let f_i be the count of word/feature i in document d, and let $P(v_j \mid c)$ be the probability of v_j occurring in a document of class c. Then

$$\begin{aligned} c_{NB} &= \text{argmax}_{c \in C} \quad P(c) \cdot \prod_{i=0}^{D-1} P(v_i \mid c)^{f_i} \\ &= \text{argmax}_{c \in C} \quad \log P(c) + \sum_{i=0}^{D-1} f_i \log P(v_i \mid c) \end{aligned}$$


Don't lunch.
your forget

	Have disease	Don't have disease
New test says have disease	TP=1	FP = 9
New test says don't have disease	FN=9	TN=9981

	Have disease	Don't have disease
New test says have disease	TP=10	FP = 9990
New test says don't have disease	FN=0	TN=0

	Have disease	Don't have disease
New test says have disease	TP=0	FP = 0
New test says don't have disease	FN=10	TN=9990

	Have disease	Don't have disease
New test says have disease	TP=10	FP = 10
New test says don't have disease	FN=0	TN=9980

Coming up:

- Do bag-of-words programming assignment (Mon, Nov 3)
- Take NBC quiz (Tues, Nov 4)
- Do NBC programming assignment (Mon, Nov 10)
- Read stylometry overview/survey paper (Tues, Nov 4)
- Read another stylometry paper (Fri, Nov 7)
- Take stylometry quiz (Tues, Nov 11)