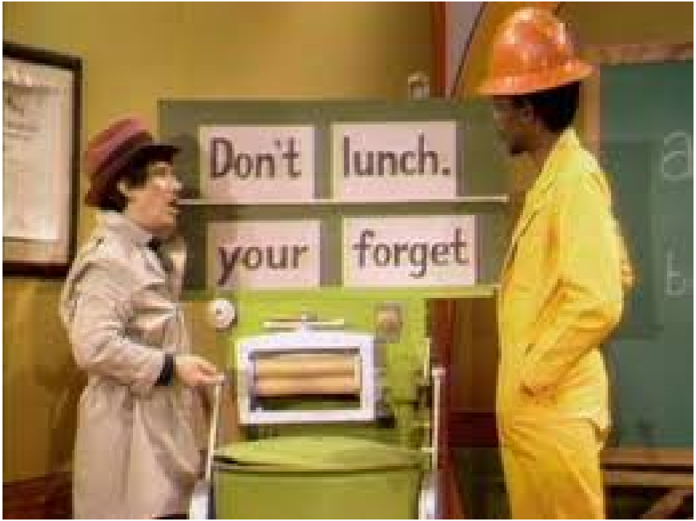
Naive Bayes classification and Stylometry units

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

Today:

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

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


Don't
lunch.
your
forget
a
b

	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 NBC programming assignment (Fri, Nov 3)

- Take AA/S basics quiz (Tues, Oct 31)
- Read AA/S survey paper (Wed, Nov 1)
- Take AA/S details quiz (Thurs, Nov 2)
- Read other AA/S papers (Fri, Nov 3)

AA/S = *authorship attribution and stylometry*