Naive Bayes
In-class exercise, Sept 10

Name: _____

Write out quantities as numbers, but don't simplify fractions. so it's easier.

DATA

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

TRAINING
doc label prior

P(+) = P(–) =

TRAINING
word likelihoods

Num tokens in neg texts (-) =
Num tokens in pos texts (+) =
V (vocab size) = 20 …. why is the less than the sum above?

Fix pseudocount = 0.1

$P(\text{“predictable”}|-) =$ $P(\text{“predictable”}|+) =$

$P(\text{“with”}|-) =$ $P(\text{“with”}|+) =$

$P(\text{“no”}|-) =$ $P(\text{“no”}|+) =$

$P(\text{“originality”}|-) =$ $P(\text{“originality”}|+) =$

PREDICTION

$P(+)\ P(S|+) =$

$P(-)\ P(S|-) =$

$\arg\max_{y \in \{+,-\}} P(y \mid S) =$

QUESTION: what would happen if pseudocount=0?

FUN BONUS QUESTION
Guess: what are the 10 most common words in English,
at least in the Brown corpus (comprised of news, fiction, and essays)