Write out quantities as numbers, but don’t simplify fractions. So it’s easier.

### Naive Bayes

In-class exercise, Sept 10

<table>
<thead>
<tr>
<th>Cat</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>just plain boring</td>
</tr>
<tr>
<td></td>
<td>entirely predictable and lacks energy</td>
</tr>
<tr>
<td></td>
<td>no surprises and very few laughs</td>
</tr>
<tr>
<td>+</td>
<td>very powerful</td>
</tr>
<tr>
<td>+</td>
<td>the most fun film of the summer</td>
</tr>
<tr>
<td>Test</td>
<td>? predictable with no originality</td>
</tr>
</tbody>
</table>

**TRAINING**

**doc label prior**

\[ P(+) = \quad P(-) = \]

**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

\[
P(\text{“predictable”}|-) = \quad P(\text{“predictable”}|+) =
\]

\[
P(\text{“with”}|-) = \quad P(\text{“with”}|+) =
\]

\[
P(\text{“no”}|-) = \quad P(\text{“no”}|+) =
\]

\[
P(\text{“originality”}|-) = \quad P(\text{“originality”}|+) =
\]

### PREDICTION

\[
P(+) \quad P(S|+) =
\]

\[
P(-) \quad P(S|-) =
\]

\[
\arg \max_{y \in \{+, -\}} P(y | S) =
\]

**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)