1. Document similarity

a vocabulary of 3 words:
\( W = \{\text{woof}, \text{meow}, \text{squeak}\} \)

each document is characterized by a vector of word counts:
\( V_1 = [2, 1, 0] \)
\( V_2 = [2, 0, 1] \)

a) \( \sum a_i \times b_i = [2, 1, 0] \cdot [2, 0, 1] = 4 + 0 + 0 = 4 \)
\( \sum a_i^2 = 4 + 1 + 0 = 5 \)
\( \sum b_i^2 = 4 + 0 + 1 = 5 \)
\( \text{Sim}(A, B) = \frac{4}{\sqrt{5} \times \sqrt{5}} = \frac{4}{5} = 0.8 \)

b) \( \text{IDF}_{\text{woof}} = \log \left( \frac{N}{n_{\text{woof}}} \right) = \log \left( \frac{2}{2} \right) = 0 \)
\( w_1 = 0 \)
\( w_2 = 0 \)
\( \text{IDF}_{\text{meow}} = \log \left( \frac{2}{1} \right) = .30103 \)
\( w_1 = .30103 \)
\( w_2 = 0 \)
\( \text{IDF}_{\text{squeak}} = \log \left( \frac{2}{1} \right) = .30103 \)
\( w_1 = 0 \)
\( w_2 = .30103 \)
\( V_1 = [0, .30103, 0] \)
\( V_2 = [0, 0, .30103] \)
\( \sum a_i \times b_i = [0, .30103, 0] \cdot [0, 0, .30103] = 0 \)
\( \text{Sim}(A, B) = 0 \)

(the only word the documents have in common is “woof”, but “woof”
appears in every document and so gets an IDF weight of 0)

c) Word counts for the third document: \( V_3 = [0, 1, 1] \)
\( \text{IDF}_{\text{woof}} = \log \left( \frac{3}{2} \right) = .17609 \)
\( w_1 = .35218 \)
\( w_2 = .35218 \)
\( w_3 = 0 \)
\( \text{IDF}_{\text{meow}} = \log \left( \frac{3}{2} \right) = .17609 \)
\[ w_1 = .17609 \]
\[ w_2 = 0 \]
\[ w_3 = .17609 \]
\[ \text{IDF}_{\text{squeak}} = \log(3/2) = .17609 \]
\[ w_1 = 0 \]
\[ w_2 = .17609 \]
\[ w_3 = .17609 \]
\[ V_1 = [.35218, .17609, 0] \]
\[ V_2 = [.35218, 0, .17609] \]
\[ V_3 = [0, .17609, .17609] \]

\[ \Sigma a_i \times b_i = [.35218, .17609, 0] \times [.35218, 0, .17609] \]
\[ = .12403 + 0 + 0 = .12403 \]

\[ \Sigma a_i^2 = .35218^2 + .17609^2 + 0 = .12403 + .03101 \]
\[ = .15504 \]
\[ \Sigma b_i^2 = .35218^2 + 0 + .17609^2 \]
\[ = .15504 \]

\[ \text{Sim}[A, B] = .12403 / (\sqrt{.15504} * \sqrt{.15504}) = .12403 / .15504 \]
\[ = .7999 \]

("woof" no longer appears in every document, and so has non-zero IDF in the larger document collection)
2. In applying Naïve Bayes to a 'bag of words' we have to make a choice as to how we model a bag: whether we just take into account whether a bag contains or doesn’t contain a word, or also take into account the frequency of the word in the bag. Both options are ‘correct’, although one choice may be better than the other for a particular application. As discussed in class, we will follow the procedure for Bernoulli Naïve Bayes, based on the presence or absence of each term. Following the algorithm in Figure 13.1 of the Stanford IR book (http://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html) we include factors for both the probability that a word appears in a document (denoted P(word)) and the probability that it doesn’t appear in a document. (denoted P(word))

(a) A priori probabilities without smoothing:
P(+) = 5/10 = 0.5
P(−) = 5/10 = 0.5

Conditional probabilities without smoothing:
P (great | +) = 5/5 P(great| - ) = 0 / 5
P (food | +) = 5 / 5 P (food - ) = 5 / 5
P (served | +) = 0 / 5 P(served | - ) = 1 / 5
P (terrible | +) = 0 / 5 P (terrible | - ) = 5 / 5

P (great | +) = 0 / 5 P (great | - ) = 5 / 5
P (food | +) = 0 / 5 P (food | - ) = 0 / 5
P (served | +) = 5 / 5 P (served | - ) = 4 / 5
P (terrible | +) = 5 / 5 P (terrible | - ) = 0 / 5

Using Naïve Bayes formula and dropping the denominator:

P (+ | “great food served”) = P (great | +) P(food | +) P(served | +) P(terrible | +) P( +) = 1 * 1 * 0 * 1 * (5/10) = 0

P (- | “great food served”) = P (great | -) P(food | -) P(served | -) P(terrible | -) P( +) = 0 * 1 * 0.2 * 0 * (5/10) = 0

So can’t decide with unsmoothed probabilities.
(b) Conditional probabilities with smoothing:

For each opinion (+ or -) and each word, there are two possibilities: the review contains the word, or the review does not contain the word. In adding one smoothing, we add one instance of each possibility. We started with 5 ‘+’ reviews, so now we have 7 (and also 7 ‘-’ reviews).

\[
\begin{align*}
P(\text{great} \mid +) &= 6/7 & P(\text{great} \mid -) &= 1/7 \\
P(\text{food} \mid +) &= 6/7 & P(\text{food} \mid -) &= 6/7 \\
P(\text{served} \mid +) &= 1/7 & P(\text{served} \mid -) &= 2/7 \\
P(\text{terrible} \mid +) &= 1/7 & P(\text{terrible} \mid -) &= 6/7 \\
\end{align*}
\]

Using Bayes’ Rule and dropping the denominator (which is the same for the ‘+’ and ‘-‘ cases)

\[
\begin{align*}
P(+ \mid \text{“great food served”}) &= P(\text{great} \mid +) P(\text{food} \mid +) P(\text{served} \mid +) P(\text{terrible} \mid +) P(+) \\
&= (6/7) \times (6/7) \times (1/7) \times (6/7) \times (5/10) = 0.045 \\
\end{align*}
\]

\[
\begin{align*}
P(- \mid \text{“great food served”}) &= P(\text{great} \mid -) P(\text{food} \mid -) P(\text{served} \mid -) P(\text{terrible} \mid -) P(+) \\
&= (1/7) \times (6/7) \times (2/7) \times (1/7) \times (5/10) = 0.0025 \\
\end{align*}
\]

\[P(+) \mid \text{“great food served”}) \text{ would be larger with Laplace smoothing.}\]