Problem 1 (Near-Duplicates).
Consider the following three documents \( d_1, d_2, \) and \( d_3 \), consisting of just one sentence each:
\[
d_1 = \langle \text{I saw Susie sitting in a shoeshine shop} \rangle \\
d_2 = \langle \text{Susie shines shoes in a shoeshine shop} \rangle \\
d_3 = \langle \text{I saw Susie sitting and shining shoes in a shoe shop} \rangle
\]
a) Extract a set of shingles of size \( n=3 \) (i.e., word-level 3-grams) for \( d_1, d_2, \) and \( d_3 \).
b) Compute the resemblance (pairwise Jaccard similarities) and the containment (relative overlap) between the sets of shingles you obtained from (a).
c) Apply the Min-Wise Independent Permutations (Min-Hash) technique to the set of shingles you obtained from (a). Define a set of 3 random permutations as appropriate hash functions to compare the pairwise similarities of the documents. Calculate the resemblance between documents, based on the results of your Min-Hash functions.

Problem 2 (Rocchio Method with Relevance Feedback). Suppose we want to search the following collection of Christmas cookie recipes. Assume that the numbers in the table indicate raw term frequencies.

<table>
<thead>
<tr>
<th></th>
<th>milk</th>
<th>pepper</th>
<th>raisins</th>
<th>sugar</th>
<th>cinnamon</th>
<th>apples</th>
<th>flour</th>
<th>eggs</th>
<th>clove</th>
<th>jelly</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( d_3 )</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( d_4 )</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>( d_5 )</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( d_6 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>( d_7 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( d_8 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Assume that we use the vector space model with the following TF*IDF variant
\[
tf.idf(t, d) = tf(t, d) \times \log \frac{|D|}{df(t)} .
\]
and rank documents according to cosine similarity.

(a) Determine the top-3 documents for the query vector
\[
q_3 = \langle 1001100000 \rangle .
\]

(b) Using Rocchio’s method \((\alpha = 0.5, \beta = 0.3, \gamma = 0.1)\) determine the query vector \( q'_3 \), assuming \( D^+ = \{d_1, d_2\} \) and \( D^- = \{d_3\} \) as positive and negative relevance feedback. Remember that Rocchio’s method determines the new query vector as
\[
q' = \alpha \ q + \frac{\beta}{|D^+|} \sum_{d \in D^+} d - \frac{\gamma}{|D^-|} \sum_{d \in D^-} d .
\]

(c) Determine the top-3 documents for the query vector \( q'_3 \).
Problem 3 (KL-Divergence and Statistical Language Models). Show that the multinomial language model (LM) for unigrams (i.e., independent occurrences of single words) ranks documents \( d \) for a given query \( q \) in ascending order of the Kullback-Leibler divergence (relative entropy) between the query LM and the document LM.

That is, the best result is the document \( d \) with minimum \( D(LM(q) \mid\mid LM(d)) \) where

\[
D(f \mid\mid g) = \sum_{x} f(x) \cdot \log \frac{f(x)}{g(x)}
\]

Some useful formulas:

- Multinomial distribution with parameters \( n, p_1, \ldots, p_m \):
  \[
P\left[ X_1 = x_1, \ldots, X_m = x_m \left| \sum_{i=1}^{m} x_i = n \right. \right] = \binom{n}{x_1 \ldots x_m} \cdot \prod_{i=1}^{m} p_i^{x_i}
  \]

- Entropy of discrete probability distribution \( f \):
  \[
  H(f) = -\sum_{x} f(x) \log f(x)
  \]

- Cross-entropy of discrete distributions \( f \) and \( g \):
  \[
  H(f, g) = -\sum_{x} f(x) \log g(x)
  \]

Problem 4 (Word Sense).

Consider the following terms and their corresponding set of semantically related terms given by WordNet:

<table>
<thead>
<tr>
<th>Term</th>
<th>Related terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>fall</td>
<td>autumn, tumble, descend, surrender, drop</td>
</tr>
<tr>
<td>run</td>
<td>test, trial, race, operate, political campaign</td>
</tr>
<tr>
<td>card</td>
<td>identity card, poster, calling card, menu, circuit board</td>
</tr>
<tr>
<td>strike</td>
<td>work stoppage, hit, affect, impress, collide into</td>
</tr>
</tbody>
</table>

a) Give an example of an information need using one or more of these terms, a short corresponding search query, and an example of how the query can be re-written or expanded with some of the related terms such that the new query improves search results.

b) Give an example of an information need using one or more of these terms, a short corresponding search query, and an example of how it can be re-written or expanded with one of the related terms such that the new query yields worse search results.

c) Briefly explain how information extraction techniques, such as part-of-speech or word sense tagging, could be used in each case (a) and (b) to improve results.