Homework 8:

Due: November 4, 2025 at 2:30p.m.

This homework must be typed in LaTeX and submitted via Gradescope.

Please ensure that your solutions are complete, concise, and communicated clearly. Use full sentences and plan your presentation before your write. Except where indicated, consider every problem as asking for a proof.

Problem 1. In document similarity analysis, the *shingling* algorithm is used to represent a document as a set of contiguous sequences (called *shingles*) to capture its structure and content. A k-character shingle uses contiguous substrings of length k characters, while a k-word shingle uses contiguous sequences of k words.
You are given the following two short documents:

- **Document A:** "The quick brown fox jumps over the lazy dog."
- **Document B:** "A quick brown dog jumps over a lazy fox."

1. Generate the set of **3-character shingles** for both documents (ignore punctuation and use lowercase).
   - List the first 10 distinct shingles for each document.
2. Generate the set of **3-word shingles** for both documents.
   - Show all 3-word shingles for each document.
3. Compute the **Jaccard similarity** between the two documents using:

   label=(i) 3-character shingles

   lbbel=(ii) 3-word shingles
4. Discuss the **differences** between using k-character and k-word shingles for text similarity.
   - Which method is more sensitive to small changes in wording or punctuation?
   - Which method better preserves semantic meaning?
   - Which would you choose for near-duplicate web page detection, and why?

Solution. 1.
- Document A: "the", "he ", "e q", " qu", "qui", "uic", "ick", "ck ", "k b", " br"
- Document B: "a q", " qu", "qui", "uic", "ick", "ck ", "k b", " br", "bro", "row".

2. The quick brown, quick brown fox, brown fox jumps, fox jumps over, jumps over the, over the lazy, the lazy dog
   A quick brown, quick brown dog, brown dog jumps, dog jumps over, jumps over a, over a

lazy, a lazy fox

3. the, he , e q, qu,qui, uic, ick, ck , k b, br, bro, row, own, wn , n f, fo, fox, ox , x j, ju, jum, ump, mps, ps ,s o, ov,ove,ver, er , r t, th, e l, la,laz,azy,zy , y d, do, dog
   a q, qu, qui, uic,ick, ck , k b, br,bro,row,own,wn ,n d, do, dog,og , g j, ju, jum, ump, mps,ps , s o, ov, ove,ver, er , r a, a , a l, la,laz,azy,zy ,y f, fo, fox

   label= 29/76-29 = 29/47 ii) label= 0

4. The word-shingle is more sensitive to articles like a and the. The character shingle better preserves similarity in semantic meaning, but the word shingle preserves the difference in the literal meaning overall. Would choose character shingle as you learn more about the similar words and meanings within texts like this which are very similar despite saying different things. Some justification along these lines should be fine.

□

Problem 2.

1. Let $F^k = \{f : \{0,1\}^k \to \{0,1\}\}$ be the set of all Boolean functions on k Boolean inputs. Give a very simple argument that the set F^k is countable.
2. Let $F^* = \cup_{i=1}^{\infty} F^k$. Show that the set F^* is countable.
3. Prove that all decision questions on finite graphs are decidable.
4. Let F^∞ be the set of Boolean functions with a countable number of Boolean inputs. Prove that F^∞ is uncountable.
5. (*) Give an example of a function in $F^\infty \setminus F^*$

Solution.

1. A function $f : \{0,1\}^k \to \{0,1\}$ is unique determined by the value on each of its inputs. It has 2^k inputs and for each input, it has 2 possible outputs 0 and 1. Therefore, there are 2^{2^k} such Boolean functions, $|F_k| = 2^{2^k}$. F_k is finite, hence countable.
2. We know that F_k is finite by part 1. Thus we can enumerate F^* by listing all elements of F_1, then all elements of F_2, and so on. This shows that F^* is countable.

   Note: More generally, you can show that a countable union of countable (which implies finite) sets is countable. This follows the proof for countability of $\mathbb{Q}$. Precisely, we can list on a two-dimensional array by diagonals where the (i,j)-th cell contains the $j-th$ function in F^i.
3. A graph with n vertices is determined by whether there is an edge between each of its $\binom{n}{2}$ vertex pairs, hence there are $2^{\binom{n}{2}}$ graphs on n vertices. Thus, a decision question on graphs with n vertices is a Boolean function on $2^{\binom{n}{2}}$ inputs. Since there is a finite set of such Boolean functions by part 1, the decision question is decidable.
4. Suppose by contradiction that F^∞ is countable. Then we can enumerate elements in F^∞ as $f_1, f_2, \dots$ where each $f_i \in F^\infty$ for some $n \in \mathbb{N}$. For $i \in \mathbb{N}$, let $e_i = (0, \dots, 0, 1_i, 0, \dots) \in \{0,1\}^{\mathbb{N}}$, the element with only 1 in the $i-$th coordinate and 0 everywhere else. We can define function $f : \{0,1\}^{\mathbb{N}} \to \{0,1\}$ where $f(e_i) = 1 - f_i(e_i)$ for all $i \in \mathbb{N}$. Then $f(e_i) \neq f_i(e_i)\ \forall i$, so $f \neq f_i\ \forall i$, which is a contradiction. Thus F^∞ is uncountable.
5. Define $f : \{0,1\}^{\mathbb{N}} \to \{0,1\}$ by

$$f(a_1, a_2, \dots) \begin{cases} 0 & \text{if there are finitely many 1's in } (a_i)_{i \in \mathbb{N}} \\ 1 & \text{otherwise} \end{cases}$$

   Clearly $f \in F^\infty$. Now assume that $f \in F^*$, then $f \in F^n$ for some $n \in \mathbb{N}$ Now consider $x = (a_1, \dots, a_n, 0, 0, \dots)$ and $y = (a_1, \dots, a_n, 1, 1, \dots)$ (x have all zeros after the first n coordinates, y have all ones after the first n coordinates). Then $f(x) = 0$ and $f(y) = 1$. However, any function g in F^n is determined by the first n coordinates, so $g(x) = g(y)$. Thus f cannot be a function in F^n, which is a contradiction. Therefore, $f \notin F^*$. We have defined a function in $F^\infty \setminus F^*$

□