

Short Problem Set (Module 2)

Note: in general, when I ask you to calculate something, please show work; it often allows me to give partial credit for answers that are close by not entirely correct. And I may not give full credit for answers that are actually correct, but where there is insufficient or no justification.

1. [40%] Character n-gram overlap is used for both automated spelling correction and personal name matching (i.e., deciding whether two names might be the same, a common database problem known as “record linkage”). Using a character 3-gram representation, how many n-grams do “MISSISSIPPI” and “MISSISSIPI” have in common (the latter is missing a 'P')? What is the Dice-coefficient score for these two strings using 3-grams? What is the Dice score using 2-grams instead? Which score is higher? Note: although there is nothing conceptually wrong in doing so, for this problem, do not use “padded” n-grams (e.g., ‘\$’ or ‘_’ symbols marking the beginning and end of the strings).

2. [30%] Compute the edit distance (or Levenshtein distance) for these two pairs of strings: (a) “EYESCREAM” and “ICECREAM”; and (b) “BROKENSTONE” and “BOOKSTORES”. Then report a sequence of transformations for that cost that converts one string into the other. You should use unit costs for each operation: insertion, deletion, or substitution; that is, each step has a cost of 1. Note, you do not need to write a program or produce any code for this problem – these examples can be easily determined by pen and paper – you do not need to construct a table as the example in the textbook.

Trivia: Computing edit distance is a classic dynamic programming problem with an $O(N^2)$ complexity. However, the decision problem “Do strings x and y (say each of length N) have an edit distance $\leq k$ can be solved in $O(kN)$, which is subquadratic.

3. [30%] Following the method described in the textbook (or lecture materials), what are the Soundex codes for the strings: (a) "Jelinek" and (b) “Khudanpur”? Show your intermediate steps to produce the code.