Module 2 Short Problem Set

Max Robinson

1 Problem 1

Question: 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).

3-Grams for "MISSISSIPPI": MIS, ISS, SSI, SIS, ISS, SSI, SIP, IPP, PPI
3-Grams for "MISSISSIPI": MIS, ISS, SSI, SIS, ISS, SSI, SIP, IPI
There are **7** 3-grams in common: MIS, ISS, SSI, SIS, ISS, SSI, SIP
Dice Co-efficient $= 7/(9+8) = 7/17 = 0.412$

2-Grams for "MISSISSIPPI": MI IS SS SI IS SS SI IP PP PI
2-Grams for "MISSISSIPI": MI IS SS SI IS SS SI IP PI
Dice Co-efficient for 2-grams $= 8/(10+9) = 8/19 = 0.421$

The score for the 2-grams is higher.

2 Problem 2

Question: Compute the edit distance (or Levenshtein distance) for these two pairs of strings: (a) "EYESCREAM" and "ICECREAM"; and (b) "BROKEN-STONE" 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.

EYESCREAM to ICECREAM
Distance = 3
delete pos 3 → EYECREAM
replace pos 0 with I → IYECREAM
replace pos 1 with C → ICECREAM

BROKENSTONE to BOOKSTORES
Distance = 5
replace pos 1 with O → BOOKENSTONE
delete pos 4 → BOOKNSTONE
delete pos 4 → BOOKSTONE
replace pos 7 with R → BOOKSTORE
insert pos 9 (at end) an S → BOOKSTORES

3 Problem3

Question: 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.

"Jelinek"
without AEIOUHWY: J0l0n0k
replace CGJKQSXZ: J0l0n02 (Note: we leave the first uppercase letter)
replace L: J040n02
replace MN: J040502
Remove all 0's: J452
Final "Jelinek" → J452

"Khudanpur"
without AEIOUHWY: K00d0np0r
replace DT: K0030np0r
replace MN: K00305p0r
replace B, F, P, V: K0030510r
replace R: K00305106
Remove all 0's: K3516
Truncate to get 4 chars: K351
Final "Khudanpur" → K351