CSC 583: Assignment 2

February 18, 2017 [redacted]

Problem 1

Suggest what tokenization and normalized form(s) should be used for these words (including the word itself as a possibility). Justify your decision.

- 'Cos
  because – This should be changed to the correct spelling / real meaning.
- Shi'ite
  shiite – People are likely to leave out the apostrophe when they search.
- cont'd
  continued – This should be changed to the complete spelling.
- Hawai'i
  hawaii – People usually leave out the apostrophe when they spell Hawai'i.
- O'Rourke
  orourke – Again, people may omit the apostrophe when searching for this name.
- ain't → Or "be not"
  is not – This should be changed to the proper form of the meaning.
- me@privacy.net
  me@privacy.net – This should be left as-is because anyone searching for this probably wants an exact match.
- <html>Some text </html>
  <**html**>**Some text** <**/html**> – Should also be left as-is because anyone searching for this string is probably a technical user who wants an exact match.

Problem 2

Assume a biword index. Give an example of a document (could be a made up paragraph) which will be returned for a query of "New York University" but is actually a false positive which should not be returned.

"**York University** (French: Université York) is a public research university in Toronto, Ontario, Canada. York University has approximately 52,300 students, 7,000 faculty and staff, and 295,000 alumni worldwide. Although a large number of alumni live in Ontario, a significant number live in British Columbia, Nova Scotia, Alberta, **New York**, and Washington, D.C."

Problem 3

Shown below is a portion of a positional index in the format: term: doc1: ⟨position1, position2, ...⟩; doc2: ⟨position1, position2, ...⟩; etc.

angels: 2: ⟨36, 174, 252, 651⟩; 4: ⟨12, 22, 102, 432⟩; 7: ⟨17⟩;
fools: 2: ⟨1, 17, 74, 222⟩; 4: ⟨8, 78, 108, 458⟩; 7: ⟨3, 13, 23, 193⟩;
fear: 2: ⟨87, 704, 722, 901⟩; 4: ⟨13, 43, 113, 433⟩; 7: ⟨18, 328, 528⟩;
in: 2: ⟨3, 37, 76, 444, 851⟩; 4: ⟨10, 20, 110, 470, 500⟩; 7: ⟨5, 15, 25, 195⟩;
rush: 2: ⟨2, 66, 194, 321, 702⟩; 4: ⟨9, 69, 149, 429, 569⟩; 7: ⟨4, 14, 404⟩;
to: 2: ⟨47, 86, 234, 999⟩; 4: ⟨14, 24, 774, 944⟩; 7: ⟨199, 319, 599, 709⟩;
tread: 2: ⟨57, 94, 333⟩; 4: ⟨15, 35, 155⟩; 7: ⟨20, 320⟩;
where: 2: ⟨67, 124, 393, 1001⟩; 4: ⟨11, 41, 101, 421, 431⟩; 7: ⟨16, 36, 736⟩;

Which document(s) if any match each of the following queries, where each expression within quotes is a phrase query?

1. "fools rush in"

   Document 2: fools [1], rush [2], in[3].
   Document 4: fools [8], rush [9], in [10].
   Document 7: fools [3], rush [4], in [5]; and fools[13], rush [14], in[15].

   So **documents 2, 4, and 7** match the query.

2. "fools rush in" `AND` "angels fear to tread"

   Since all documents match the first phrase, we just find the documents which match the second phrase.

   Not document 2, since "angels" and "fear" are not adjacent.
   Document 4: angels [12], fear [13], to [14], tread [15].
   Not document 7, since "angels" only appears at [17], and there is no "to" at [19].

   So only **document 4** matches the query.

Problem 4

Write down the entries in the permuterm index dictionary that are generated by the term "hope".

Add the terminal symbol and perform all possible rotations:

- hope$
- $hope
- e$hop
- pe$ho
- ope$h

Problem 5

Compute the edit distance between "paris" and "arid". What are the N (rows) and M (columns) dimensions of the edit distance matrix? Write down the N $\times$ M array of distances between all prefixes as computed by the edit distance algorithm in Figure 3.5 in IIR. For each cell in the matrix, use the four-number representation to keep track of your intermediate results.

The matrix is 6 $\times$ 5 (including initialization row and column). The edit distance is 2: delete 'p' from "paris", and replace 's' with 'd' from "arid".

		a	r	i	d
	0	1 **1**	2 **2**	3 **3**	4 **4**
p	1 **1**	1 2 / 2 **1**	2 3 / 2 **2**	3 4 / 3 **3**	4 5 / 4 **4**
a	2 **2**	1 2 / 3 **1**	2 3 / 2 **2**	3 4 / 3 **3**	4 5 / 4 **4**
r	3 **3**	3 2 / 4 **2**	1 3 / 3 **1**	3 4 / 2 **2**	4 5 / 3 **3**
i	4 **4**	4 3 / 5 **3**	3 2 / 4 **2**	1 3 / 3 **1**	3 4 / 2 **2**
s	5 **5**	5 4 / 6 **4**	4 3 / 5 **3**	3 2 / 4 **2**	2 3 / 3 **2**

Problem 6

Consider the following fragment of a positional index with the format:
word: document: ⟨position, position, ...⟩; document: ⟨position, ...⟩
...

Gates: 1: ⟨3⟩; 2: ⟨6⟩; 3: ⟨2, 17⟩; 4: ⟨1⟩;
IBM: 4: ⟨3⟩; 7: ⟨14⟩;
Microsoft: 1: ⟨1⟩; 2: ⟨1, 21⟩; 3: ⟨3⟩; 5: ⟨16, 22, 51⟩;

The $/k$ operator, word1 $/k$ word2 finds occurrences of word1 within k words of word2 (on either side), where k is a positive integer argument. Thus $k = 1$ demands that word1 be adjacent to word2.

1. Describe the set of documents that satisfy the query Gates /2 Microsoft.

   The documents which match the query are **1 and 3**, since document 1 has Gates [3] and Microsoft [1], and document 3 has Gates[2] and Microsoft [3].

2. Describe each set of values for k for which the query Gates $/k$ Microsoft returns a different set of documents as the answer.

   If $k = 1$, then the set of documents will no longer contain document 1.

   If $k \geq 5$, then the set of documents will grow to include document 2 (Gates [6] and Microsoft [1]). There are no other documents containing both Gates and Microsoft, so further changes increases to k won't change the results.