Assignment 2 Solution

Problem 1

- ’Cos: **because** – This should be changed to the common spelling of the complete word “because”.
- 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: **is not** or **be not** – This should be changed to the unabbreviated form.
- me@privacy.net: **me@privacy.net** – This should be left as-is because anyone searching for this probably wants an exact match of the address.
- <html>Some text </html>: **<html> some text </html>** - Keep the HTML tags intact; tokenize the rest.

Problem 2

“**York University** (French: Universite 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

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 that 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

pandemic$

$pandemic

c$pandemi

ic$pandem

mic$pande

emic$pand

demic$pan

ndemic$pa

andemic$p

Problem 5

		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

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

   The documents that 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 increases to *k* won't change the results.