CSE 586 Midterm Exam

Olcay Taner YILDIZ

I. Question (12 points)

Consider these documents:

Doc1: new home sales top forecasts
Doc2: home sales rise in july
Doc3: increase in home sales in july
Doc4: july new home sales rise

1) Draw the term-document incidence matrix for this document collection.
2) Draw the inverted index representation for this collection.

II. Question (12 points)

Write the algorithm that handles the boolean query x XOR y using inverted index.

III. Question (12 points)

Apply filters

- Case-folding
- Stemmer
- Stop word removal

to the document

```
Such an analysis can reveal features
that are not easily visible from the
Variations in the individual genes
and can lead to a Picture of
expression that is more biologically
Transparent and accessible to
interpretation
```

IV. Question (10 points)

Create a portion of a positional index such that the first query “fools rush in” returns documents with id’s 4, 7, 10, 14; the second query “fools rush in” AND “fools fear” returns documents with id’s 4, 14; and the third query “fools fear” returns documents with id’s 1, 4, 12, 14.

V. Question (12 points)

Give two words of length 5 whose edit distance is 3. Compute their edit distance using dynamic programming.

VI. Question (12 points)

Write the pseudocode for front coding of N tokens. For example, if we compress the words

```
automata
automate
automatic
automation
```

using front coding, we get

```
8automat$a1#e2#ic#3#ion
```

VII. Question (10 points)

If there are 4 zones in a document and the weights of these zones are g_0 = 0.1, g_1 = 0.2, g_2 = 0.3 and g_3 = 0.4, what are all the distinct score values a document may get?

VIII. Question (12 points)

Let say we use nnn (direct term frequency, 1 as document frequency and no normalization) as tf-idf variant in query ranking. Construct a query such that the resulting ranking will be 1, 3, 2, 4 for the documents in Question 1.

IX. Question (8 points)

An IR system returns 8 relevant documents, and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?