CSCE 561 Assignment#2, Fall 2017

Vijay V Raghavan

Assigned: 18th October 2017

Due: 27th October 2017

---

1. *All details of work for each question must be submitted.*
2. *Staple the question and answer sheet together.*
3. *Make a cover with Name, CLID.*
4. *Number all pages and give an index to each question.*
5. *Any sort of cheating will not be tolerated. More information on cheating policy can be found on class webpage.*

Q1: 20 Points

Let the retrieval output for a query, using a retrieval system be

$$\Delta = (+ - + \mid + - + \mid - + - - + \mid - + -)$$

a) Find recall and fallout values, after retrieving 3, 6, 11 and 14 documents. Draw the R/F graph.
b) What is the expected recall and expected fallout after retrieving 7 documents (read off from the graph)?
c) What is the expected precision after retrieving 7 documents (you need to use the formula that expresses P as a function of R, F, and G)?
d) Using the theorem that R-norm is given by the area of the R/F graph that lies above R/F curve, but below the line represented by F = 1, compute R-norm (use the graph from part a. to compute the area of polygon of interest).

Q2: 15 Points

Let the retrieval output for some query, using a retrieval system be:

$$\Delta = (+ - + - | - - + + | + -+)$$

a) Calculate the PRECISION (defined as PRECALL) using preferred interpolation at recall values of 0.4, 0.7, 0.9, and 1.
b) Determine the PRECISION (defined as PRECALL) using ceiling interpolation at recall values of 0.4, 0.7, 0.9, and 1.

Q3: 10 Points

Assume that the documents have five relevance levels mostly relevant, relevant, mildly relevant, less relevant, non relevant in that order.
5 – Mostly relevant 4 –
Relevant
3 – Mildly relevant 2 – Less
relevant

1 – Non relevant

There are a total of 23 documents as shown in the following figure, for an arbitrary user need, the number of documents judged as mostly relevant, relevant, mildly relevant, less relevant, non relevant are four, six, five, five and three respectively.

$\Delta =$

5 4 3 2 2 1	5 4 3 2 1	5 4 4 4 3 2 1	5 4 3 3 2

Compute the R-Norm value according to the above information.

Q4: (15 Points)

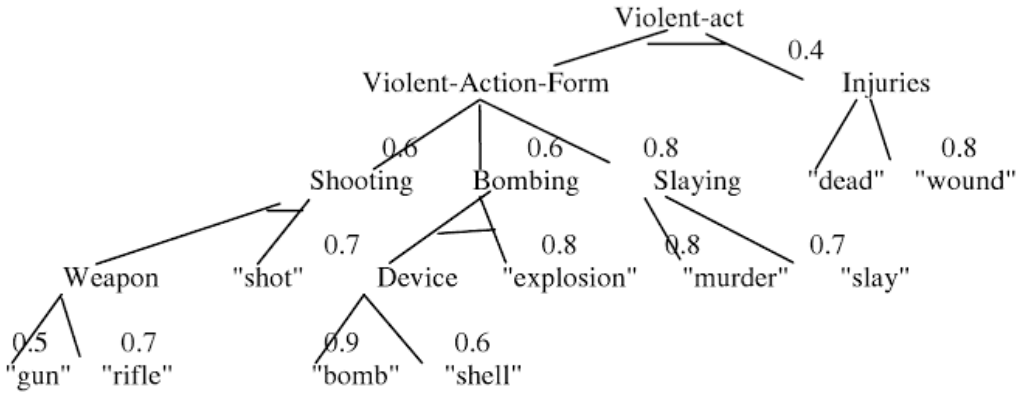
a) Write down the rules corresponding to the tree structure representing the concept of Violent-act (figure below)

b) Find the RSV of the document

D = (“rifle”,”bomb”,”slay”, “wound”)

with respect to the following concepts (using the computations as in the RUBRIC system): Bombing, Violent-Action-Form, Violent-act.

c) Generate a table as in the last page of RUBRIC (pdf) notes that shows the RSV for all the minimal (w.r.t. RSV) combinations of text expressions having non- zero RSV values relative to the concept Violent-act.



Q5: (40 Points)

A program that accepts any query from the query file and searches for documents based on the data structures from Assignment 1. Implement the search algorithm based on vector space model of retrieval. Your program should be able to return ranked results for a query. The program should display up to 20 results in descending order of their RSV values along with the total number of results[r1] (non-zero RSV values), snippet of text from document and time taken to retrieve those results.

Weighting Function

Use normalized frequency based on weight of term *i* in the document *j* is calculated as

$tf_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$, Where $1 \leq k \leq M$, M is the total number of words in document j and f_{ij} is the frequency of i^{th} term in j^{th} document.

Steps

1) Implement Vector Space retrieval model and inverted index based search algorithm

2) User should be able to select a query interactively and obtain ranked results

Deliverables

Code and accompanying documentation: Your code needs to be well documented with comments. The comments should not literally verbalize what the code is doing; instead, it should provide a high-level summary of what the code is supposed to accomplish, listing aspects such as tacit assumptions and invariants (if you are using sophisticated data structures). You also need to include a README.txt file that describes how to compile your program and run it on various datasets.

Implementation

You may code your project in any programming language (such as C#, C++, Java, Python, Perl, VB).