# COMP 4321 Search Engine for Web and Enterprise Data

# Homework Assignment 2

Released: Apr 24, 2017

# Due: May 4, 2017 (11:59pm)

**Submission Method: Answer the questions on papers and submit hardcopy to me during the lecture on May 2 or to the TAs directly.**

1. **[30]** Using the KMP algorithm, determine the number of character positions to shift in the shift array for the following pattern.
   * 1. Basic (before applying cascade mismatch):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **pattern** | **A** | **B** | **B** | **A** |
| **shift** |  |  |  |  |

* + 1. After applying cascade mismatch (ref: slides 28 and 29 of textscan-all.ppt):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **pattern** | **A** | **B** | **B** | **A** |
| **shift** |  |  |  |  |

Based on the shift table obtained in (b), show the steps in matching the pattern against the text string in the following table. In each step, underline the pattern character that causes the mismatch. You can see that there are three matches. Continue the match to identify all three matches (i.e., do not stop after finding the first match). You can add more rows if needed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Text -> | **a** | **b** | **c** | **a** | **b** | **b** | **a** | **a** | **b** | **b** | **a** | **b** | **b** | **a** | **a** |
| Pattern -> | a | b | **b** | a |  |  |  |  |  |  |  |  |  |  |  |
| Step 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. **[25]** 
   * 1. Fill in the following table. There are a total of 100 documents, and all of the relevant documents are shown in the table; they are marked with a √ in the first column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Rank** | **doc ID** | **Recall** | **Precision** | **Fallout** |
|  | 1 | 1001 |  |  |  |
|  | 2 | 2873 |  |  |  |
|  | 3 | 3916 |  |  |  |
|  | 4 | 0983 |  |  |  |
|  | 5 | 8310 |  |  |  |
| √ | 6 | 7892 |  |  |  |
|  | 7 | 4562 |  |  |  |
| √ | 8 | 4921 |  |  |  |
| √ | 9 | 7934 |  |  |  |
|  | 10 | 9248 |  |  |  |
| ... | . . . | ... | ... | ... | ... |
|  | 98 | 1688 |  |  |  |
| √ | 99 | 0926 |  |  |  |
|  | 100 | 3861 |  |  |  |

* + 1. Draw the precision/recall graph as in slide 18 of the lecture notes and show the graph after interpolation (ref: slide 2 of 000-performance-x.ppt).

* + 1. Consider ONLY the top 10 documents: what are (i) the top-10 precision, (ii) Average Precision (AP) using the Area-under-Curve (AUC) formulation, and (iii) nDCG10 where relevant documents get score 1 and non-relevant documents get 0.

1. **[20]** The following is a document-term matrix showing the weights of the terms in each document:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **T1** | **T2** | **T3** | **T4** |
| **D1** | 2 | 1 | 2 | 0 |
| **D2** | 1 | 3 | 0 | 1 |
| **D3** | 0 | 2 | 1 | 0 |

Initially, T1 and T2 have been chosen as the index terms.

* + 1. Compute the centroid of the three documents based on T1 and T2.
    2. Compute the average similarity of the three documents to the centroid based on inner product.
    3. Now you want to select T3 or T4 as the next index term. By recomputing average similarity as in (b), determine if T3 or T4 is a better index term based on their term discrimination values.

1. **[20]** Compute the mutual information of (i) “talent” and “pool”, and (ii) “legal” and “environment” in the following passage, considering a window size of 2 and word orders as given (i.e., talent must be in front of pool, and legal must be in front of environment).

Thirdly, talent is the key for Hong Kong's future development, both in the quality and quantity of the talent pool. This is the crux of where the city's future lies. Fourthly, environment, not merely a sound legal environment but also good ecological environment.

1. [25]The bipartite graph represents the places visited by a traveler.

Peter

Ocean Park

Disney

John

Mary

Peak

Buddha

* 1. Consider the graph as a web graph. Compute the hub and authority weights of each node (both person and place) assuming the initial weights of all nodes are equal to 1 in iteration 0. Carry out the computation in the first two iterations (iterations 1 and 2). Which person has the highest hub weight? Which place has the highest authority weight? In the computation, it is not necessary to normalize the weights in each iteration.
  2. How would you interpret the meaning of hub and authority weights in this traveler-place scenario if each link is a vote of the value of the place? Both John and Mary have visited two places, what can you say about differences of their traveling experience. Likewise, both Peak and Disney has attracted one visitor, which place has higher value? Why?