**Assignment 1**

1. In class, we briefly discussed pre-processing techniques such as stemming, stop-word removal and thesaurus construction. Given a text document, suggest any three additional pre-processing techniques that may be used. Explain the approach and outline the potential benefit of the approach.

**Answer:** There are numerous other techniques that could be used to pre-process data such that it is easy to retrieve

1. Dimensionality Reduction: In this pre-processing technique, we are concerned with reducing number of features to our system. When a system has a lot of data with a lot of features it affects the system performance later, this phenomenon is called curse of dimensionality. To tackle this problem, we need to reduce the number of features or “dimensions” and maintain variation in data.

Most widely used dimensionality reduction technique is called Principal Component Analysis. In this technique, we try to transform the data into a new co-ordinate system based on the initial data such that the variation of the initial data stays same.

1. Feature Engineering/Selection:

Suppose in your information system you have housing price data stored with various dimensions like age of the house, locality, length of the house plot, width, value, etc. We can make out over here that length and width of the house plot are related to the size hence we can combine the length and width dimensions to just one dimension area = length \* width. In this technique, we try to create better features that would help us create a better system.

1. Handling missing values: In many retrieval systems you can find data that has missing values like NA, NaN or NULL. You can handle such values using various statistical measures like mean, median or assigning them minimum or maximum values from the dataset.

2. Given the following small sample document collection:

(a) D1: Shipment of gold damaged in a fire

(b) D2: Delivery of silver arrived in a silver truck

(c) D3: Shipment of gold arrived in a truck

Calculate the term weightings for terms in D1. Show your workings and state any assumptions you make. (10 marks)

**Answer:**

Steps taken to generate term weighting for terms in D1.

1. Remove stop words from all the document as they don’t add much. And use stemming to get to the root word.

(a) D1: Shipment gold damage fire

(b) D2: Delivery silver arrive silver truck

(c) D3: Shipment gold arrive truck

2. Get the frequency of terms in the document.

Total number of terms: 13

Term Frequencies:

1. Use the calculated weights and apply them to terms in D1

D1: Shipment of gold damaged in a fire

Remove stop words and apply stemming.

D1: Shipment gold damage fire

Term 1: Shipment = 2/13

Term 2: Gold = 2/13

Term 3: Damage = 1/13

Term 4: Fire = 1/13

1. In class we discussed the document collection as term-document matrix, where each cell in the matrix indicates the usefulness of term i in describing document j. We also discussed how we could evaluate the similarity of a query and document. Outline a suitable indexing structure to store the information in the matrix (note that matrix is sparse). (10 marks) Outline at a high level, in pseudo-code, an algorithm to calculate the similarity of a document to a query. (10 marks)