Student Name:
NetID:

University of Texas at Dallas
CS 6322.001 : Information Retrieval
Spring 2016
Take Home Quiz # 3

Instructor: Dr. Sanda Harabagiu
Grader: Ramon Maldonado

Issued: April 18th 2016
Due April 20th 2016 in class

Query Expansion QUIZ

Use Automatic Local Analysis based on Metric Clusters to expand the following query:

Original QUERY: **earthquake Ecuador**

– When the local collection has 4 documents (N=4) V_l=??? S_l=???

Document 1: One person was dead after a magnitude-7.8 earthquake occurred Saturday evening on the coast of Ecuador.

Document 2: An earthquake with magnitude 7.4 occurred near Esmeraldas, Ecuador at 23:58:37.40 UTC on Apr 16, 2016.

Document 3: A 7.8-magnitude earthquake struck near Ecuador's coast Saturday, shaking homes 100 miles away in the capital of Quito and leaving 28 people dead.

Document 4: A powerful 7.8 magnitude earthquake shook Ecuador's central coast Saturday, cracking buildings and rattling homes as far away as the capital of Quito.

– Find what is the local vocabulary V_l=??? as well as the local stems S_l=???

Then, compute the distance between stems, given that *the distance $r(k_i, k_j)$ between two stem keywords k_i and k_j is given by the number of words between them in the same document. If k_i and k_j are in distinct documents we take $r(k_i, k_j)=\infty$.* This allows you to compute the correlations terms, using the formula:

$$\boxed{c_{u,v} = \sum_{k_i \in V(s_u)} \sum_{k_j \in V(s_u)} \frac{1}{r(k_i, k_j)}}$$

The correlation values between the stems from the local collection inform a correlation matrix from which you can infer the metric clusters for the keywords “earthquake” and “Ecuador”. You are required to generate the clusters of size 3, i.e. find the clusters *$S_u(n)$ defined as local metric clusters around each stem s_u from the original query, as we assume n=3 for each metric cluster!!!*

Show how you have built the metric clusters for keywords “earthquake” and “Ecuador” and show the expanded query that they determine.

In addition, generate the normalized metric clusters and the resulting expanded query when you consider the normalization formula:

$$s_{u,v} = \frac{c_{u,v}}{|V(s_u)| \times |V(s_v)|}$$