Problem Set 3

Assigned: Feb. 27
Due: Mar. 20

Problem 1


Problem 2

Web pages can be clustered according to link structure instead of textual content, or in addition to textual content. Consider the following clustering algorithm, starting with a directed graph of links

function ClusterByLink(in G: directed graph of links, k: integer; out S: set of k clusters) {
    make all the edges in G undirected;
    N = number of vertices in G;
    for each vertex V in G construct the N-dimensional vector \( \vec{V} \) where
    \[
    \vec{V}[U] = \begin{cases} 
    2 & \text{if } U = V \\
    1 & \text{if there is an edge } V - U \\
    0 & \text{otherwise}
    \end{cases}
    \]
    S = kMeansCluster(\{ \vec{V}_1 \ldots \vec{V}_N \}, k);
}

Show a trace of this algorithm running on the graph shown below. Assume that the initial clustering in the k-means algorithm is \{ \{ A,E,I \}, \{ B,F,G \} \{ C,D,H \} \}. Your trace should show the successive values of the clusters and of the centers.

![Diagram of a directed graph with vertices A, B, C, D, E, F, G, H, I and edges connecting them.]

1
Problem 3

Suppose you are evaluating a clustering scheme against a gold standard (see MR&S section 16.3). The gold standard has 3 clusters: A, B, C; the cluster has found two X, Y. These are related as follows

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>10</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Y</td>
<td>50</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

A. Evaluate the clustering in terms of purity.

B. Compute the quantities TP, TN, FP, FN. Note that since there are N=150 documents in the entire collection, we should have TP+TN+FP+FN = N(N-1)/2 = 11175. Using the equations in MSR, compute the Rand index, $P$, $R$, and $F_\beta$ with $\beta = 1$. 