TEXT MINING

ASSIGNMENT 4

EROL ÖZKAN

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

Introduction

In this assignment we try to implement a semi-supervised, transductive learning approach which assumes that each data point can be linearly reconstructed from its neighborhood. We propagate the labels from the labeled points to the whole dataset using their neighborhoods. We test our code on a small dataset as well as on the suplied Reuters dataset.

Graph Construction

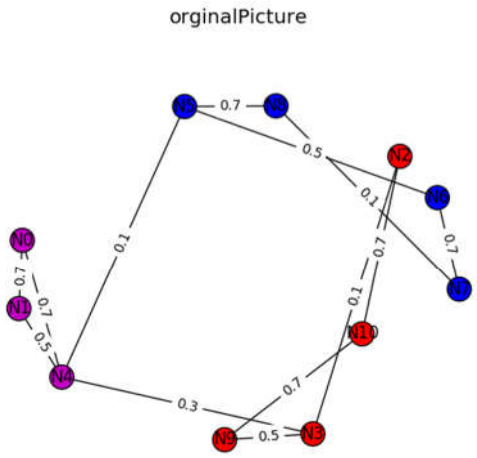
We start with building a graph G = {V, E} where the set of nodes V represents set of documents and E is the set of edges whose weights is the similarities between these documents. We build our graph based on k-NN graph scheme by assigning the most similar documents as edges to every node. So, in our graph, every node pair share an undirected edge when two nodes are k-nearest neighbors. We select k value as 3 in our tests. Also we do not take into account edges with weight below 0.

Table 1 shows an example input matrix.

Tablo 1 : Input Matrix

0	0.7	0	0	0.7	0	0	0	0	0	0
0.7	0	0	0	0.5	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0.7
0	0	1	0	0.3	0	0	0	0	0.5	0
0.7	0.5	0	0.3	0	1	0	0	0	0	0
0	0	0	0	1	0	0.5	0	0.7	0	0
0	0	0	0	0	0.5	0	0.7	0	0	0
0	0	0	0	0	0	0.7	0	1	0	0
0	0	0	0	0	0.7	0	1	0	0	0
0	0	0	0.5	0	0	0	0	0	0	0.7
0	0	0.7	0	0	0	0	0	0	0.7	0

Figure 1 shows the parsed graph based on this input matrix.



Şekil 1: Parsed Graph Visualization From Input Matrix

Sampling

For this example, there are 11 vertices in our graph, with the nodes labeled with one of three categories: namely; acq, earn, and interest.

Table 2 shows true labels for this example.

Tablo 2 : True Labels

Node Name	Node Label
N0	interest
N1	interest
N2	earn
N3	earn
N4	interest
N5	acq
N6	acq
N7	acq
N8	acq
N9	earn
N10	earn

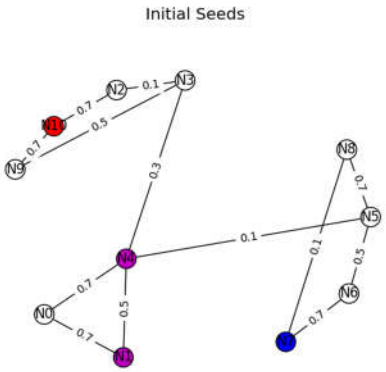
We sample this set by taking values after desired modulo operation.

Table 3 shows sampled set with “x%3==1”.

Tablo 3 : Sampled Set with "x%3==1”

Node Name	Node Label
N1	interest
N4	interest
N7	acq
N10	earn

Figure 2 shows sampled nodes in other words initial seeds in our graph. Here red node represents “earn” class while purple and blue nodes represensts “interest” and “ack” classes respectively.



Şekil 2 : Graph Visualization - Initial Seeds with "x%3==1”

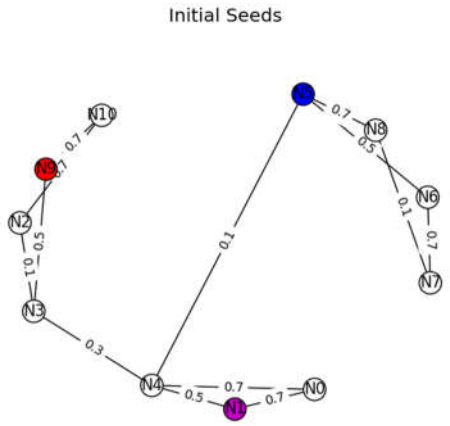
We also sample this set with a different modulo operation.

Table 4 shows sampled set with “x%4==1”.

Tablo 4 : Sampled Set with x%4==1

Node Name	Node Label
N1	interest
N5	acq
N9	earn

Figure 3 shows sampled nodes in our graph. Again, red node represents “earn” class while purple and blue nodes represensts “interest” and “ack” classes respectively.



Şekil 3 : Graph Visualization - Initial Seeds with "x%4==1”

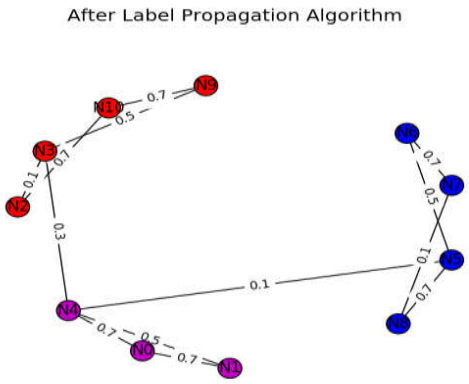
Results

We evaluate our label propagation algorithm and get the following result as shown in Table 6. We are able to get an accurancy of 1.

Tablo 5: Results For “x%3 == 1” & ACCURANCY = 1

time	nodeName	seedLabel	assignedLabel	trueLabel	result
2017-01-15 09:45:25.350074	N10	earn	earn	earn	True
2017-01-15 09:45:25.350134	N8	-	acq	acq	True
2017-01-15 09:45:25.350161	N9	-	earn	earn	True
2017-01-15 09:45:25.350177	N0	-	interest	interest	True
2017-01-15 09:45:25.350190	N1	interest	interest	interest	True
2017-01-15 09:45:25.350204	N2	-	earn	earn	True
2017-01-15 09:45:25.350216	N3	-	earn	earn	True
2017-01-15 09:45:25.350229	N4	interest	interest	interest	True
2017-01-15 09:45:25.350242	N5	-	acq	acq	True
2017-01-15 09:45:25.350254	N6	-	acq	acq	True
2017-01-15 09:45:25.350267	N7	acq	acq	acq	True

Figure 4 shows labeled graph after label propagation algorithm. Sampled nodes were "x%3==1".

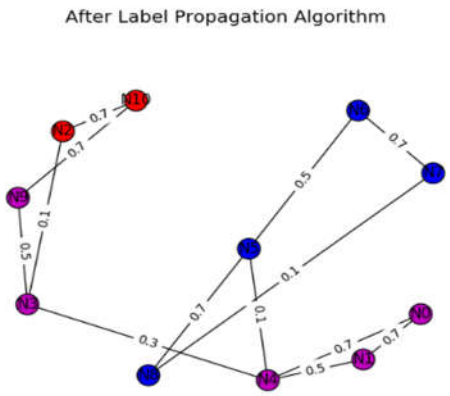


Şekil 4 : Graph Visualization - Results For "x%3==1"

We also evaluate our label propagation algorithm on set sampled with "x%4 == 1". We are able to get a accurancy of 0,81 here.

Tablo 6 : Results For "x%4 == 1" & ACCURANCY = 0,81

time	nodeName	seedLabel	assignedLabel	trueLabel	result
2017-01-15 09:53:29.098965	N10	-	earn	earn	True
2017-01-15 09:53:29.099050	N8	-	acq	acq	True
2017-01-15 09:53:29.099089	N9	earn	interest	earn	False
2017-01-15 09:53:29.099117	N0	-	interest	interest	True
2017-01-15 09:53:29.099141	N1	interest	interest	interest	True
2017-01-15 09:53:29.099163	N2	-	earn	earn	True
2017-01-15 09:53:29.099186	N3	-	interest	earn	False
2017-01-15 09:53:29.099210	N4	-	interest	interest	True
2017-01-15 09:53:29.099235	N5	acq	acq	acq	True
2017-01-15 09:53:29.099260	N6	-	acq	acq	True
2017-01-15 09:53:29.099285	N7	-	acq	acq	True



Şekil 5 : Graph Visualization - Results For "x%4==1"

Reuters Dataset Results

Further, we test our code with suplied Reuters dataset. In suplied dataset there are a total of 5.485 vertices in our graph, with the nodes labeled with one of eight categories: namely; acq, crude, earn, grain, interest, money-fx, ship, trade. We apply different modulo operations (x%2, x%20, x%100, x%500) for sampling and we get the initial seeds that has the following structure.

Tablo 7 : Modulo 2 Operation Seeds Information

Label	Number Of Seeds	Percentage In Seeds
earn	1415	%51.6046681255
money-fx	117	%4.26695842451
trade	135	%4.92341356674
acq	788	%28.7381473377
grain	22	%0.802334062728
interest	105	%3.82932166302
crude	108	%3.93873085339
ship	52	%1.89642596645

Tablo 8 : Modulo 20 Operation Seeds Information

Label	Number Of Seeds	Percentage In Seeds
earn	145	%52.7272727273
money-fx	9	%3.27272727273
trade	16	%5.81818181818
acq	77	%28.0
grain	1	%0.363636363636
interest	7	%2.54545454545
crude	12	%4.36363636364
ship	8	%2.90909090909

Tablo 9 : Modulo 100 Operation Seeds Information

Label	Number Of Seeds	Percentage In Seeds
earn	35	%63.6363636364
money-fx	2	%3.63636363636
trade	2	%3.63636363636
acq	11	%20.0
interest	1	%1.81818181818
crude	4	%7.27272727273

Tablo 10 : Modulo 500 Operation Seeds Information

Label	Number Of Seeds	Percentage In Seeds
earn	6	%54.5454545455
acq	4	%36.3636363636
interest	1	%9.09090909091

We calculate accurancy after applying label propagaion algorithm and get the following results.

Tablo 11 : Modulo 2 Operation Accurancy - 0.211850501367

Label	Number Of Nodes
trade	997
crude	829
earn	794
interest	704
acq	648
grain	590
money-fx	480
ship	443

Tablo 12 : Modulo 20 Operation Accurancy - 0.238468550593

Label	Number Of Nodes
acq	1446
earn	1033
trade	922
interest	700
ship	538
crude	480
money-fx	313
grain	53

Tablo 13 : Modulo 100 Operation Accurancy - 0.360984503191

Label	Number Of Nodes
earn	2141
acq	1828
crude	685
interest	432
money-fx	247
trade	152

Tablo 14 : Modulo 500 Operation Accurancy - 0.457064721969

Label	Number Of Nodes
earn	2973
acq	2319
interest	193

What we choose will propagate more than others…