Assignment 2

Problem 1: Signed Networks.

Consider the following simple model for constructing random signed networks, which we'll call the G^+ model. Start with a complete graph on n nodes, the sign of each edge e is marked as positive with probability p (and thus negative with probability $1-p$). All edges are undirected.

Let G_B denote the event that a graph G is balanced. In this question, we'll show that $P(G_B) \to 0$ as $n \to \infty$ for graphs generated according to the G^+ model. Assume that $p = \frac{1}{2}$.

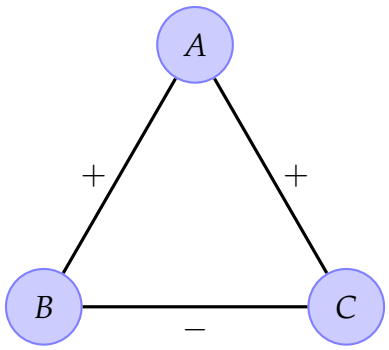
Q1: Let T be a maximum set of disjoint-edge triangles in G. A"disjoint-edge" set of triangles is one in which every edge is in exactly one triangle. Give a simple lower bound for $|T|$, the number of triangles that don't share any edges in G (the bound should be an increasing function of n).

Q2: For any triangle in G, what is the probability that it is balanced?

Q3: Using the simple lower bound from Q1, give an upper bound on the probability that **all** of the triangles in T are balanced. Show that this probability approaches 0 as $n \to \infty$.

Q4: Explain why Q3 implies that $P(G_B) \to 0$ as $n \to \infty$.

Q5: Now we want to construct a balanced signed network by dynamically growing the network, as new nodes join the network and create new signed edges to nodes already in the network. Consider the network shown in the following figure. Is it possible to add a node D such that it forms signed edges with all exisiting nodes (A, B and C), but isn't itself part of any unbalanced triangles? Justify your answer.



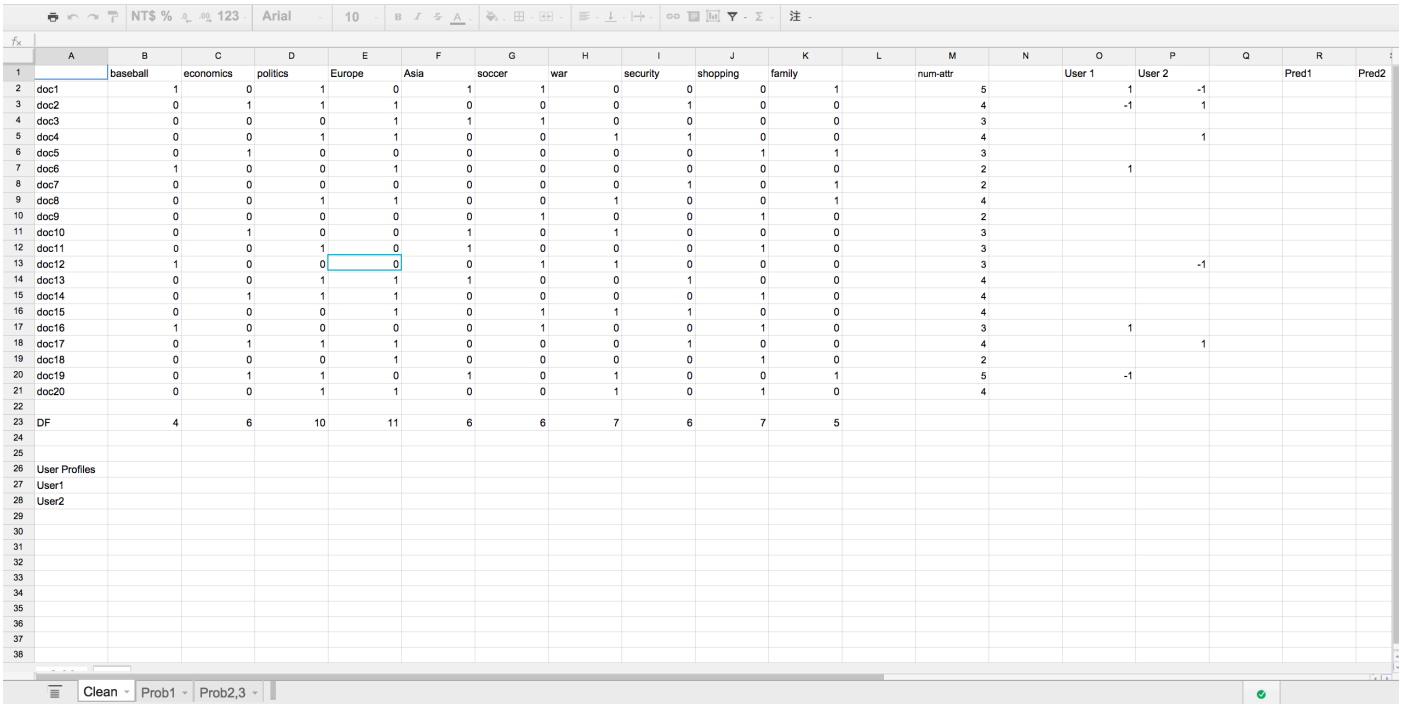
Q6: Using your answer from Q5, consider the following question. Take any complete signed network, on any number of nodes, that is unbalanced. When (if ever) is it possible for a new node X able to join the network and form edges to all existing nodes in such a way that it does not become involved in any unbalanced triangles? If it's possible, give an example. If it's not, explain why.

Problem 2: Recommender Systems.

In this problem, we will practice with content-based recommender system. To get yourself prepared for this problem, you may review the course and lab contents about recommender system. Also the following two links (clickable) are recommended to check:

- Machine learning Andrew NG @ Coursera Week 9
- Intro to Recommender Systems @ Coursera Week 3

The dataset will be worksheets as in the following picture



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		baseball	economics	politics	Europe	Asia	soccer	war	security	shopping	family		num-attr		User 1	User 2		Pred1	Pred2
2	doc1	1	0	1	0	1	1	0	0	0	1		5		1	-1			
3	doc2	0	1	1	1	0	0	0	1	0	0		4		-1	1			
4	doc3	0	0	0	1	1	1	0	0	0	0		3						
5	doc4	0	0	1	1	0	0	1	1	0	0		4			1			
6	doc5	0	1	0	0	0	0	0	0	1	1		3						
7	doc6	1	0	0	1	0	0	0	0	0	0		2		1				
8	doc7	0	0	0	0	0	0	0	1	0	1		2						
9	doc8	0	0	1	1	0	0	1	0	0	1		4						
10	doc9	0	0	0	0	0	1	0	0	1	0		2						
11	doc10	0	1	0	0	1	0	1	0	0	0		3						
12	doc11	0	0	1	0	1	0	0	0	1	0		3						
13	doc12	1	0	0	0	0	1	1	0	0	0		3			-1			
14	doc13	0	0	1	1	1	0	0	1	0	0		4						
15	doc14	0	1	1	1	0	0	0	0	1	0		4						
16	doc15	0	0	0	1	0	1	1	1	0	0		4						
17	doc16	1	0	0	0	0	1	0	0	1	0		3		1				
18	doc17	0	1	1	1	0	0	0	1	0	0		4			1			
19	doc18	0	0	0	1	0	0	0	0	1	0		2						
20	doc19	0	1	1	0	1	0	1	0	0	1		5		-1				
21	doc20	0	0	1	1	0	0	1	0	1	0		4						
22																			
23	DF	4	6	10	11	6	6	7	6	7	5								
24																			
25																			
26	User Profiles																		
27	User1																		
28	User2																		
29																			
30																			
31																			
32																			
33																			
34																			
35																			
36																			
37																			
38																			

There are 20 documents, each documents contains some features such as "baseball", "economics" etc, 1 means it contains this feature, 0 means does not contain this feature. So each document's profile is given in the worksheet, we need to first build user profiles for User 1 and User 2 based on their ratings, 1 means like and -1 means dislike and there are miss ratings here. After building user profiles, we need to get priction of ratings for all the documents for each user.

IMPORTANT: Please watch this video Assignment Details-Video and check this doc Assignment Details-Doc carefully.

Q1: Simple model: when building user profile, simply apply dot product between every feature column and user rating column (missing rating can be seemed as 0). What document does user 1 will like the most? And what is the prediction score of user 1 on that document? How many documents does this model predict user 2 will dislik (with negative prediction score)? You can use "Prob1" worksheet for this problem, for the next two problems, you can use "Prob2,3" worksheet.

Q2: Simple+Normalization model: Normalization here means we normalize each document profile, this can be seemed as taking TF score (weigh of each feature for each document) into account. The normalized document profiles are given in "Prob2,3" worksheet. Basically you only need to repeat the same steps of Q1. Which document is now in second for user 1 with this model? And what is the prediction score?

Q3: Simple+Normalization+IDF model: Now we are going to take IDF(Inverse Document Frequency) score into account. DF (Document Frequency) score is given in "Prob2,3" worksheet. Here we define IDF as 1/DF (more common is 1/log(DF) as taught in the course), but here we need a dramatic value to see differences with such small dataset. With this model, what is user 1's prediction score on document 9 (Compare document 1 and document 9 for user 1, you will see a dramatic difference from last model's prediction)? For user 2, Look at document 6, it was moderately positive before and now is slightly negative, what is the reason (single choice) ?

(a) because all the numbers went down in the new model

(b) because the article wasn't actually that good to begin with

(a) because the user's profile keeps changing

(a) because document 6 has two attributes, a common one the user really liks (Europe) and a rare one the user dislikes (Baseball). In prior models, the fact that user liked Europe more than s/he disliked baseball was decisive, but this model recognizes that Baseball is rarer than Europe, and therefore should have more weight (after all, there are plenty of other articles about Europe)