CSE 5337/7337 Information Retrieval and Web Search LJ Brown
Spring 2018: Homework 6
Due: 5/3/2018
Notation: (5337/7337 points)

1. (25/25 points) Consider the following web pages and the set of web pages that they link to:
   Page A points to pages B and C.
   Page B points to page C.
   All other pages have no outgoing links.
   Apply the PageRank algorithm to this subgraph of pages. Assume $\alpha = 0.15$. Simulate the algorithm for three iterations.

   See "Question_1.pdf".

2. (25/25 points) Assume the following training set:
   Food: "cherry pie"
   Food: "buffalo wings"
   Beverage: "cream soda"
   Beverage: "orange soda"
   Apply 3-nearest-neighbor (kNN) text categorization to the name "cherry soda". Show all the similarity calculations needed to classify the name, and the final categorization. Assume simple term-frequency weights (no IDF) with cosine similarity.

   Classification: "Beverage", see "Question_2.py" for computation.

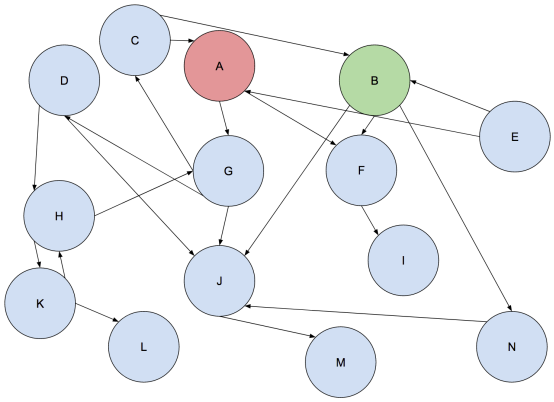
3. (25/25 points) You need to categorize vehicles into the following categories: truck, suv, and sedan using 5 features. Perform a naïve Bayes classification with the following probabilities. Show your work.

c	TRUCK	SUV	SEDAN
P(c)	0.35	0.4	0.25
P(f1|c)	0.2	0.01	0.2
P(f2|c)	0.01	0.1	0.05
P(f3|c)	0.1	0.001	0.005
P(f4|c)	0.001	0.2	0.005
P(f5|c)	0.005	0.008	0.01

   a. What category would you assign to the vehicle (f1, f2, f3)?
      TRUCK
   b. What category would you assign to the vehicle (f1, f2, f4, f5)?
      SUV

   See "Question_3.pdf" for explanation.

4. (25/25 points) Websites can have a good or bad reputation. What pages a site links to, can affect its reputation. There are sites often seen as “Good” or “Bad” based on their content and hyperlinks between pages. If the following graph represents the relationships between some webpages, the arrows between them are links, the A and B are marked as Good and Bad respectively, mark as many of the websites as possible as Good, Bad, or unknown reputations.



(END)