CAP 6777 Web Mining

2016 Summer

**Opinion Mining, Web Indexing and Search [15 points, Due: 06/23 (Firm)]**

**Part 1: Questions and Answers [0.75 pt/each]**

1. ~~What is sentiment~~
2. Please outline major steps of building an automated online review spam detection system for a website (such as Amazon.com).
3. What is a bag-of-word representation model? What is N-gram for representing a document? Explain the advantage and disadvantage of Bag-of-Word model vs. the N-gram model.

1. What is a Boolean retrieval model? What is a vector space model? Please explain their difference.
2. What is TF-IDF? What is the advantage of using TF-IDF for information retrieval?
3. Please explain the PageRank score formula. How does Pagerank score assess the importance of a webpage?
4. Please explain the disadvantage of using TF-IDF and Pagerank, alone, for information retrieval. Please propose one solution to combine TF-IDF and Pagerank to retrieve information online.
5. What is inverted index? When using inverted index for retrieval, what is the difference between document-at-a-time *vs.* query-at-a-time models.

**Part 2: Calculation** [**9 pts**]

**Question 1 [3 pts]:** Given following five documents, and a query Q, please use TF-IDF to calculate the matching score between query Q and each document (Please only consider following terms in your calculation Σ = {dish, duck, rabbit, Beijing, recipe, roast}.). Please show TF table and IDF table [2 pts], and show the TF-IDF scores between query Q and each document [1 pt]

Q: “Beijing duck recipe”

D1: “walks like a duck and quacks like a duck, must be a duck”

D2: “Beijing Duck is prized for the thin, crispy duck skin with authentic versions of the dish serving”

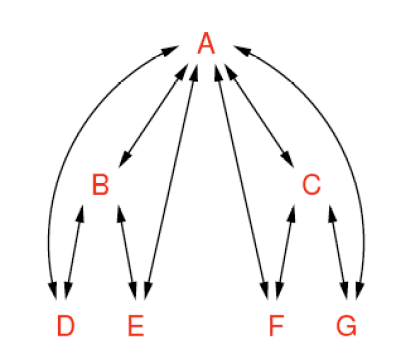
D3: “Bugs' ascension to stardom also prompted the Warner animators to recast Daffy Duck as the rabbit's rival, intensely jealous and determined to steal back the spotlight while Bugs remained indifferent to the duck's jealousy, or used it to his advantage. This turned out to be the recipe for the success of the duo.”

D4: “I found this great recipe for rabbit Braised in Wine on cookingforengineers.com”

D5: “Last week Li has shown you how to make the Sechuan duck. Today we'll show you a recipe for making Chinese dumplings, a popular dish that I had a chance to try last summer in Beijing.“

**Question 2 [4 pts]:** Given seven web pages with the following link structure

* please show the adjacency matrix and normalize the adjacency matrix [1 pt]
* Please use “Simple Iteration” to calculate the importance score for each website, after the first and the second iterations [The initial importance scores can be set as 1/n=0.15]. Please also report the difference of the pagerank score (using L1 norm) after the first and the second iterations, respectively. [2 pts]
* Please also use eigenvector based Pagerank algorithm to calculate the pagerank scores for each web page [1 pt] (please show your solutions.)



**Question 3 [2 pts]:** Because Pagerank scores play important role for search engines, some disreputable sites try to obtain a high pagerank score by playing different types of tricks. One example is to have many spam pages to point to a Page A, and further point Page A to Page B, as shown in the following figure.

* Please calculate pagerank scores for all pages (using either simple iteration or eigenvector based methods) [1 pt].
* Please explain how this approach can be used to mislead the search engine [0.5 pt], and explain how search engine can detect this type of misuse [0.5 pt]

