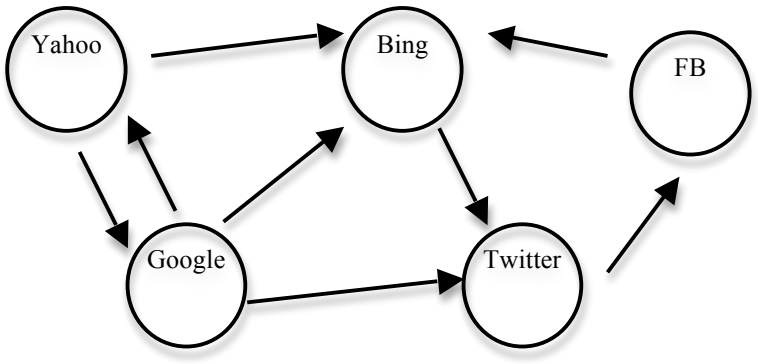
Homework #5 (due in 2 weeks)

Questions (40 points)

[1] Explain how detection of exact duplicate documents (web or otherwise) can be performed more efficiently than near duplicate detection, and without directly comparing the full text of each document to all other documents. Directly comparing against an entire collection of documents would be prohibitively expensive. (Hint: the method I am looking for can be applied to images or videos as well as text documents.)
[2] Explain how *shingling* is used to identify near duplicate documents in a large collection.
[3] Given the following directed graph of webpages, perform two iterations of PageRank computations. The arcs indicate outbound links between webpages. Initially give each page a PageRank score of 0.2. Use a 'teleport' (or transition) probability of 0.10. (Put differently, 90% of the time the random surfer follows a link to get to a new page). After each iteration you should normalize the scores. Normalizing means you should divide the PageRank score of each page by the sum of all PageRank scores so that the new total sums to 1.0. Show the PageRank scores for all five pages after each of the two iterations.



Web Query Log Analysis (60 points)

For this assignment you will work with a web query log from an Internet search engine (Excite) consisting of about two and a half million web queries. The main goal is to analyze the web queries and learn about user queries. You are allowed great flexibility in how you approach the task. You may use spreadsheets, databases, Unix commands (e.g., grep, wc), or custom written software programs.

Data
The dataset is available as a single gzipped tab separated value (TSV) file. Each line of the file consists of 4 tab delimited fields described below. The file is about 100MB uncompressed. The data are encoded as UTF-8 and most queries are in English. The data in the log file is unfiltered and may contain objectionable content.

Col.	Field	Description
1	Timestamp	6 digits indicating hour, minutes, and seconds (e.g., 9:30am is 093000). (The day is 12/20/1999.)
2	Userid	A hashed "user" ID, most likely an IP address.
3	First rank	0-based start of desired ranked list. The search engine would provide ten documents starting at this rank. 0 indicates the top 10 documents. 30 indicates ranks 30 to 39.
4	Query	The user provided search query. It may contain punctuation and whitespace.

First 15 lines:

090000	B0A0F80A06A3AB6C	0	In what year did baseball become an offical sport?
090000	95A33E619934A39B	0	wirehair pointing griffon
090000	E613C21C535BC636	30	ncic
090000	00CD4DE085A391DD	0	+ER +home +TV +Romano +picture
090000	5F48819400DB52D7	0	adolescent won't sleep in own bed
090000	D87CE5C149126B4B	0	where can i find free porno passwords
090000	47F6F715137F7C8D	0	play station codes . com
090000	40B1AACE633D9AC9	30	birth control and depression
090000	43D7E2332D3631DC	0	government
090000	87BE88FDCB1F7629	0	"WorkAbility I"+conference
090000	687340036669C45D	0	kitchen appliances
090000	E43DD6D82BFBD0B8	0	where can I find a chines rosewood
090000	CA52ECD1524E737D	0	jennifer love hewitt naked
090000	2B4FAF545C0E6EF0	40	pageant trim
090000	6456584F5B316AAE	100	tiger electronics

Last 15 lines:

165959	5F083C02AF42D762	10	"master boot record, fdisk"
165959	9F57839B10170414	0	beastiality
165959	2302407F00A4D6F9	870	www. lynn white.com
165959	4AD556DDCA079EB8	0	Where can i find crafts over the internet?
165959	FD647F92E62C1999	0	Where can I find a child's lilac sweater?
165959	3DF4E9B0AFF6B808	0	windows fix irq
165959	590F4121ABC62C02	0	what are the longterm physical effects of methamphetamine use in women?
165959	E43DD6D82BFBD0B8	0	chugach mountain
165959	4EB35F3114240AEE	110	alphabet hawiian
165959	8BA362CFA3B96117	0	"free internet access"
165959	C81EA097CF872C51	0	yahoo
165959	3365AF9F3B5CB4D9	0	alta vista
165959	32E290F942064B3A	0	body surface area drug dosage
165959	4D71604181DC294A	10	SLSA AND Australia
165959	302D2CA498C522F4	0	start up win95

Basic Analysis (45 points**)**
Please answer any ten (10) of the following questions (Q1 to Q13).
Q1. What is the average number of queries per user id?
Q2. Report the mean and median query length in both words and characters.
Q3. What percentage of queries are mixed case? All upper case? All lower case?
Q4. What percent of the time does a user request only the top 10 results?
Q5. What percent of unique queries are in the form of an explicit question (*i.e.,* look for patterns such as starting with Wh-words, or ending with a '?' symbol). What is the most common type of question?
Q6. What are the 20-most common queries issued?
Q7. What are the 20 most common non-stopwords appearing in queries?
Q8. What percent of queries contain stopwords like ‘and’, ‘the’, ‘of’, ‘in’, ‘at’?
Q9. What are the 10 most common non-stopwords appearing in queries that contain the word *download*?
Q10. What percentage of queries were asked by only one user?
Q11. Find 10 examples of misspelled words (but not 10 examples of the same misspelled word)
Q12. Which occurs in queries more often "Al Gore" or "Johns Hopkins"? "Johns Hopkins" or "John Hopkins"?
Q13. How often do URLs appear in queries?

Other Analysis (15 points)
Answer any three (3) of the following questions (Q14 to Q20).
Q14. Estimate the percentage of queries that contain a person's name?

Q15. Can you find addresses, phone numbers, and other identifiers in the log file? Is it likely that this web query log puts anyone's privacy at risk? Justify your response.
Q16. How often is search engine "query" syntax used, like phrases in quotes, Boolean operators, or '+' or '-' signs?
Q17. How often is a consecutive query a reformulation of the previous one? (Not the same query to greater depth.)
Q18. How does query volume change throughout the day?
Q19. What are the most popular websites mentioned in the queries?
Q20. Estimate the percentage of queries that are about sports?

Extra Credit (a few meager points)
You might come up with other interesting questions. Very interesting analysis might get a couple of measly extra-credit points.

What to turn in
Briefly describe the methods and tools that you used, and summarize any conclusions you reach. In addition to your analysis and any supporting data, examples, charts, graphs, etc..., please also provide any source code that you write for the assignment. Code quality is not very important for this programming assignment – the emphasis should be on the quality of the analysis and the clarity of the results. I don't expect this, but if you have significant trouble processing the entire file, work with a subset (e.g., a half, a quarter) and state that you couldn't process the entire file – I'll take off only a few points for not processing the complete file.