CS/INFO 1305 AI Lab 2
Due Monday, Aug 1, at 9:00pm

Information retrieval and web search

For this assignment, submit one document to CMS before the due time. This document should contain answers to all the enumerated questions. Each answer should be marked with the question number to which it corresponds and be 1–3 sentences long.

1 The vector space model

In this section we will explore information retained and lost using the vector space model.

1. What is the vector representation of this document: "Roses are red, violets are blue"?
2. What are two reasonable documents that could correspond to this document vector? {1:bites 1:dog 1:man} What is an unlikely document that could correspond to this vector? Write your example documents as English sentences, not as vectors.
3. What is this document about (more specifically than that it is about a Harry Potter book)? {1:24 1:8.3 1:Deathly 1:Hallows 1:Harry 1:Inc. 1:J.K. 1:Potter 1:Rowlings 1:Scholastic 1:States, 1:United 1:according 2:and 1:copies 1:final 1:first 1:hours 3:in 1:installment 1:its 1:million 1:on 1:popular 1:publisher 1:sale 1:series, 1:seventh 1:sold 3:the 1:to 1:wildly }
4. Suppose the user issued the query "What is the name of the first Harry Potter book?". How many words in this query overlap with the document represented by the above vector? Why is it not so important that the above document vector matches the query well?
5. Just for fun (no need to submit an answer): Is the following a positive or negative book review? Do not search for the original text of this review online. {1:Pride 1:Austen, 1:Everytime 6:I 1:Jane 1:Prejudice 3:and 1:beat 1:begin. 1:books 1:but 1:cant 1:conceal 1:criticise 1:dig 1:every 1:frenzy 1:from 1:have 4: her 1:madden 1:me 1:my 1:often 1:over 1:own 1:read 1:reader 1:shin-bone 1:skull 1:so 1:stop 1:that 2:the 1:therefore 1:time 3:to 1:up 2:want 1:with }

2 Comparing search engines

In this section we will compare two web search engines: yahoo.com and ask.com. Begin by choosing three queries. If possible, use real queries that you have made recently. Pose your queries to both search engines. Compare the returned pages.

6. What queries did you use?
7. Are the "real" search results returned by the two engines different, and if so how? ("Real" as in not ads and not sponsored.)
8. How are the sponsored links distinguished from the search results?

Now try some more advanced search features.

9. How would you search (using each of the two search engines) just Cornell-based websites for the phrase "information retrieval"?
10. How would you search (using each of the two search engines) for pages containing "Michael Jordan" but not the word "basketball"?
11. Have you ever used these two features (or any other "advanced" features) before?