CS/INFO 1305 AI Lab
Due Wednesday, July 18, at 10:00pm

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 (you don't need to be correct to get credit): 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: bing.com and google.com. Begin by choosing two queries. Pick ones that result in ads and/or sponsored documents being returned. Pose your queries to both search engines. Compare the returned pages.

6. What queries did you use?
7. Were the top three "real"—not ads and not sponsored—search results returned by the two engines different, and if so how?
8. How were the sponsored links distinguished from the real search results?

Now try an advanced search feature.

9. Using either search engine, how would you search for pages containing the words "tour" and "France" but not "de"? Describe where to find the advance search feature and which fields and/or operators you used. Was it easy to find the advanced search features?

3 NACLO: Maasai or Hawaiian

10. Submit your answers to the questions on *either* Maasai or Hawaiian.