CSE 5337/7337 Information Retrieval and Web Search L.J. Brown  
Spring 2018: Homework 5  
Due: 4/19/2018   
Notation: (5337/7337 points)

1. (15/15 points) Run the following ambiguous and under-specified queries on a web search engine of your choice. For each term, how many different interpretations appear in the top ten results?
   1. Jaguar

2

* 1. Windows

2

* 1. Hello

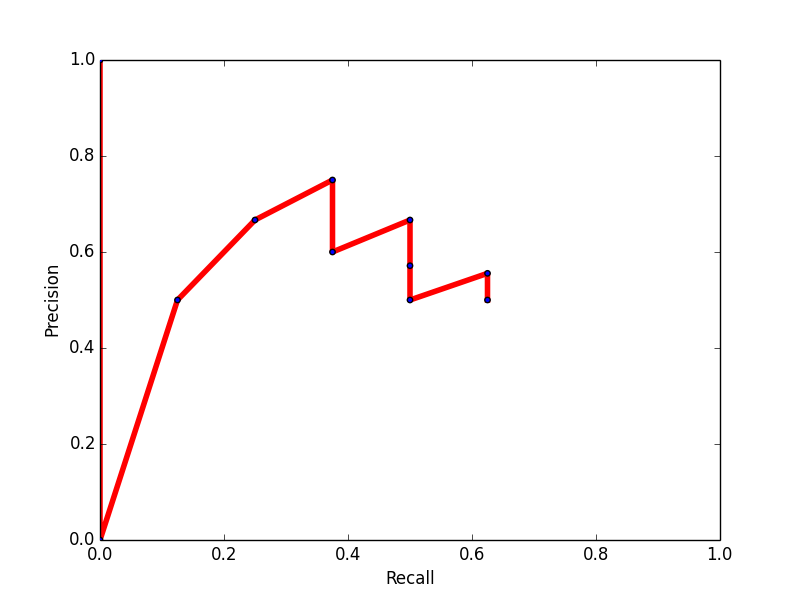
8

* 1. Trim

3

* 1. Book

5

1. (15/10 points) Consider a set of ranked documents R returned for the query q. Let R = { -, +, +, +, -, +, -, -, +, -} where “+” indicates a relevant document and “-” indicates a nonrelevant document. The system contains 8 relevant documents.
   1. Create the Precision versus Recall Curve for q. 
   2. What is the MAP value?

MAP: 0.573665

1. (15/15 points) The query-likelihood model scores documents based on the probability that the query is “generated” using the document’s language model. Suppose you have a collection of 4 documents (show below). Given the query “humpty dumpty”, what would be the score given to each document (assume case insensitive matching)?
   1. Doc1: Humpty Dumpty sat on a wall,

1/6 \* 1/6 = 1/36

* 1. Doc2: Humpty Dumpty had a great fall.

1/6 \* 1/6 = 1/36

* 1. Doc3: All the king's horses and all the king's men

0 \* 0 = 0

* 1. Doc4: Couldn't put Humpty together again.

1/5 \* 0 = 0

1. (25/20 points) Perform a K-means clustering (K=2) for the documents listed below. How many iterations are needed for the K-means to converge? List the docIDs for each cluster.

|  |  |
| --- | --- |
| docID document text | docID document text |
| 1 hot chocolate cocoa beans | 7 sweet sugar |
| 2 cocoa south america | 8 sugar cane brazil |
| 3 beans harvest south | 9 sweet sugar beet |
| 4 cocoa butter | 10 sweet cake icing |
| 5 butter truffles | 11 cake black forest |
| 6 sweet chocolate |  |

2 iterations to converge

Cluster 1: 1

Cluster 2: 2,3,4,5,6,7,8,9,10,11

1. (30/30 points) The following dataset identifies if a person will be able to play or not:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weather | Temp | Humidity | Windy | Play |
| Rainy | Cool | Normal | False | yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Hot | Normal | False | Yes |
| Overcast | Mild | High | True | Yes |
| Sunny | Mild | High | True | No |
| Sunny | Cool | High | False | ? |

Explain and show your work to determine how a Bayes Classifier would decide the last row.

Answer worked out in Question\_5.pdf

0.496 -> Yes

0.264 -> No

Play = Yes

1. (0/10 points) Consider two different search engines searching the same set of documents for the same query. System 1 returns {+,-,+,-,-,-,-,-,+,+}while System 2 returns {-,+,-,-,+,+,+,-,-,-}. “+” indicates a relevant document and “-” indicates a nonrelevant document. There are four relevant documents in the collection.
   1. What is MAPsystem1?
   2. What is MAPsystem2?
   3. Which system is “better”?

(END)