Information Retrieval Supervision 2 (2017/18)

Exercise 1

1.1 Consider making a language model from the following training text: *the martian has landed on the latin pop sensation ricky martin*

1.2 How might a language model be used in a spelling correction system? In particular, consider the case of context-sensitive spelling correction, and correcting incorrect usages of words, such as *Are you their?*

Exercise 2

doc1	phone ring person happy person
doc2	dog pet happy run jump
doc3	cat purr pet person happy
doc4	life smile run happy
doc5	life laugh walk run run

2.1 Smoothing is crucial in the language modeling approach to information retrieval. Why is smoothing important and how is it typically achieved?

2.2 Given the query {happy person smile}, show how a unigram language modeling approach would rank the documents outlined above. Choose a suitable form of smoothing and include all your workings. State any other assumptions made.

Exercise 3

The following list of Rs and Ns represents relevant (R) and non-relevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list is on the left of the list. The list shows 6 relevant documents. Assume that there are 8 relevant documents in the collection.

R R N N N N N N R N R N N N R N N N N R

3.1 What is the precision of the system in the top twenty?
3.2 What is the F1 on the top twenty?
3.3 What is the (uninterpolated) precision of the system at 25% recall?
3.4 What is the interpolated precision at 33% recall?
3.5 Assume that these twenty documents are the complete result set of the system. What is the AP for the query?
3.6 What is the largest possible MAP that this system could have?
3.7 What is the smallest possible MAP that this system could have?

Exercise 4

doc1	hot chocolate cocoa beans
doc2	cocoa ghana africa
doc3	beans harvest ghana
doc4	cocoa butter
doc5	butter truffles
doc6	sweet chocolate
doc7	sweet sugar
doc8	sugar cane brazil
doc9	sweet sugar beet
doc10	sweet cake icing
doc11	cake black forest

4.1 Perform K-means clustering for the documents in the table above.
4.2 After how many iterations does K-means converge?

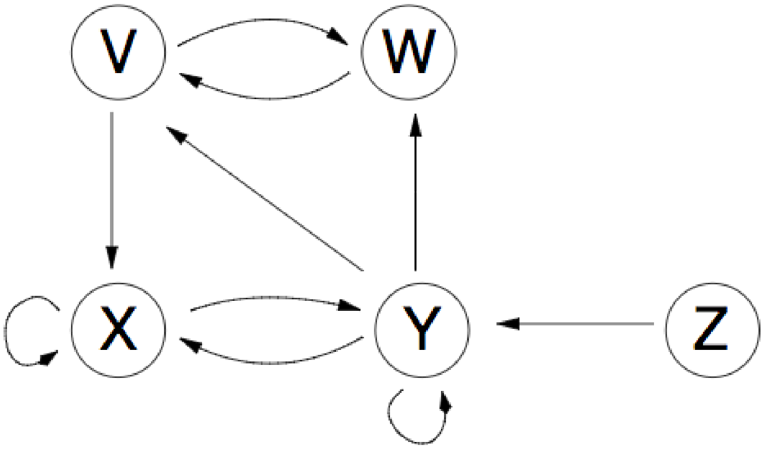
Exercise 5

The PageRank R of a website u is defined as:

$$R(u) = (1 - q) + q \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Here, B_u is the set of pages that points to u, N_u is the number of pages that u points to, and q is the probability of staying locally on the web page.

5.1 Explain the concept of PageRank, and how it is calculated.
5.2 Why is it relevant for web search?
5.3 Give, and briefly explain, the corresponding matrix notation of the PageRank computation.
5.4 Give the linkage matrix A of the network given in the diagram below.



5.5 Show the final matrix that will be subjected to the PageRank calculation, if $q = 0.8$ is used.