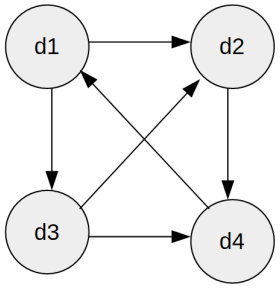
DATA.STAT.840 Statistical Methods for Text Data Analysis
Exercises for Lecture 10: Information retrieval

Exercise 10.1: Retrieval using TF-IDF and unigram language models.
Consider these four artificial documents:
d1: the robot is insightful but you are strong and i may answer and the wise fox is insightful and you are insightful and i am insightful but i will explain the insightful bird
d2: the bird is insightful
d3: when will they explain the friendly insightful strong insightful bird and is the bird strong and is a strong robot insightful
d4: a cat is strong but you are cautious and i may help but a fox is insightful but are they strong and when may you answer
In total, these four documents have the following vocabulary of 25 words:
'a', 'am', 'and', 'answer', 'are', 'bird', 'but', 'cat', 'cautious', 'explain', 'fox', 'friendly', 'help', 'i', 'insightful', 'is', 'may', 'robot', 'strong', 'the', 'they', 'when', 'will', 'wise', 'you'
Consider the **query: "**insightful bird".

a) Compute TF-IDF vectors for d1 and d2, using "raw count" for term frequency and "logarithmic inverse document frequency" for inverse document frequency, as described on Lecture 3 slides 37-38. Compute also a TF-IDF vector for the query. Then compute the cosine similarity between the query and d1 and d2. Which document is closer to the query?
b) Compute the unigram probability for the query given by d1 and d2, as discussed on Lecture 10 slide 14. In this exercise you do not need to apply smoothing to the probabilities. Which document gives larger probability to the query?

Report your computations and results.

Exercise 10.2: Pagerank.
Suppose the four documents d1, d2, d3, d4 of exercise 10.1 are webpages that link to each other with hyperlinks as shown in the picture below.



Use the equations of Lecture 10, slides 18-21 to solve a Pagerank prior for the document probabilities. Report your computation and the resulting prior probabilities of the four documents.

Exercise 10.3: Limitations and improvements of Pagerank.
a) Ask a large language model (see previous exercises for some possibilities) two questions: 1. what the limitations of the Pagerank algorithm are, and 2. in what ways the algorithm can be improved.

b) Find either from the course material or online material information that would support/agree with the answers that the large language model gave (or information that disputes/disagrees with the answers), and modify the answers to add the supporting/disputing information, with references to where you found it. Note: you do not need to add a massive amount of additional information, but try to add something for at least most of the points the large language model makes.

Provide the large language model's answers and your modifications including the added references; clearly mark the modifications you have made to the answers.