Name: Toghrul Tahirov Student Number: G47609664

EXERCISE 3 - MATHEMATICAL FUNCTIONS

Given the following four documents with the words in them which is the total set of documents we have:

D1: computer computer math computer computer science

D2: physics physics math math physics computer math

D3: physics math biology math science science biology computer biology

D4: physics physics physics physics math computer math computer

*PS. The code for all of the questions can be found in this github repository under weekly-assignments: https://github.com/DataMonarch/gwu-information-retrieval-systems

Questions:

1. **Across all four documents (looking at all the words in all of the documents) show the probability of computer, science, math, physics and biology (can leave in fraction form) occurring**

   Probability of the word computer: 0.26666666666666666
   Probability of the word math: 0.26666666666666666
   Probability of the word science: 0.1
   Probability of the word physics: 0.26666666666666666
   Probability of the word biology: 0.1

2. **What is the probability that each word occurs in a document (what per cent of the documents does each word occur in)**

   Probability of the word computer occurring in a document: 1.0
   Probability of the word math occurring in a document: 1.0
   Probability of the word science occurring in a document: 0.5
   Probability of the word physics occurring in a document: 0.75
   Probability of the word biology occurring in a document: 0.25

3. **What is average information for the words given probabilities in 1. Above. Three words are 8/30 and 2 words are 3/30.**

Average information for the word computer: 1.9068905956085187
Average information for the word math: 1.9068905956085187
Average information for the word science: 3.321928094887362
Average information for the word physics: 1.9068905956085187
Average information for the word biology: 3.321928094887362

4. **If all five words the words occurred equal likely instead of the above – what would the average information be and compare it to the average information from 3. above**

Avg_Info(x) = - log2(P(x))
Average information for the word computer (equal probs | real probas): 0.2 | 1.9068905956085187
Average information for the word math (equal probs | real probas): 0.2 | 1.9068905956085187
Average information for the word science (equal probs | real probas): 0.2 | 3.321928094887362
Average information for the word physics (equal probs | real probas): 0.2 | 1.9068905956085187
Average information for the word biology (equal probs | real probas): 0.2 | 3.321928094887362

5. **Can you calculate average information for 2 above explain your answer.**

Yes, we can calculate the average information for each word based on the probabilities of the words occurring in each document (as calculated in question 2). The average information measures the amount of uncertainty (in bits) associated with each word, and it is defined as the negative logarithm (base 2) of the probability of the word occurring. Here are the average info calculations for each word based on the 2nd question:

Average information associated with the word computer depending on its occurence in a document: -0.0
Average information associated with the word math depending on its occurence in a document: -0.0
Average information associated with the word science depending on its occurence in a document: 1.0

Name: Toghrul Tahirov Student Number: G47609664

Average information associated with the word physics depending on its occurence in a document: 0.4150374992788438
Average information associated with the word biology depending on its occurence in a document: 2.0

As we can see, the average information for the words "computer", "math", and "physics" (which occur in 3 out of 4 documents) is relatively low, which means that these words carry less information or more uncertainty (noise) than the words "science" and "biology" (which occur in only 1 or 2 documents). This is because the more often a word appears, the less uncertain we are about its occurrence in any given document. This concept is similar to TF-IDF, which is a measure of how relevant a word is in a given corpus considering its occurrence across the documents in the corpus.