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-assignment03: 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 (8/30)
   Probability of the word math: 0.26666666666666666 (8/30)
   Probability of the word science: 0.1 (3/30)
   Probability of the word physics: 0.26666666666666666 (8/30)
   Probability of the word biology: 0.1 (3/30)

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 .**
   Info(x) = - log2(P(x))

$$\mathrm{AVE\_INFO} = - \sum_{k=1}^{n} p_k \, \mathrm{Log}_2 \, (p_k)$$

Information value (IDF) of the word computer: 1.9068905956085185
Information value (IDF) of the word math: 1.9068905956085185
Information value (IDF) of the word science: 3.321928094887362
Information value (IDF) of the word physics: 1.9068905956085185
Information value (IDF) of the word biology: 3.321928094887362

Average information value of the whole corpus: 2.1898980954642875

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**

   Information value (IDF) of the word computer (equal probas | real probas): 2.321928094887362 | 1.9068905956085185
   Information value (IDF) of the word math (equal probas | real probas): 2.321928094887362 | 1.9068905956085185
   Information value (IDF) of the word science (equal probas | real probas): 2.321928094887362 | 3.321928094887362
   Information value (IDF) of the word physics (equal probas | real probas): 2.321928094887362 | 1.9068905956085185
   Information value (IDF) of the word biology (equal probas | real probas): 2.321928094887362 | 3.321928094887362

   Average information value of the whole corpus (equal probas | real probas): 2.321928094887362 | 2.1898980954642875

5. **Can you calculate average information for 2 above explain your answer.**
   It is not possible to calculate the average information value in this case if we assume that the probabilities of occurrence for words need to add up to 1.

   However, I believe the probabilities of occurrence for words in a corpus do not necessarily need to add up to 1 to calculate the average information value.

The calculation of average information involves taking the logarithm of the reciprocal of the probability of a word occurring. As long as the probabilities are positive, the calculation of the average information will be valid.

Nevertheless, the probabilities of all words in the corpus should be relative to each other. This means that the sum of probabilities does not have to equal 1, but the relative size of the probabilities should, mathematically, make sense. In other words, if one word has a higher probability than another, then it should also have a higher contribution to the average information value.

Despite this argument, I should also mention that if the probabilities do not add up to 1, the resulting average information values will not be directly comparable to the values obtained using different sets of probabilities

If we take into account the above argument, then, 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). Here are the average info calculations for each word based on the 2nd question:

```
Information value (IDF) of the word computer: -0.0
Information value (IDF) of the word math: -0.0
Information value (IDF) of the word science: 1.0
Information value (IDF) of the word physics: 0.4150374992788438
Information value (IDF) of the word biology: 2.0


Average information value of the whole corpus: 0.41067666647435835
```

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.

Name: Toghrul Tahirov   Student Number: G47609664