Assignment 2

Bhavneet Soni

1. What is TF.IDF (define or explain the concept)?

   TF.IDF stands for Term Frequency – Inverse Document Frequency, it’s industry standard parameter used to reflect importance of a word to a document in a collection. About 83% of text-based recommender systems use TF.IDF to give suggestions [1] . Term frequency (TF) is the frequency or number of times a word appear in a document, however using just the term frequency is misleading as some of the words such as “a”, “the”, “of” “in” etc. are used very often in documents however are of little significance. So to offset this issue we check how many times the word appear in other documents with in the collection, which is represented by Inverse Document Frequency (IDF). Its calculated by log of (total number of documents/ documents containing the word). What this term does is it gives a higher weightage to a term that is not so common. Hence the product term TF.IDF gives us a fair estimation of a words significance within the collection.

2. Suppose our collection include 2^{30} documents, and the word W appears in 2^{15} of them, calculate IDF of W? In document J, the word W appears 20 times, and the maximum number of occurrences of any word in this document is 40, calculate TF_{WJ}? How about TF.IDF for W in document J ? (Use base 2 logarithms in your calculations).

   a. Since word W appear in 2^{15} out of total documents of 2^{30}, IDF would be given by $\log_2(2^{30}/2^{15})$. So IDF of W is ➔ **15**.
   b. TF_{WJ} is the frequency of the word W in document J is given by number of f(w)/f(max). So for our problem TF will be given by 20/40 ➔ **0.5**
   c. TF.IDF is simply TW x IDF [0.5 x 15] ➔ **7.5**

3. Explain what the primary storage is and what the secondary storage is? Which one is faster? Why?.

   Primary storage also known as Random Access Memory (RAM), is the memory that the CPU can access directly. These usually consists of chips on the mother board (SDRAM, DDR3, DDR 4) and in some cases cache (L1, L2, L3) on CPU. Primary storage is indexed by memory registers and CPU u can directly access any address in the storage without delay and moving to get to any address hence the name random access. Primary storage are volatile memory blocks that lose the data once the power is cutoff. Whereas secondary storage is used to store data for longer duration like hard disk, tapes etc. Data (state) is preserved even when the power has been turned off, traditionally these used to be of moving parts Magnetic cylinders and tapes and to access them a physical read write had to be moved to the proper address to be able to be read. But recently solid state drives are becoming more common which do not have any moving parts making them faster than ever. But since they are further away from the CPU, secondary storage is considerably slower in comparison to primary storage.

4. A hash function h takes a hash-key value as an argument and produces a bucket number as a result. A common and simple hash function is h(x) = x mod B, what if x is not an integer, say a letter?

   A character or string is converted into its ASCI code and that is used to calculate its hash function

5. Explain what the power law is and give some examples.

   Power Law is a relationship between two things such that change in one leads to a proportional change in the other. Its very useful in understanding how increasing size of data increases its time complexity. eg most prominent use of these laws is in computer science and driving factor for development of new algorithms, if are we are trying to sort some data using insertion sort which has a time complexity of O(n^2), if the number of data points increase from 10 to 100, time complexity increase from 100 to 10,000, so we can say that there is square relationship between the two. Another example for such relationship is in the biological systems where the growth of population is closer to exponential over the generations close to doubling every generation

6. Based on the properties of the base of natural logarithms, calculate the approximation for the following:

$$1+2+\frac{4}{2}+\frac{8}{6}+\frac{16}{24}+\frac{32}{120}+\ldots$$

The term is of the form of Taylor equation of 2

$$\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\frac{x^5}{5!}+\cdots=1+x+\frac{x^2}{2}+\frac{x^3}{6}+\frac{x^4}{24}+\frac{x^5}{120}+\cdots=\sum_{n=0}^{\infty}\frac{x^n}{n!}.$$

Where x is 2. It can be expanded to e^2 ➔ **7.389**

References