INFO-533 Homework 2 (100 points)

Deadline: TBD, 2024 11:59 pm in Brightspace

1. Explain Binary Search Tree and its key properties. Highlight the differences between BST and Hash Table.

2. What is the Jaccard Coefficient? Using Bigrams, find out the Jaccard Coefficient between each pair in the following list of words:
   Jukebox
   Juicebox
   Jumbo
   Note: Do not use boundary symbols.

3. Using SPIMI Index construction method, build postings list for following documents:
   a. whispers of the wind echo through trees.
   b. the trees murmur with gentle calming whispers.
   c. echoes of the wind linger among rustling leaves.

   Show the results before sorting and after sorting.
   Note: Don't stem or do any preprocessing on terms, ignore memory handling & consider only 1 block

4. Using Map Reduce method, build postings list for following documents:
   a. whispers of the wind echo through trees.
   b. the trees murmur with gentle calming whispers.
   c. echoes of the wind linger among rustling leaves.

   Consider two parsers and two inverters with the first two documents in split 1 and third document in split 2. Each parser will partition terms in two disjoint sets, one starting with [a-i] and other starting with [j-z]
   Show the results of each parser for each split in the form of (term, docId).
   Show the postings list generated by each inverter

5. For this question, use Heaps' law, estimate the size of the vocabulary for the given collection of 10,000,000,000 documents containing 100 tokens on average. Consider there are 1,000 unique terms in the first 10,000 tokens and 10,000 unique terms in the first 1,000,000 tokens. Show the main steps of the computation.