*Midterm Problems -- Information Retrieval CSCI 6452*

1. Given the following: 0101101101110111 (**16 points** )

1. Generate 9 sistrings – the first is given below. Underline or circle the unique prefix for each sistring (**3 points**)

0 1 0 1 1 0 1 1 0 1 1 1 0 1 1 1

0101101101110111

101101101110111

01101101110111

1101101110111

101101110111

01101110111

1101110111

101110111

01110111

Unique Prefixes:  
010

10110110

0110110

110110

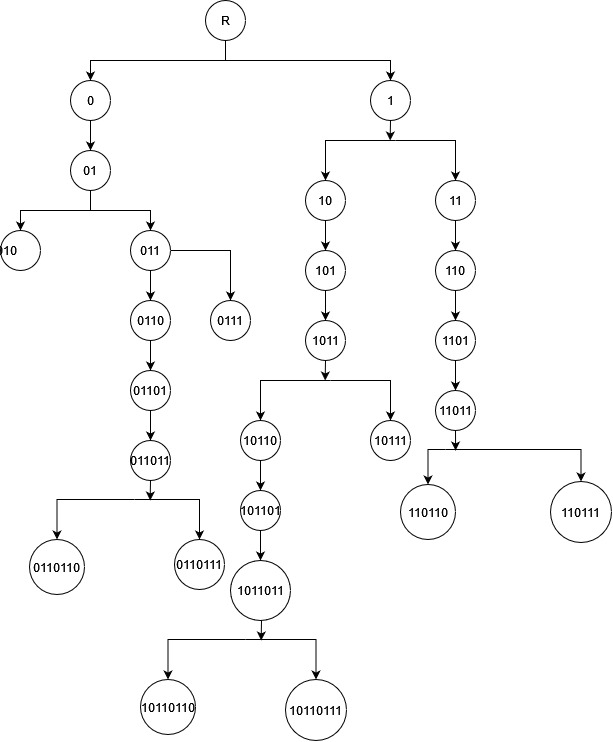
10110111

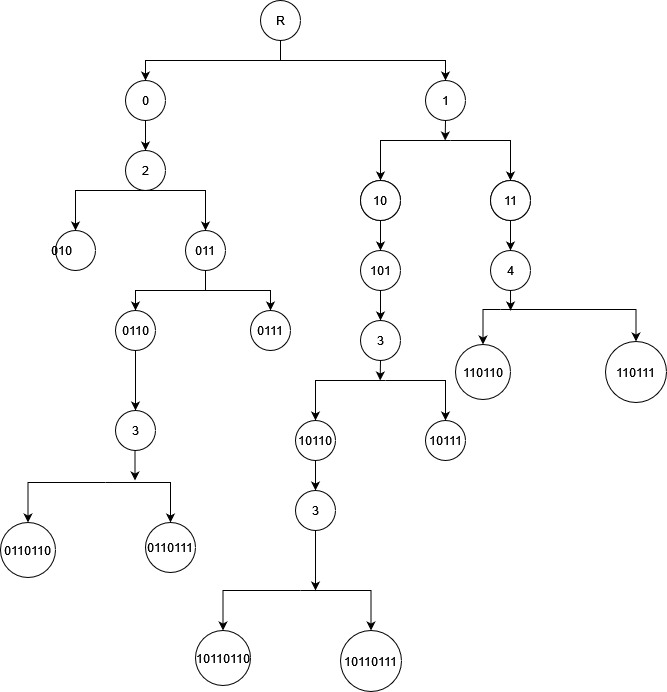
0110111

110111

10111

0111

1. Create the PAT Trie (4 points)
2. Create the reduced PAT trie (**4 points**)



1. Given following search string show number of compares using PAT and reduced PAT trie and result of search from above: 101101111 and what the result of the search – sistring satisfies it or not. (**5 points)**

**This string would require 8 compares when searched in the full PAT trie whilst requiring only 4 compares in the case of reduced PAT trie. Even though the prefix is found 101101110111 in the trie, the next set of characters do match. Thus, the sistring does not satisfy the search**

1. Given the following documents determine the weights for Naïve Bayesian category for document to be about “fruits”. Calculate the probabilities for all words for both in and not in the category – you can leave it as a fraction. Given the new document listed determine if it should be given the category or not. **LEAVE ALL CALCULATIONS AS A FRACTION** **(10 points - 7 points calculate weights; 3 points new document)**

Doc1 lemon, banana, banana, plum, plum, pear is member of fruits

Doc2 pear, banana, banana, plum, banana, plum is member of fruits

Doc3 lemon, banana, lemon, plum, lemon is member of fruits

Doc4 lemon, banana, lemon, lemon, plum, plum, lemon is member of fruits

Doc5 banana, lemon, pear, banana is NOT member of fruits

Doc6 lemon, banana, banana, lemon is NOT member of fruits

Doc7 banana, lemon, lemon, lemon is NOT member of fruits

Doc8 lemon, lemon is NOT member of fruits

**NEW DOCUMENT:** banana, lemon, banana

Text

Description automatically generated

Text

Description automatically generated

For the new doc:

P(fruit) = 1 / 2 \* [8/28 \* 9/28 \* 8/28] = 576 / 43904

P(not fruit) = 1 / 2 \* [6/18 \* 9/18 \* 6/18] = 324 / 11664

P(not fruit) > P(fruit) => new document is not a member of fruit

1. Given two documents – using 3 word shingles (e.g., w1w6w8 = 168) **(14 Pts: a and c = 5 points; b and d = 2 points)**.
   1. List the shingles for each document in 2 columns (one for each document)
   2. Determine the shingles in both documents and calculate the resemblance **leave as fraction**
   3. Select the 5 lowest shingle numbers from each document – write out them
   4. determine the resemblance – **leave as fraction**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W3 | W6 | W3 | W6 | W7 | W8 | W9 | W1 | W4 | W7 | W3 | W2 | W1 | W9 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W4 | W7 | W3 | W6 | W7 | W8 | W9 | W1 | W4 | W7 | W5 | W3 | W8 | W9 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. A screenshot of a computer

   Description automatically generated with medium confidenceGenerated the shingles using the below python code (results are given on the right)

A picture containing text, road, scoreboard, meter

Description automatically generated

1. A picture containing text

   Description automatically generatedResemblance calculated using the below python code (results displayed at the output) – **7/17**:
2. Generated using the below python code (results displayed at the output)
3. A screenshot of a computer

   Description automatically generated with medium confidenceA screenshot of a computer

   Description automatically generated with medium confidenceResemblance calculated using the below python code (results displayed at the output) – **2/8**

# Code for task 1

Text

Description automatically generated

Text

Description automatically generated