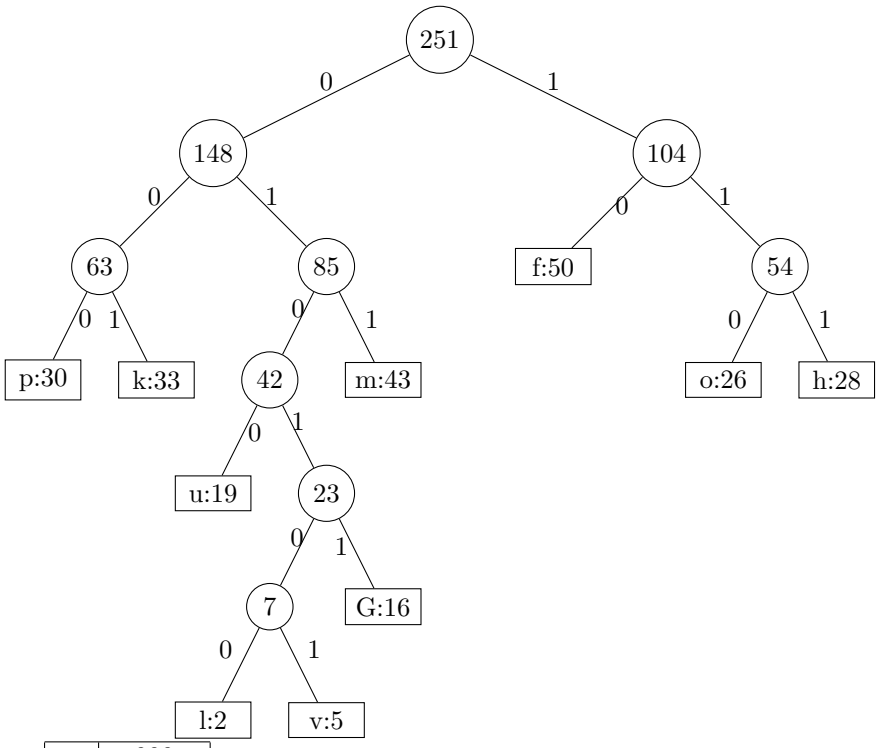
Information Retrieval: Assignment 4

Adrian Langseth

October 2019

Task 1: Text Compression



p	000
k	001
u	0100
l	010100
v	010101
g	01011
m	011
f	10
o	110
h	111

Canonical:

f	10
p	000
k	001
m	011
o	110
h	111
u	0100
g	01011
l	010100
v	010101

Decode the strings using the given huffman codes

100011000111001000111110111110100010011100 = INFORMATION
000101011100001000101011011011001 = RETRIEVAL

Task 2 : Index Analysis Using Lucene

Give a short explanation of Lucene

Lucene is a library used for text search. Through an API one can perform indexing, makes available search algorithms and search rankings.

Index the files found in the documents folder available on Blackboard. Copy the console output after doing the indexing. Explain the steps involved in the indexing process according to the source code.

Indexing to directory './svada.index'...
adding ./src/documents/doc9.txt
adding ./src/documents/doc8.txt
adding ./src/documents/doc10.txt
adding ./src/documents/doc6.txt
adding ./src/documents/doc7.txt
adding ./src/documents/doc5.txt
adding ./src/documents/doc4.txt
adding ./src/documents/doc1.txt
adding ./src/documents/doc3.txt
adding ./src/documents/doc2.txt
471 total milliseconds

It creates a standardAnalyzer, and the configuration of the index writer based on this analyzer. It then creates a indexwriter based on this configuration and indexes the documents in the directory.

After indexing the documents files, a search can be performed. This can be done by running the SearchFiles class.

Query: "Dog"

7 total matching documents
1. ./src/documents/doc2.txt
2. ./src/documents/doc3.txt
3. ./src/documents/doc8.txt
4. ./src/documents/doc10.txt
5. ./src/documents/doc6.txt
6. ./src/documents/doc1.txt
7. ./src/documents/doc4.txt

Query: "Small Big"

6 total matching documents
1. ./src/documents/doc10.txt

2. ./src/documents/doc9.txt
3. ./src/documents/doc5.txt
4. ./src/documents/doc6.txt
5. ./src/documents/doc1.txt
6. ./src/documents/doc4.txt

Query: "Big Cat Dog"

6 total matching documents

1. ./src/documents/doc7.txt
2. ./src/documents/doc5.txt
3. ./src/documents/doc4.txt
4. ./src/documents/doc10.txt
5. ./src/documents/doc6.txt
6. ./src/documents/doc1.txt

What query model is used here?

Boolean

Do returned documents contain all query terms together (AND) or are all documents containing any of them returned (OR)?

All returned documents contain one of them (OR)

Provide the query/queries you chose and time needed to perform the search on the Enron DB.

Searching for: dirt

Time: 24ms

422 total matching documents

1. ./src/maildir/germany-c/sent_items/200.
2. ./src/maildir/hendrickson-s/inbox/30.
3. ./src/maildir/hendrickson-s/sent_items/56.
4. ./src/maildir/davis-d/notes_inbox/50.
5. ./src/maildir/davis-d/discussion_threads/298.
6. ./src/maildir/davis-d/all_documents/70.
7. ./src/maildir/sanders-r/pacific_virgo/25.
8. ./src/maildir/sanders-r/all_documents/396.
9. ./src/maildir/sanders-r/pacific_virgo/28.
10. ./src/maildir/sanders-r/all_documents/405.