INFORMATION RETRIEVAL

Information Retrieval System

(Assignment 3)

Prepared by:

NAME	BRANCH	REG NO	ROLL NO
Suvigya Agrawal	CCE	150953130	41
Harsh Tandon	IT	150911144	16
Chaudhary Shyamal	CCE	150953066	35
Anvesha Saxena	IT	150911156	18
Bhavesh Bhansali	IT	150911124	13

Instructor : Poornalatha G Ma’am

Course : Information Retrieval

Date : 02-11-17

1. Problem Statement :

Develop a simple Information Retrieval System using Boolean retrieval model. Following are the functionalities to be implemented:

- Read the content of the document (minimum 10).
- Create inverted index where in postings list store gaps instead of doc id.
- The gap is represented using VB code compression technique.
- For a given query consisting of only one term, display compressed postings list consisting of gaps, also reconstruct doc ids from the postings list.
- Create appropriate interfaces for taking input and displaying results.

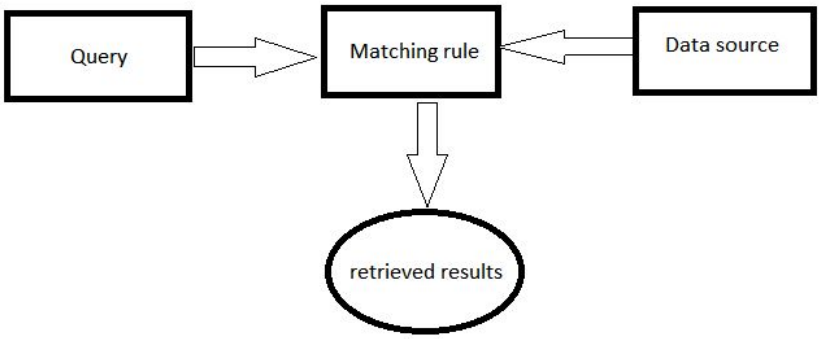
2. Explanation :

- In this project we designed and implemented a text based information retrieval system.
- Basically read the content of the document and convert inverted index where in posting list store gaps instead of document id. The gap list is created of the posting list.
- The gap is represented using VB code compression technique.
- For a given query consisting of terms, the corresponding VB Encoded gap list is retrieved. The gap list of top ranking documents are retrieved.
- Compressed postings list consisting of gaps are displayed.
- Document id from posting is reconstructed. And appropriate interfaces for taking input and displaying results is created.

3. Implementation :

Information Retrieval:

Information Retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for metadata that describe data, and for databases of texts, images or sounds.



basic model of an information retrieval system

Boolean Retrieval Model:

The boolean retrieval is the most simple of these retrieval methods and relies on the use of Boolean operators. The terms in a query are linked together with AND, OR and NOT. This method is often used in search engines on the Internet because it is fast and can therefore be used online. This method has also its problems. The user has to have some knowledge to the search topic for the search to be efficient, e.g., a wrong word in a query could rank a relevant document non relevant. The retrieved documents are all equally ranked with respect to relevance and the number of retrieved documents can only be changed by reformulating the query.

The boolean model of information retrieval is a classical information retrieval (IR) model and is the first and most adopted one. It is used by virtually all commercial IR systems today.

Each document id is called a posting and a set of document ids is a posting list. So, the most basic inverted index is a dictionary of terms each of which is associated with a posting list. It goes without saying that an inverted index is built in advance to support future queries.

Variable Byte (VB) Encoding:

Variable byte (VB) encoding uses an integral number of bytes to encode a gap. The last 7 bits of a byte are ‘’payload’’ and encode part of the gap. The first bit of the byte is a *continuation bit*. It is set to 1 for the last byte of the encoded gap and to 0 otherwise. To decode a variable byte code, we read a sequence of bytes with continuation bit 0 terminated by a byte with continuation bit 1. We then extract and concatenate the 7-bit parts. The idea of VB encoding can also be applied to larger or smaller units. For most IR systems variable byte codes offer an excellent tradeoff between time and space. They are also simple to implement.

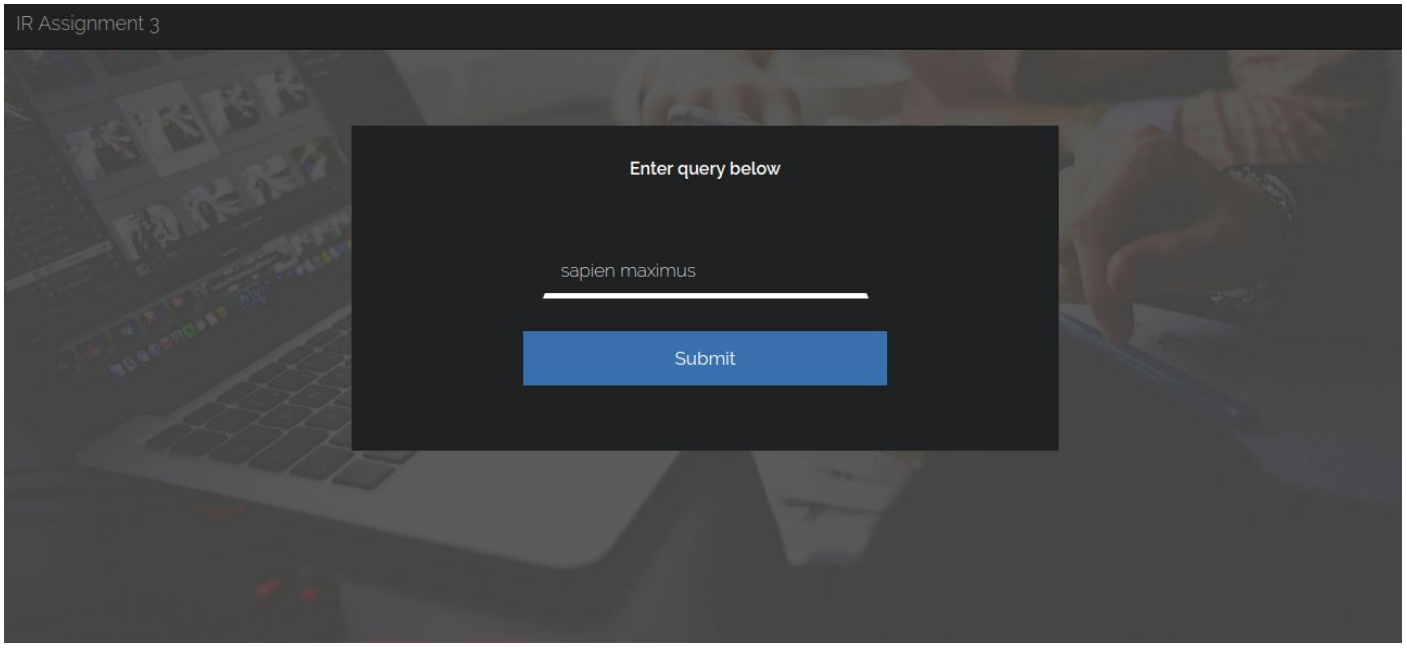
4. Psuedo Code :

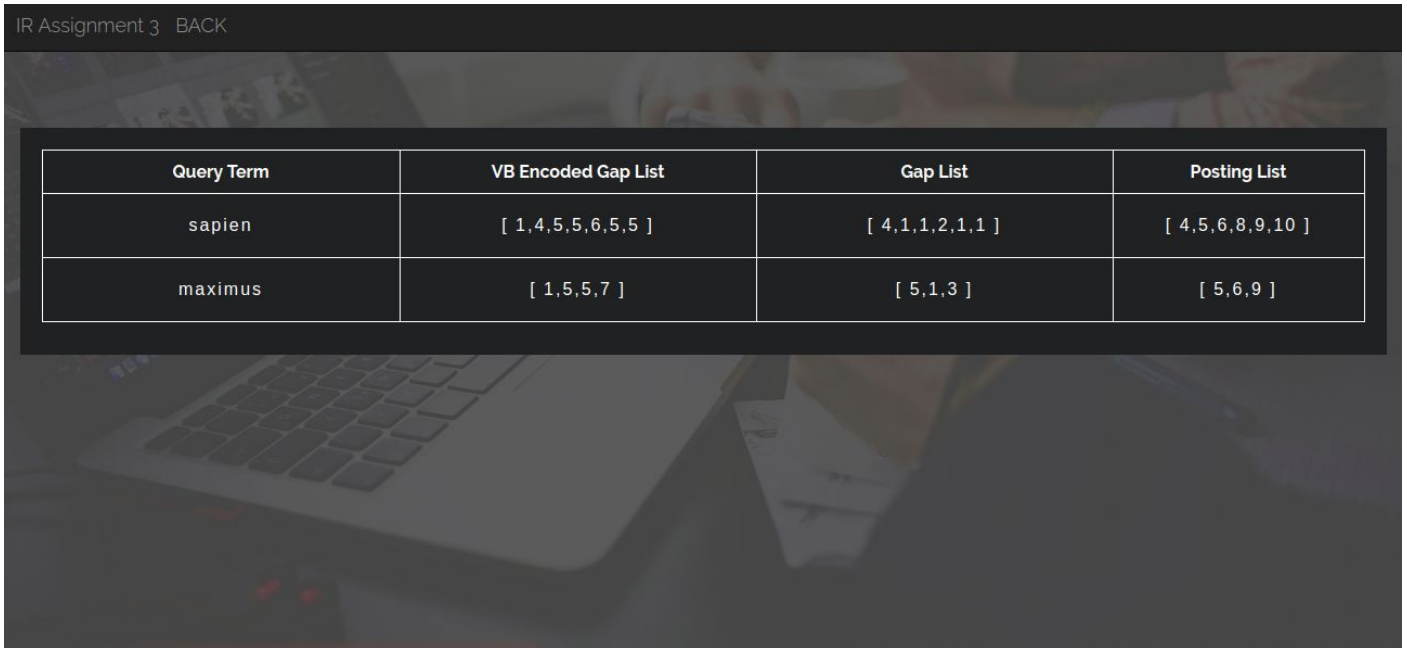
```
VBEncodeNumber(n)
1  bytes ← hi;
2  while (true)
3  do Pretend ( bytes, n mod 128 )
4   if (n < 128):
5       then break;
6   n ← n / 128;
7  bytes [len(bytes)] += 128;
8  return bytes;
```

```
VBEncode(numbers)
1  bytestream ← hi;
2  for each n ∈ numbers:
3  do bytes ← VBEncodeNumber(n)
4   bytestream ← Extend(bytestream, bytes);
5  return bytestream;
```

```
VBDecode(bytestream)
1  numbers ← hi;
2  n ← 0;
3  for i ← 1 to len(bytestream):
4  do if (bytestream [ i ] < 128):
5       then n ← 128 × n + bytestream [ i ];
6   else n ← 128 × n + ( bytestream [ i ] - 128 );
7       Append(numbers, n);
8       n ← 0;
9  return numbers;
```

5. Screenshots :





Query Term	VB Encoded Gap List	Gap List	Posting List
sapien	[1,4,5,5,6,5,5]	[4,1,1,2,1,1]	[4,5,6,8,9,10]
maximus	[1,5,5,7]	[5,1,3]	[5,6,9]

6. Conclusion :

In this project at first we define the meaning of Information Retrieval. Then how we implemented the information retrieval using boolean retrieval model using posting list and VB code. Using this technique we can retrieve documents related to the user queries. To process large document collections quickly. This is restricted to only documents but the amount of online data has grown at least as quickly as the speed of computers, and we would now like to be able to search collections that total in the order of billions to trillions of words.

7. Refrences :

https://www.researchgate.net/publication/232614940_A_Boolean_Model_in_Information_Retrieval_for_Search_Engines_PDF

https://nlp.stanford.edu/IR-book/pdf/01bool.pdf

https://www.codeproject.com/Articles/375219/Boolean-Retrieval-Model