Information Retrieval Index Implementation Report

Contents

[1 Directory structure of code 1](#_Toc513249845)

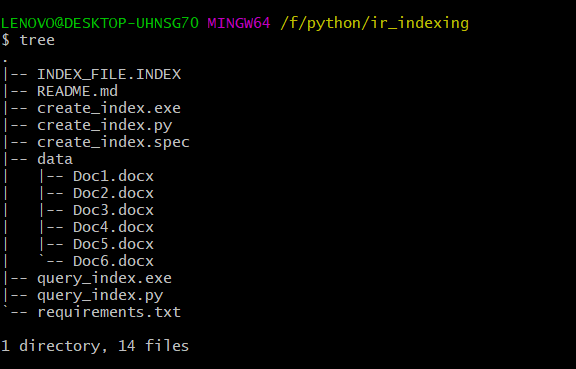
[2 Creating index 2](#_Toc513249846)

[2.1 Output of index operation 2](#_Toc513249847)

[3 Query 3](#_Toc513249848)

[3.1 Query outputs 3](#_Toc513249849)

# Directory structure of code



# Creating index

For each document:

1. Stop words are removed
2. Each word is lemmatized
3. Term is then added to the index in this format:

Because this is a small dataset, we can store and load entire index from one file.

{

"would": {

"df": 1,

"postings": ["Doc3.docx"]

},

"year": {

"df": 2,

"postings": ["Doc2.docx", "Doc6.docx"]

},

"young": {

"df": 1,

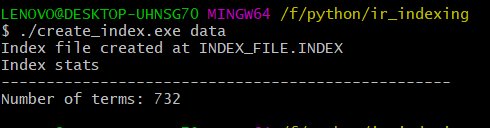
"postings": ["Doc3.docx"]

}

}

After removing stop words and lemmatization total number of terms obtained is **732**

## Output of index operation



# Query

For each term in the query , following operations are performed:

1. Stop words are removed
2. Terms are lemmatized
3. Retrieved the posting list for this term from the index (which is loaded into memory)

Then intersect the posting lists to obtain the result , because it’s a simple conjunctive Boolean query.

## Query outputs

