Information Retrieval Project Report

UTSAV GUPTA

|  |  |  |
| --- | --- | --- |
|  |  |  |
| BT09CSE079  [utsavgupta7@gmail.com](mailto:utsavgupta7@gmail.com)  **Tel** +91-8793966447 |  |  |
|  |  |  |

Contents

[Introduction 1](#_Toc370582410)

[Inverted Index & Schema Independent Inverted Index 2](#_Toc370582411)

[Phrase Matching 5](#_Toc370582412)

[Edit Distance Algorithm 7](#_Toc370582413)

[Contact information 8](#_Toc370582414)

# Introduction

“Brief Description Of the contents of the project.”

This project implements:-

* inverted index
* schema independent inverted index
* phrase matching
* edit distance algorithm

On a sample corpus to understand various information retrieval techniques and the underlying concepts.

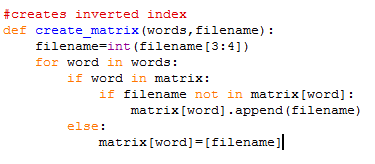
The entire project was implemented using python 2.7 and some basic modules provided by IDLE programming environment.

October 26, 2013

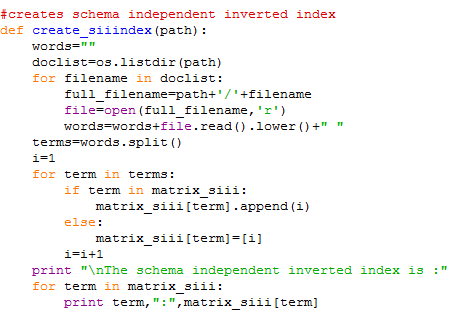
# Inverted Index & Schema Independent Inverted Index

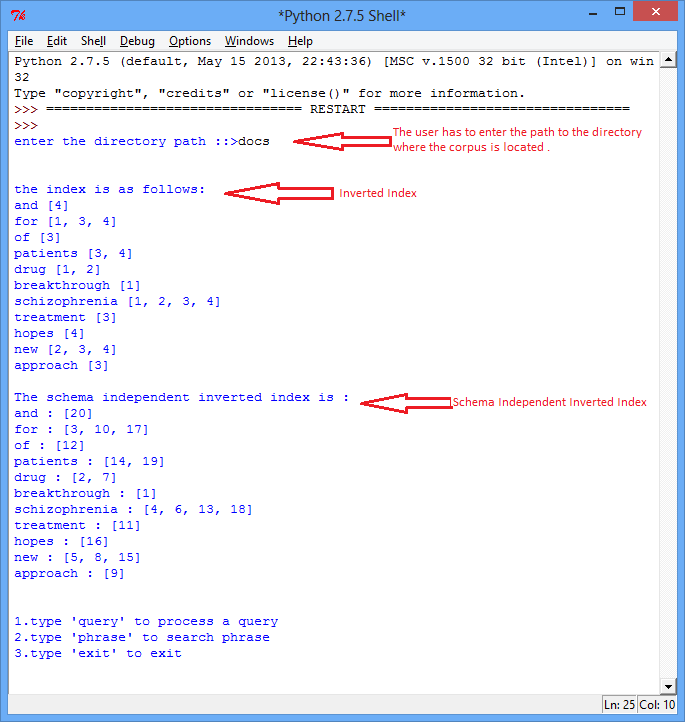
* An inverted index consists of “term vocabulary” and “posting list”. The term list consist of all the terms present in the corpus. The posting list for every term is simply the ID’s of all the documents in which the term occurs, sorted by the document ID’s.
* A schema independent inverted index simply combines all the documents to form one big document and then creates an index based on position of the terms in that one single combined document. The schema independent inverted index is used for phrase matching.

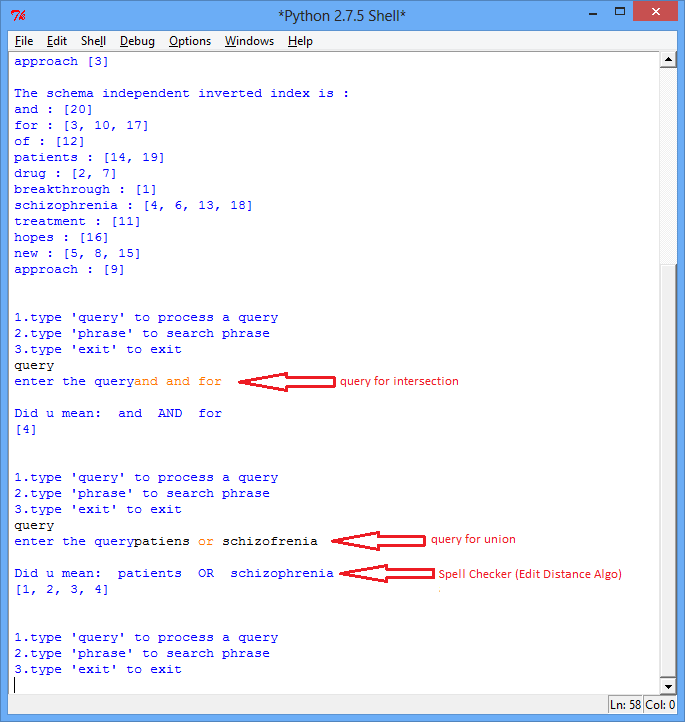
**Function for creating inverted index:**



**Function for creating schema independent inverted index:**

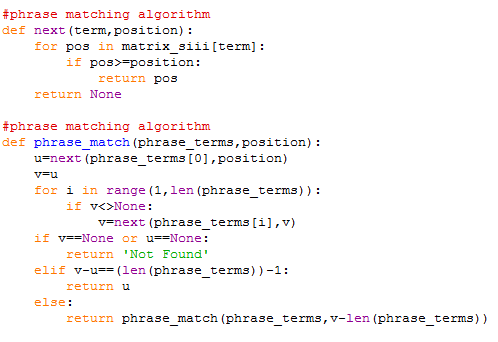
****

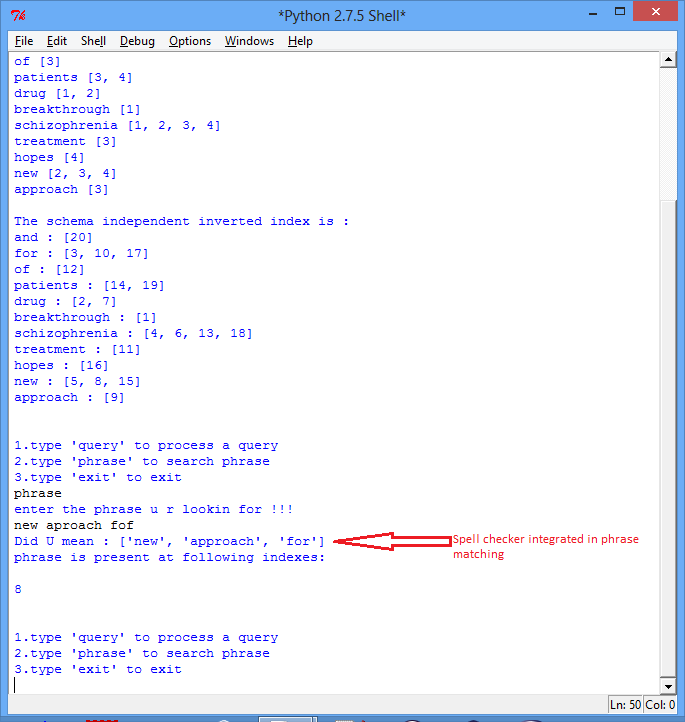




# Phrase Matching

**Function for phrase matching is as follows:**

****

****

# Edit Distance Algorithm



# Contact information

UTSAV GUPTA

BT09CSE079

Utsavgupta7@gmail.com

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
| **References**  <http://nlp.stanford.edu/IR-book/>  <http://www.python.org/> |  |  |  |