Query Likelihood Model

Maryam Mohmmad Alrashdi

443800993

Table of Contents

[INTRODUCTION 4](#_Toc88613450)

[Types of IR Models 5](#_Toc88613451)

[Components of Information Retrieval/ IR Model 6](#_Toc88613452)

[Query Likelihood Model 7](#_Toc88613453)

[The Goal of the project 7](#_Toc88613454)

[IMPLEMENTATION 8](#_Toc88613455)

[Overview of the techniques used in the project 8](#_Toc88613456)

[Code 9](#_Toc88613457)

[Result and output 11](#_Toc88613458)

[The outputs for Unigram 11](#_Toc88613459)

[The outputs for Bigram 12](#_Toc88613460)

[CONCLUSION 13](#_Toc88613461)

[REFERENCES 14](#_Toc88613462)

**List of figures**

[Figure 1 Types of IR models 5](#_Toc88596070)

[Figure 2 Components of Information Retrieval/ IR Model 6](#_Toc88596071)

# INTRODUCTION

Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e., usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.

Not only librarians, professional searchers, etc. engage themselves in the activity of information retrieval but nowadays hundreds of millions of people engage in IR every day when they use web search engines. Information Retrieval is believed to be the dominant form of Information access.

The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies regarding the existence and location of documents that might consist of the required information. Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents. The system searches over billions of documents stored on millions of computers. A spam filter, manual or automatic means are provided by email program for classifying the mails so that it can be placed directly into folders.

Information Retrieval (IR), (more precisely, text information retrieval) is a branch of computer science that deals with the processing of collections of documents containing ‘free text’, such as scientific papers, or even the contents of electronic textbooks. The objective of such processing is to facilitate rapid and accurate search of the text based on keywords of interest.

## Types of IR Models

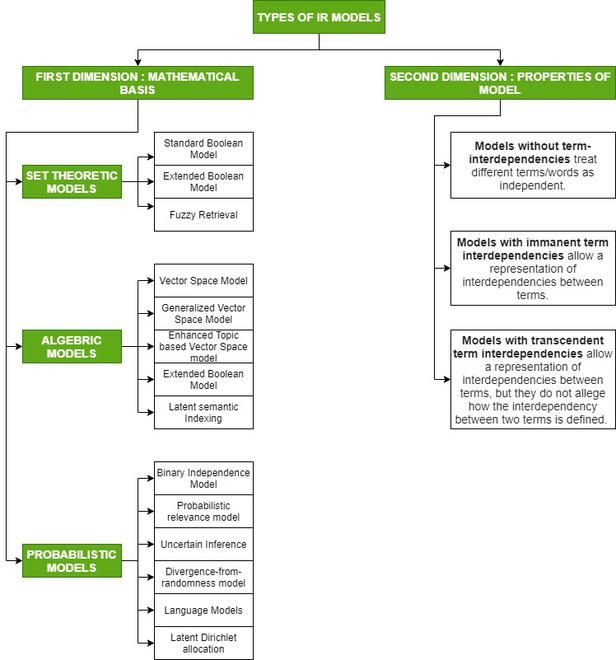


Figure 1 Types of IR models

## Components of Information Retrieval/ IR Model

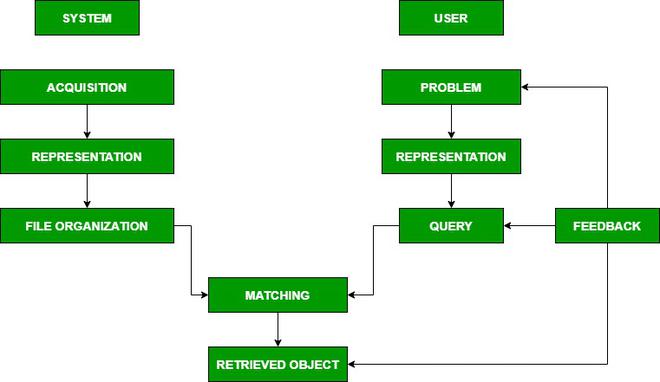


Figure 2 Components of Information Retrieval/ IR Model

In this project we will apply one of the language models (Query likelihood model).

## Query Likelihood Model

Over the past three decades, probabilistic models of document retrieval have been studied extensively. In general, these approaches can be characterized as methods of estimating the probability of relevance of documents to user queries.

The goal of a language model is to assign a probability to a sequence of words by means of a probability distribution”

The query likelihood model is a language model used in information retrieval. A language model is constructed for each document in the collection. It is then possible to rank each document by the probability of specific documents given a query. This is interpreted as being the likelihood of a document being relevant given a query.

The main purpose of this to put the core Information Retrieval concepts into practice by building and using our very own retrieval systems.

### The Aim of this project

The aim of this project is to build retrieval systems using the query likelihood model, which will show the percentage of the probability of the query being in each file.

# IMPLEMENTATION

## Overview of the techniques used in the project

* Query Likelihood: Foundation for this model is taken using Baye’s rule which states as below.

صورة تحتوي على نص, الأسلحة

تم إنشاء الوصف تلقائياً

* Removes all punctuations from a string.
* Stopping word: It’s the process of removing common words from the stream of tokens that become index terms. Stop words are taken into a list and while parsing the document for indexing, if the word is present in the stop word list, that word is not added to the inverted index.
* Tokenization: is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.
* Stemming: The task of stemming component is to group words that are derived from a common stem.
* Lemmatization: is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

Note: Lemmatization and stemming work in a similar way with a slight difference, so using one of them is enough, but in this project, I preferred to use them together for better results.

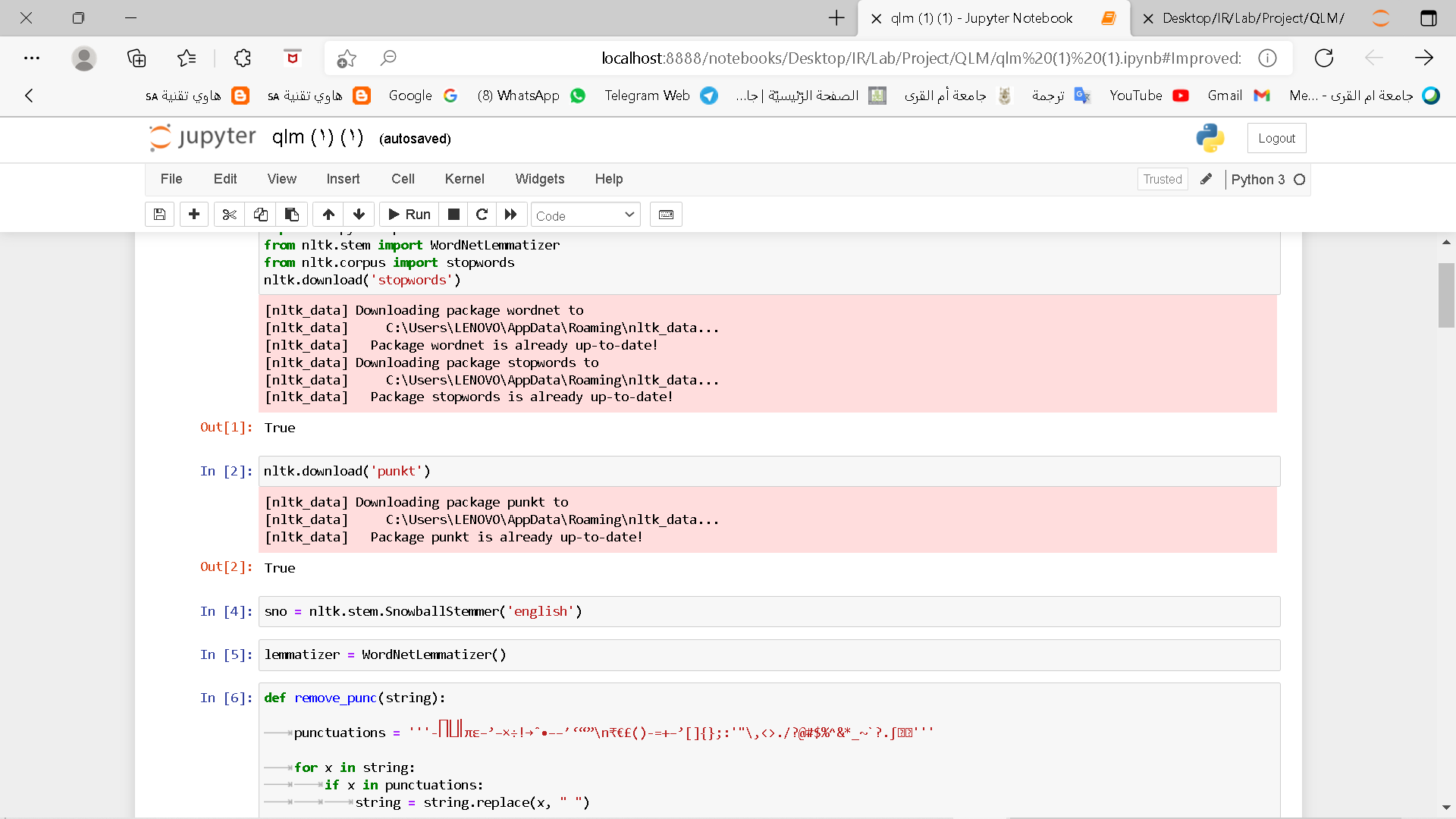
## Code

In this project, Query Likelihood Model has been implemented that contains all text processing.

First, we need to initialize and import all the libraries we need in the code.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً



Then we start by defining some of the functions we want to use on the documents.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

At this point we will start reading the files, the corpus containing all the text documents, and the query containing the queries we will apply on documents.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

After we defined the text processing above, here we begin to apply them to both the document and the query.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

Now we start executing Query Likelihood Model.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

## Result and output

The outputs and results are divided into two parts, the first for Unigram, and the other for Bigram.

### The outputs for Unigram

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

### The outputs for Bigram

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

# CONCLUSION

We have seen in this project the results of our application model, so we conclude from that the Unigram for these queries is much better than the use of Bigram, as we see in the second and third queries in Bigram, the possibilities appeared in this way because the query is one word for a Bigram model, so conditions do not apply to it.

We also see that not all the results for Bigram are correct, knowing that writing the code and the query are correct, which means that spars diagram probability occurs.

Spars diagram probability means for example the probability of finding the word "python" is higher than the probability of finding a bigram "python iter".

# REFERENCES

* NCBI: [https://www.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov/)
* Geeks for geeks: <https://www.geeksforgeeks.org>
* Microsoft Academic : [https://academic.microsoft.com](https://academic.microsoft.com/topic/2781111757/publication/search?q=Query%20likelihood%20model&qe=And(Composite(F.FId%253D2781111757)%252CTy%253D%270%27)&f=&orderBy=0)