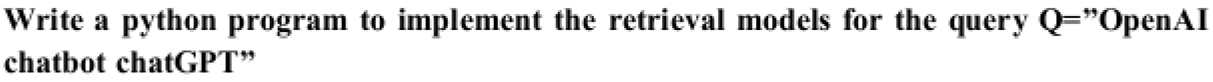
INFORMATION EXTRACTION AND RETRIEVAL

(IER)

Information Retrieval Model based on Bernolli Process

Problem Statement:



Write a python program to implement the retrieval models for the query Q="OpenAI chatbot chatGPT"

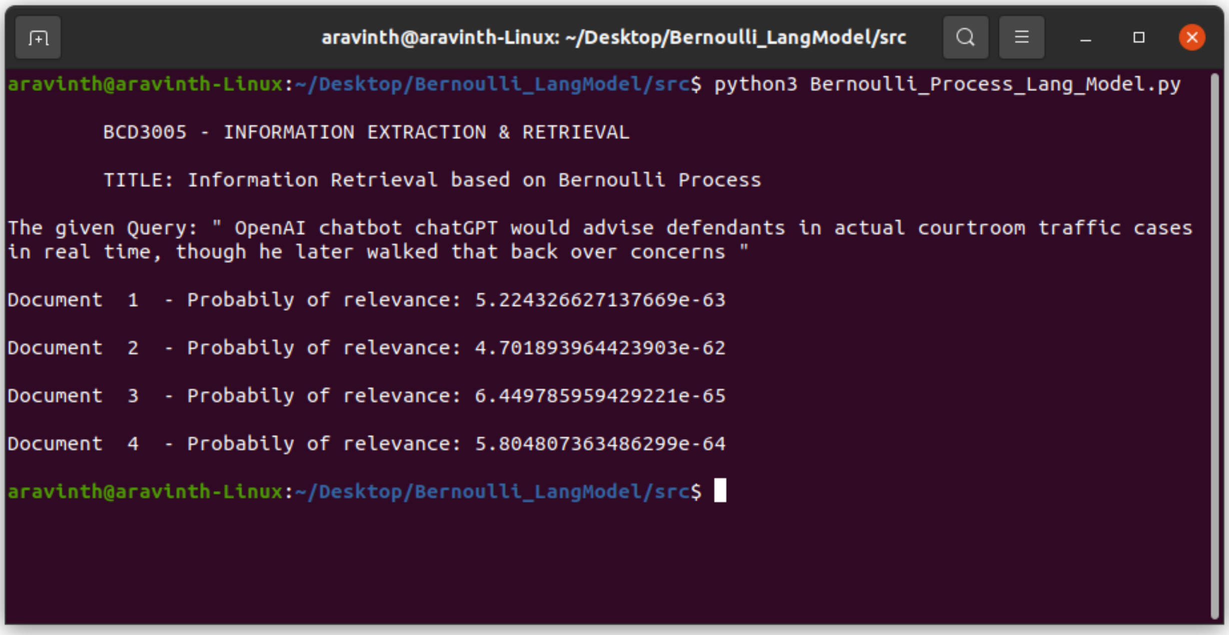
Implement Language model for information retrieval based on Bernoulli process. Consider P(ki/c) formula given by Miller , Leek and Schwartz. No need to consider concepts related to large documents and risk factors.

CODE

```
import numpy as np

# Define the documents
doc1 = "Since OpenAI released its blockbuster bot ChatGPT
in November, users have casually experimented with the
tool, with even Insider reporters trying to simulate news
stories or message potential dates.To older millennials
who grew up with IRC chat rooms – a text instant message
system – the personal tone of conversations with the bot
can evoke the experience of chatting online. But ChatGPT,
the latest in technology known as \"large language model
tools,\" doesn't speak with sentience and
doesn't \"think\" the way people do."

doc2 = "Other tech companies like Google and Meta have
developed their own large language model tools, which use
programs that take in human prompts and devise
sophisticated responses. OpenAI, in a revolutionary move,
also created a user interface that is letting the general
public experiment with it directly. Some recent efforts
to use chat bots for real-world services have proved
troubling – with odd results. The mental health company
Koko came under fire this month after its founder wrote
```

```
about how the company used GPT-3 in an experiment to
reply to users."

doc3 = "The founder of the controversial DoNotPay
service, which claims its GPT-3-driven chat bot helps
users resolve customer service disputes, also said an
AI \"lawyer\" would advise defendants in actual courtroom
traffic cases in real time, though he later walked that
back over concerns about its risks. Chat GPT is an AI
Chatbot developed by Open AI. The chatbot has a language-
based model that the developer fine-tunes for human
interaction in a conversational manner. Effectively it’s
a simulated chatbot primarily designed for customer
service; people use it for various other purposes too
though. These range from writing essays to drafting
business plans, to generating code. But what is it and
what can it really do?"

doc4 = "Chat GPT is an AI chatbot auto-generative system
created by Open AI for online customer care. It is a pre-
trained generative chat, which makes use of (NLP) Natural
Language Processing. The source of its data is textbooks,
websites, and various articles, which it uses to model
its own language for responding to human interaction. The
main feature of Chat GPT is generating responses like
those humans would provide, in a text box. Therefore, it
is suitable for chatbots, AI system conversations, and
virtual assistants. However, it can also give natural
answers to questions in a conversational tone and can
generate stories poems and more. Moreover, it can: Write
code, Write an article or blog post, Translate, Debug,
Write a story/poem, Recommend chords and lyrics"

# Define the query
query = "OpenAI chatbot chatGPT"

# Create the vocabulary
vocabulary = list(set(doc1.split() + doc2.split() +
doc3.split() + doc4.split() + query.split()))

# Calculate the document frequencies for each word
doc_freqs = np.zeros(len(vocabulary))
for doc in [doc1, doc2, doc3, doc4]:
    words = doc.split()
    for i, term in enumerate(vocabulary):
        if term in words:
            doc_freqs[i] += 1

# Calculate the probabilities of relevance for each
document
probs = np.zeros(4)
for j, doc in enumerate([doc1, doc2, doc3, doc4]):
    words = doc.split()
    p = 1
    for i, term in enumerate(vocabulary):
        if term in words:
            p *= doc_freqs[i] / 4
        else:
            p *= (1 - doc_freqs[i] / 4)
    probs[j] = p

print("\n\tINFORMATION EXTRACTION & RETRIEVAL\n\t\n\
nTITLE: Information Retrieval based on Bernoulli Process\
n")

print("The given Query: \"", query, "\"\n")
i = 1

for x in probs:
    print("Document ", i, " - Probabily of relevance:",
x, "\n")
    i+=1
```

OUTPUT



```
aravinth@aravinth-Linux:~/Desktop/Bernoulli_LangModel/src$ python3 Bernoulli_Process_Lang_Model.py

        BCD3005 - INFORMATION EXTRACTION & RETRIEVAL

        TITLE: Information Retrieval based on Bernoulli Process

The given Query: " OpenAI chatbot chatGPT would advise defendants in actual courtroom traffic cases
in real time, though he later walked that back over concerns "

Document  1  - Probabily of relevance: 5.224326627137669e-63

Document  2  - Probabily of relevance: 4.701893964423903e-62

Document  3  - Probabily of relevance: 6.449785959429221e-65

Document  4  - Probabily of relevance: 5.804807363486299e-64

aravinth@aravinth-Linux:~/Desktop/Bernoulli_LangModel/src$ 
```

~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~