ADAMAS UNIVERSITY

SCHOOL : ENGINEERING AND TECHNOLOGY.

DEPT : COMPUTER SCIENCE & ENGINEERING.

SUBJECT : ELECTIVE – VII LAB

(INFORMATION RETRIEVAL LAB)

(((LAB - 1)))

SUBJECT CODE : ECS44204

NAME : ANINDYA NAG.

COURSE : B.TECH

ROLL NO : UG/02/BTCSE/2018/005

ENROLMENT NO : AU/2018/02/0001809

SCE : (A).

YEAR : 4TH

SEMESTER : 8TH

DATE : 28.01.2021

INDEX

Lab: 01 Date:28.01.2022

Program No.:01

Program Name:

Write a Python Program to count unique words and number of their occurrence in text.

Input: **path to a text file (.txt)**

Output:

1) Total number of words.

2) Top 10 most frequent words

3) Number of occurrences for the top 10

Program Code:

```
#ANINDYA NAG
#Roll:005

import re
from collections import Counter

def count_words(path):
  with open(path, encoding='utf-8') as file:
    total_words = re.findall(r"[0-9a-zA-z-']+", file.read())
    total_words = [word.upper() for word in total_words]
    print('\nTotal Words:', len(total_words))

    word_counts = Counter()
    for word in total_words:
      word_counts[word] +=1

    print('\nTop 10 Words:')
    for word in word_counts.most_common(10):
      print(word[0], '\t', word[1])

count_words('Anindya_lab1.txt')
```

Output:

```
Total Words: 590

Top 10 Words:
THE   45
INFORMATION   20
IN      19
OF      19
RETRIEVAL       15
A        14
TO      12
FOR   12
ARE   9
BY      9
```

Input Text File:

Anindya_lab1.txt

Information retrieval (IR) in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those 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 the metadata that describes data, and for databases of texts, images or sounds.

Automated information retrieval systems are used to reduce what has been called information overload. An IR system is a software system that provides access to books, journals and other documents; stores and manages those documents. Web search engines are the most visible IR applications.
An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.[1]

Depending on the application the data objects may be, for example, text documents, images,[2] audio,[3] mind maps[4] or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

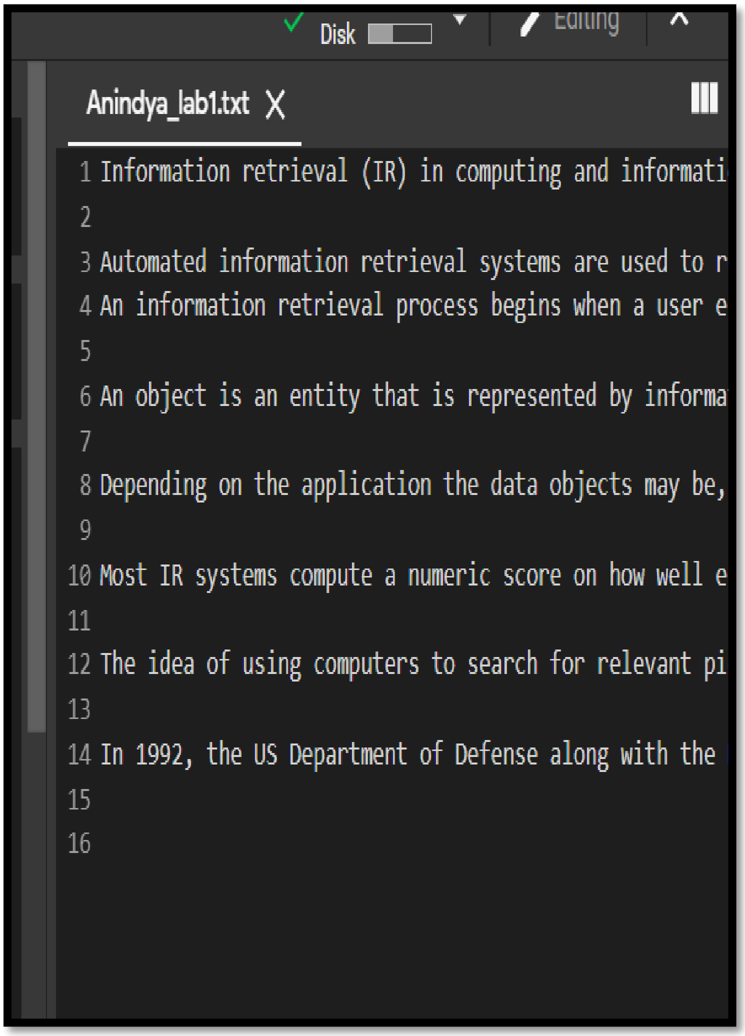
Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.[5]

The idea of using computers to search for relevant pieces of information was popularized in the article As We May Think by Vannevar Bush in 1945.[6] It would appear that Bush was inspired by patents for a 'statistical machine' - filed by Emanuel Goldberg in the 1920s and '30s - that searched for documents stored on film.[7] The first description of a computer searching for information was described by Holmstrom in 1948,[8] detailing an early mention of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one even featured in the 1957 romantic comedy, Desk Set. In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s several different retrieval techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents).[6] Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

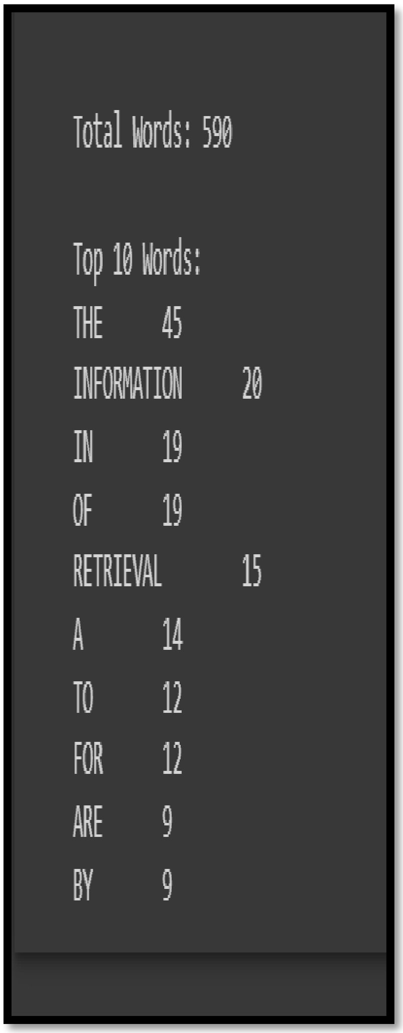
Screenshot’s:

Input: (Text File):



```
Information retrieval (IR) in computing and informati

Automated information retrieval systems are used to r
An information retrieval process begins when a user e

An object is an entity that is represented by informa

Depending on the application the data objects may be,

Most IR systems compute a numeric score on how well e

The idea of using computers to search for relevant pi

In 1992, the US Department of Defense along with the


```

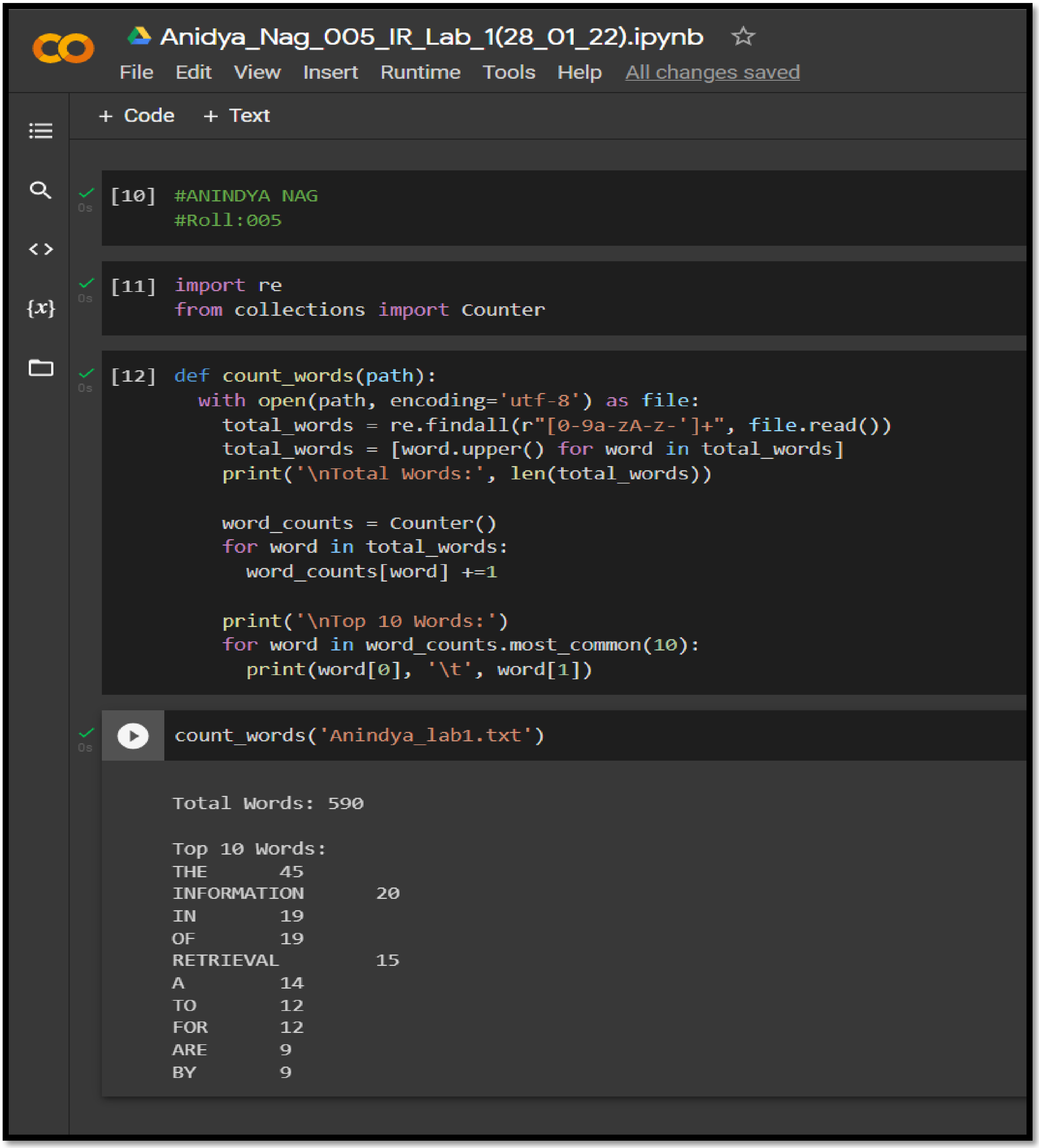
Output:



```
Total Words: 590


Top 10 Words:
THE         45
INFORMATION     20
IN          19
OF          19
RETRIEVAL       15
A           14
TO          12
FOR         12
ARE         9
BY          9
```

Screenshot's:

Code:



Anidya_Nag_005_IR_Lab_1(28_01_22).ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

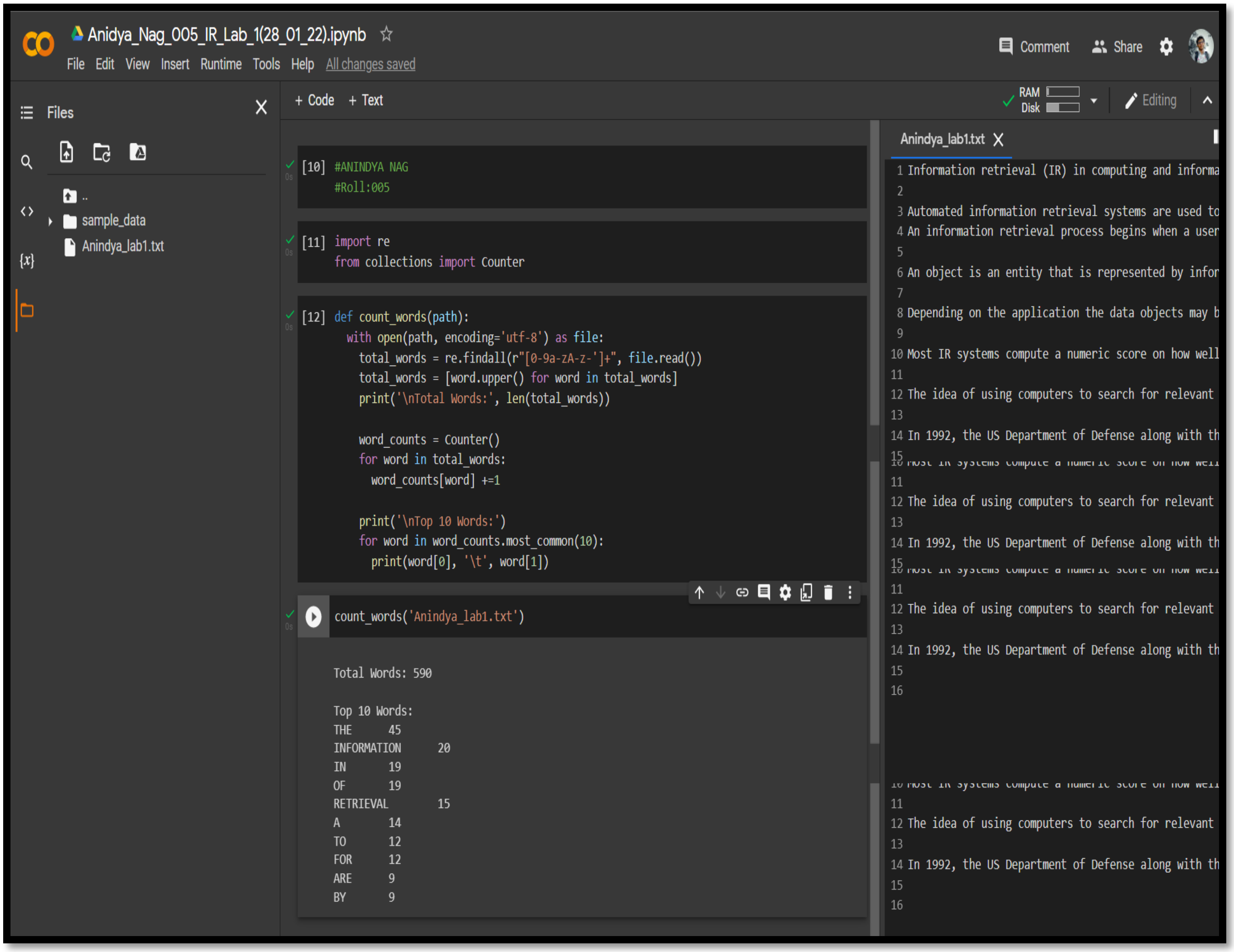
```
[10] #ANINDYA NAG
     #Roll:005
```

```
[11] import re
     from collections import Counter
```

```
[12] def count_words(path):
       with open(path, encoding='utf-8') as file:
         total_words = re.findall(r"[0-9a-zA-z-']+", file.read())
         total_words = [word.upper() for word in total_words]
         print('\nTotal Words:', len(total_words))

         word_counts = Counter()
         for word in total_words:
           word_counts[word] +=1

         print('\nTop 10 Words:')
         for word in word_counts.most_common(10):
           print(word[0], '\t', word[1])
```

```
count_words('Anindya_lab1.txt')
```

```
Total Words: 590

Top 10 Words:
THE         45
INFORMATION         20
IN          19
OF          19
RETRIEVAL           15
A           14
TO          12
FOR         12
ARE         9
BY          9
```


Anidya_Nag_005_IR_Lab_1(28_01_22).ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Files
..
sample_data
Anindya_lab1.txt
Anindya_lab1.txt

```
[10] #ANINDYA NAG
     #Roll:005
```

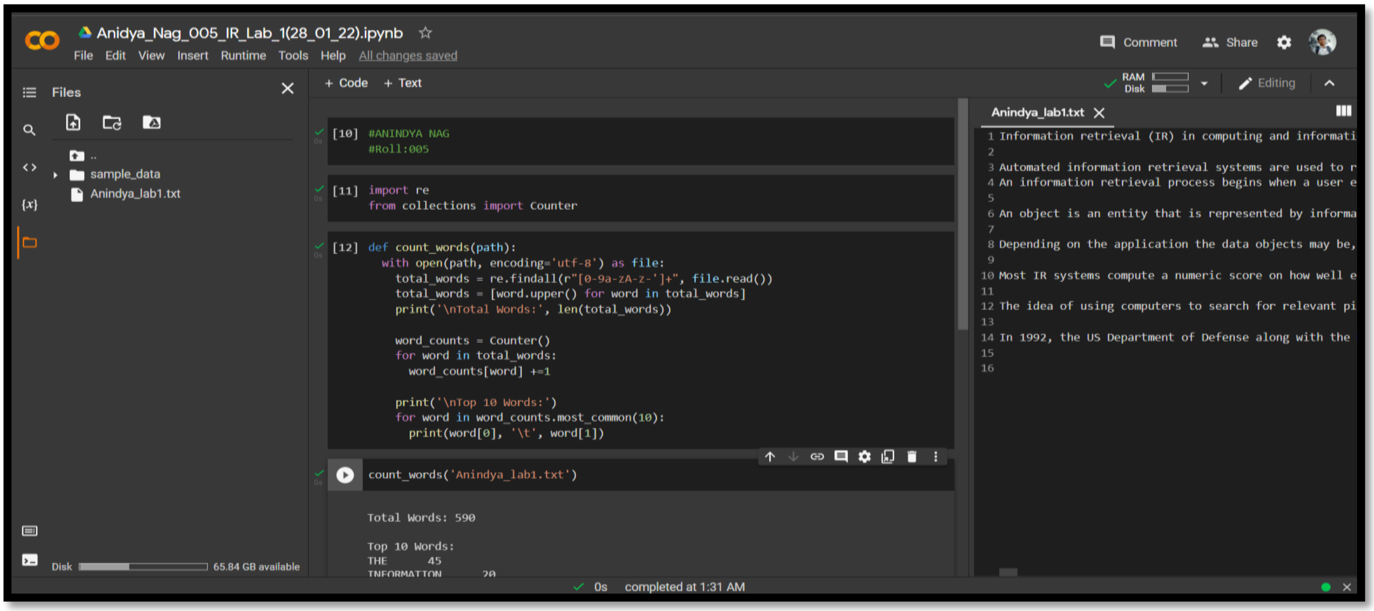
```
[11] import re
     from collections import Counter
```

```
[12] def count_words(path):
       with open(path, encoding='utf-8') as file:
         total_words = re.findall(r"[0-9a-zA-z-']+", file.read())
         total_words = [word.upper() for word in total_words]
         print('\nTotal Words:', len(total_words))

         word_counts = Counter()
         for word in total_words:
           word_counts[word] +=1

         print('\nTop 10 Words:')
         for word in word_counts.most_common(10):
           print(word[0], '\t', word[1])
```

```
count_words('Anindya_lab1.txt')
```

```
Total Words: 590

Top 10 Words:
THE      45
INFORMATION      20
IN       19
OF       19
RETRIEVAL        15
A        14
TO       12
FOR      12
ARE      9
BY       9
```

```
Information retrieval (IR) in computing and informa

Automated information retrieval systems are used to
An information retrieval process begins when a user

An object is an entity that is represented by infor

Depending on the application the data objects may b

Most IR systems compute a numeric score on how well

The idea of using computers to search for relevant

In 1992, the US Department of Defense along with th


```



```
[10] #ANINDYA NAG
     #Roll:005
```

```
[11] import re
     from collections import Counter
```

```
[12] def count_words(path):
       with open(path, encoding='utf-8') as file:
         total_words = re.findall(r"[0-9a-zA-z-']+", file.read())
         total_words = [word.upper() for word in total_words]
         print('\nTotal Words:', len(total_words))

         word_counts = Counter()
         for word in total_words:
           word_counts[word] +=1

         print('\nTop 10 Words:')
         for word in word_counts.most_common(10):
           print(word[0], '\t', word[1])
```

```
count_words('Anindya_lab1.txt')
```

```
Total Words: 590

Top 10 Words:
THE      45
INFORMATION      20
```

```
Information retrieval (IR) in computing and informati

Automated information retrieval systems are used to r
An information retrieval process begins when a user e

An object is an entity that is represented by informa

Depending on the application the data objects may be,

Most IR systems compute a numeric score on how well e

The idea of using computers to search for relevant pi

In 1992, the US Department of Defense along with the


```