**Experiment 2**

**Develop a python script to parse the pdf files using pdfminer.**

WHAT IS A PDF?

Portable document format is a file format developed by Adobe, that presents documents in a manner independent of application software, hardware, and operating systems.

WHAT IS PARSING?

Parsing in python means, the process of analyzing a string of characters like text, code or data and converting it into a structured format that a program can understand and manipulate.

WHAT IS PDFMINER?

It is an open source library and tool designed for extracting information from pdf documents, Pdfminer primary focus is on getting and analyzing text data within pdf’s.

1. **Extract text** from a PDF file.
2. **Tokenize** the text into words.
3. **Count** the frequency of each word.
4. Find words with:

**Length > 5 characters**.

**Frequency > 20 times** in the document.

1. **Plot the frequency distribution** of such words.

#code

**from nltk.tokenize import RegexpTokenizer**

**from pdfminer.high\_level import extract\_text**

**from nltk.probability import FreqDist**

**# Extract the text from PDF file**

**text = extract\_text(r"C:\Users\91984\Downloads\ISR Unit 1.pdf")**

**print("Extracted Text Sample:\n", text[:50]) # Print first 500 characters of the text**

**# Create an instance of tokenizer using NLTK RegexpTokenizer**

**tokenizer = RegexpTokenizer('\w+')**

**# Tokenize the text read from PDF**

**tokens = tokenizer.tokenize(text)**

**print("\nTotal Tokens Extracted:", len(tokens))**

**print("Sample Tokens:", tokens[:20]) # Print first 20 tokens**

**# Find Frequency Distribution**

**freqdist = FreqDist(tokens)**

**print("\nMost Common 10 Words:", freqdist.most\_common(10))**

**# Find words whose length is greater than 5 and frequency greater than 20**

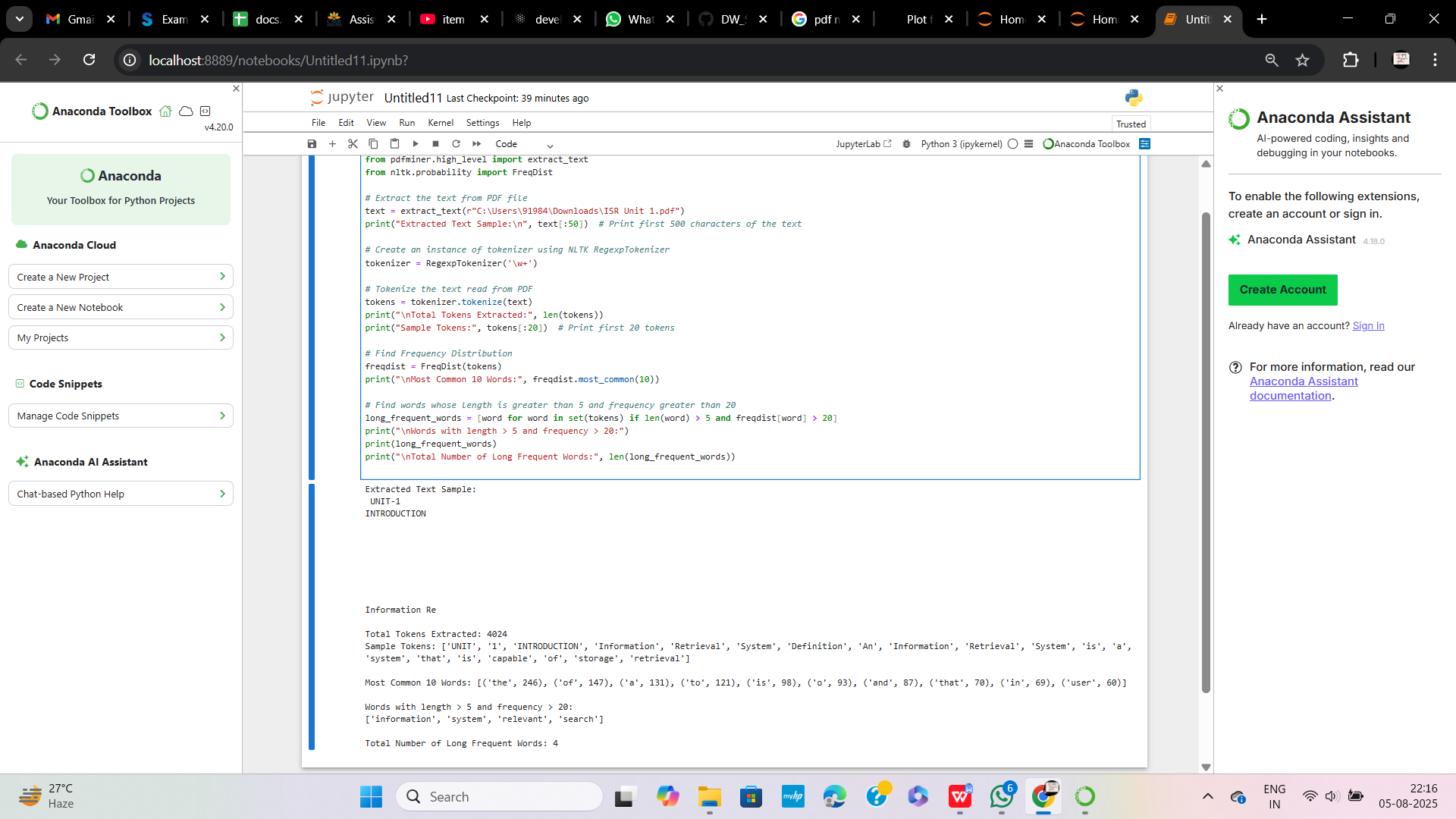
**long\_frequent\_words = [word for word in set(tokens) if len(word) > 5 and freqdist[word] > 20]**

**print("\nWords with length > 5 and frequency > 20:")**

**print(long\_frequent\_words)**

**print("\nTotal Number of Long Frequent Words:", len(long\_frequent\_words))**

OUTPUT



#code

**from nltk.tokenize import RegexpTokenizer**

**from nltk.probability import FreqDist**

**import matplotlib.pyplot as plt**

**from pdfminer.high\_level import extract\_text**

**# Extract text**

**text = extract\_text(r"C:\Users\91984\Downloads\CSE\_R22 syllabus book.pdf")**

**# Tokenize words**

**tokenizer = RegexpTokenizer(r'\w+')**

**tokens = tokenizer.tokenize(text)**

**# Compute Frequency Distribution**

**freqdist = FreqDist(tokens)**

**# Plot Top 30 Most Common Words (No filters)**

**plt.figure(figsize=(12,6))**

**freqdist.plot(30, cumulative=False)**

**plt.show()**

