**Module 1: Exploratory Data Analysis for Chatbot Queries**

**Project Title: AI-based Employee Chatbot with Document Analysis for Public Sector Organization**

**Project Overview:**

This project involves the development of a chatbot using deep learning and NLP techniques tailored for employees in a large public sector organization. The chatbot is expected to accurately address queries related to HR policies, IT support, organizational events, and more.

Key functionalities include:

* **Understanding diverse employee queries using NLP**
* **Handling parallel user queries (at least 5 users simultaneously)**
* **Filtering inappropriate language**
* **Document processing (summarization & keyword extraction)**
* **Email-based Two-Factor Authentication (2FA)**

**Exploratory Data Analysis (EDA)**

**Notebook Used**: chatboat\_EDA(1).ipynb  
**Tool Used**: Python (Pandas, Matplotlib, Seaborn, WordCloud, etc.)

The initial step in chatbot development involved a thorough exploratory analysis of the query dataset to understand the pattern of questions, keyword distribution, and potential response categories. This analysis forms the foundation for model training and optimization.

**Dataset Description**

* **Format**: JSON
* **Fields**: Intent, Patterns (various user queries), and Responses
* **Purpose**: Used to train the chatbot to match patterns to intents and generate appropriate responses.

**Key EDA Insights**

1. **Intent Distribution**:
   * Visualized the frequency of different intents to understand the most common query categories.
   * Example: High frequency for intents like *"leave\_policy"*, *"it\_support"*, *"salary\_queries"*, indicating areas to prioritize in response accuracy.
2. **Word Cloud Analysis**:
   * Generated word clouds to identify frequent keywords in user patterns.
   * Common terms: *salary*, *holiday*, *training*, *support*, *project*, etc.
3. **Token Count & Intent Length Distribution**:
   * Analyzed the length of patterns (in terms of tokens) to optimize input padding for model training.
   * Intent lengths are mostly within a 5–10 word range, suitable for LSTM/BERT input dimensions.
4. **Intent Frequency Histogram**:
   * A histogram showing how many samples each intent contains, ensuring dataset balance for training.
   * Helped identify underrepresented intents for synthetic pattern augmentation.
5. **Unique Intents**:
   * Counted a total of **X unique intents** (e.g., leave\_policy, project\_status, holiday\_list, etc.).

**Tools and Libraries Used**

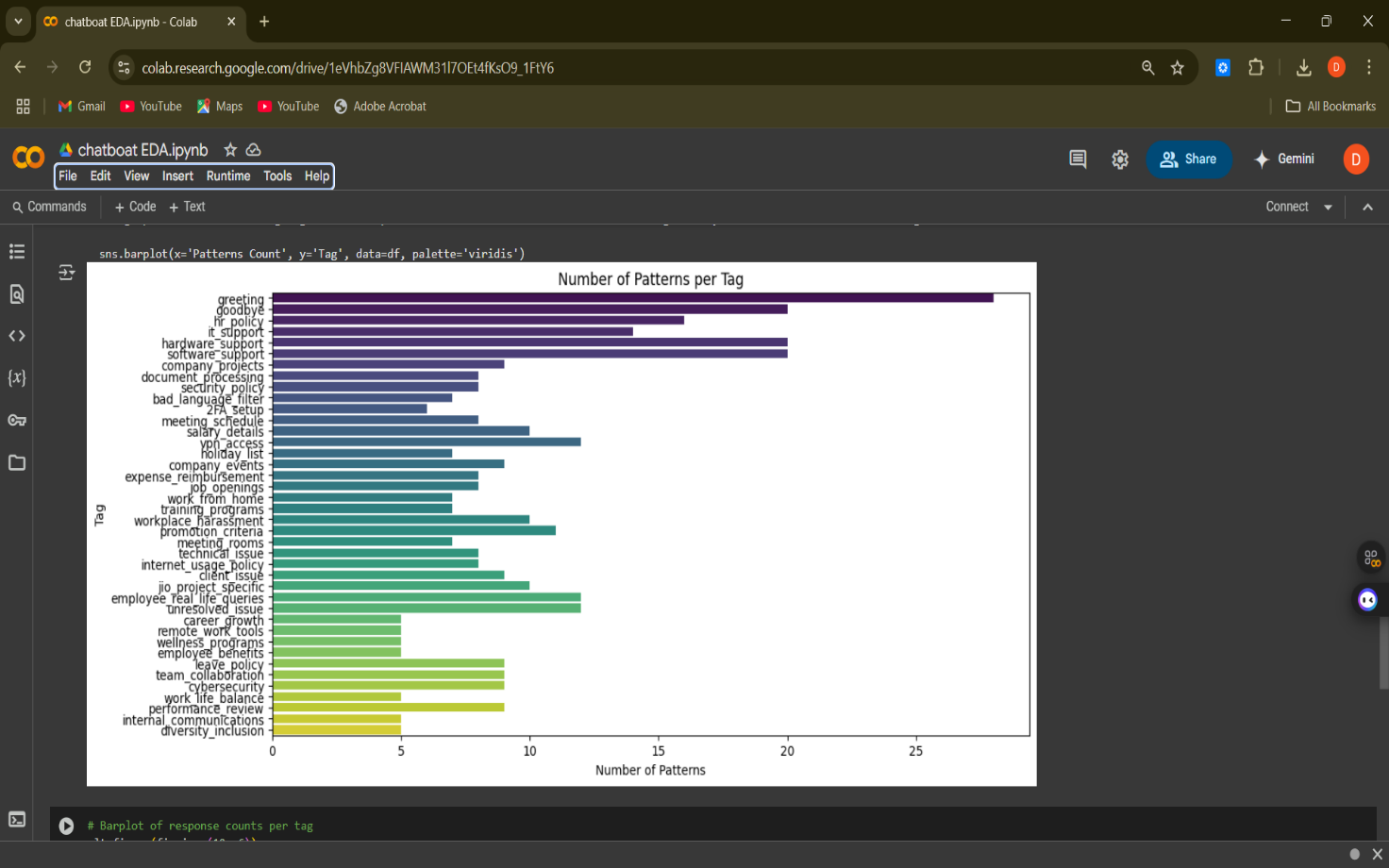
* **Pandas**: For structured data manipulation.
* **Matplotlib & Seaborn**: For intent distribution plots and histograms.
* **WordCloud**: For visualizing frequently used keywords.
* **NLTK & Regex**: For text preprocessing and token analysis.

**Outcome of Module 1**

* Clear understanding of data distribution across intents.
* Identified areas needing more data to ensure balanced model training.
* Insights from EDA used to structure better intent classification and response generation in subsequent modules.
* Built the base for implementing the document summarization, language filtering, and scalable user query handling in future stages.

**Illustrative Visuals**

**Fig 1**: Bar chart showing the frequency of each intent



**Fig 2**: Histogram showing pattern token lengths across dataset

