🧩 Project Title

"Discovering Emerging Topics in Drug Discovery Research using BERTopic"

🎯 **Project Objective**

To automatically identify and analyze **emerging research topics and trends** in the field of **drug discovery** by applying **BERTopic**, a transformer-based topic modelling algorithm.

The project aims to:

1. Extract latent topics from scientific publication abstracts related to drug discovery.
2. Optimize BERTopic hyperparameters to achieve coherent and interpretable topics.
3. Visualize how research focus areas (e.g., AI-driven molecule design, protein folding, mRNA therapeutics) have evolved over time.

💡 Business & Scientific Relevance

- **Biotech impact:** Pharmaceutical R&D teams and biotech analysts need to monitor fast-evolving areas like AI-driven drug design or molecular modelling to identify **research frontiers** and **innovation opportunities**.
- **Data science skill relevance:** Demonstrates real-world **NLP**, **transformer embeddings**, **topic modelling**, and **trend analysis** — all highly valued by biotech, fintech, and AI companies (e.g., Microsoft Research, AstraZeneca AI Labs, NVIDIA BioNeMo).

🔍 Data Source Plan

You'll need **independently collected data**, so you can easily source abstracts or titles from:

1. **PubMed** – Use the NCBI E-utilities API to download abstracts with keywords like *"drug discovery"*, *"AI drug design"*, *"molecular docking"*, *"protein folding"*, etc.

2. **arXiv** – Filter for "bioinformatics" or "computational biology" categories.

3. (Optional) **Kaggle datasets:** e.g., *"PubMed Drug Discovery Abstracts"* or *"BioNLP Publications Dataset."*

💾 You can easily preprocess and store data as .csv with columns:
[title, abstract, publication_date, authors, journal].

⚙️ **Methodology Overview (Jupyter Notebook Workflow)**

1️⃣ Data Collection & Cleaning

- Scrape or download ~2,000–5,000 abstracts.
- Clean text: remove duplicates, special characters, boilerplate text.
- Optional: filter by publication year for dynamic analysis.

2️⃣ Text Preprocessing

- Lowercasing, lemmatization.
- Remove stopwords and overly frequent scientific tokens.
- (Optional) Use domain-specific tokenizer like SciSpaCy.

3️⃣ Embedding Generation

- Compare embedding models (for hyperparameter tuning):
  - all-MiniLM-L6-v2 (lightweight)
  - BioBERT or SciBERT (domain-specific)

4️⃣ Topic Modelling with BERTopic

- Use **UMAP** for dimensionality reduction.
- Use **HDBSCAN** for clustering.
- Tune hyperparameters such as:
  - n_neighbors (5–50)
  - min_cluster_size (10–50)
  - top_n_words (5–10)

- Evaluate with **topic coherence score** and **topic diversity**.

5️⃣ Evaluation & Optimization
- Quantitatively assess coherence to select optimal parameters.
- Qualitatively review topic interpretability.
- Visualize the most frequent words per topic.

6️⃣ Temporal / Trend Analysis
- Group papers by year → observe topic shifts.
- Visualize topic proportions over time using BERTopic's topics_over_time() or a custom time series plot.

7️⃣ Result Presentation
- Word clouds or bar charts for top topics.
- Dynamic topic trends over time.
- Table of top keywords and representative documents per topic.

📈 Expected Results
- ~10–15 interpretable research topics (e.g., *AI-based molecule screening*, *mRNA vaccine development*, *quantum simulation for drug design*).
- Visualization of how each topic has grown or declined over the past 10 years.
- Insights into emerging "hot topics" in drug discovery.

🧮 Hyperparameter Optimization Plan

Parameter	Range to Test	Evaluation Metric
n_neighbors	5, 15, 30, 50	Topic coherence
min_cluster_size	10, 25, 50	Topic coherence + topic diversity
Embedding model	MiniLM vs SciBERT vs BioBERT	Topic interpretability
top_n_words	5–10	Human readability

🧠 What You'll Learn / Showcase
✅ Real-world text preprocessing for unstructured scientific data.
✅ Application of **transformer-based topic modelling (BERTopic)**.
✅ Hyperparameter tuning and model evaluation using coherence metrics.
✅ Visualization of topic evolution and trend detection.
✅ A business-scientific report format in **Jupyter Notebook**, matching your course's evaluation rubric.

🏁 Deliverables
- **Jupyter Notebook Report** with:
  - Data collection + preprocessing
  - Model training + optimization
  - Visualizations and analysis
- **Summary Slide / PDF** (optional for presentation):
  - Key findings and implications for R&D strategy