Group Assignment: Latent Topic Landscapes in Research Abstracts

Team Projects 1 and 2: Due on 4/16

Introduction

Modern scientific research spans thousands of papers and domains. But what if you could automatically uncover the **latent themes** across a collection of paper abstracts—without using any predefined labels?

In this group assignment, you'll apply **unsupervised generative models**—*Gaussian Mixture Models (GMMs)*, *Restricted Boltzmann Machines (RBMs)*, and *Autoencoders*—to learn compressed representations of research abstracts. You will explore how these models help uncover hidden structure in textual data, such as research topics, author style, or methodological similarities.

Note: This is assignment is drawn from your final project and will help you in further rethinking your work and revisiting some of the ideas that you may have already explored

Learning Outcomes

By completing this assignment, you will be able to:

- Preprocess and vectorize natural language text data for generative modeling
- Use Autoencoders, GMMs, and RBMs on non-image datasets
- Analyze and visualize the latent structure in scientific text
- Evaluate clustering and reconstruction quality in an unsupervised setting
- Collaborate to design, implement, and interpret hybrid ML pipelines

Dataset

You may choose one of the following:

- **Your own dataset and embeddings** from your term-long project

- **External datasets**, e.g., arXiv abstracts dataset, PubMed, or scraped conference proceedings

Each abstract should include:

- Title and abstract text
- Optional metadata: categories, keywords, authors (useful for qualitative comparison)

Preprocessing Requirements

Each abstract must be represented as a vector using **at least one** method:

- TF-IDF vectors
- Sentence or document embeddings (e.g., Sentence-BERT, Doc2Vec)

You may reuse any embeddings you generated for your project.

Tasks

Task 1: Dimensionality Reduction

Use either an **Autoencoder** or an **RBM** to reduce the dimensionality of your abstract vectors (e.g., from 768 or 5000+ to 10–50 latent dimensions).
Visualize the latent space (e.g., using PCA, t-SNE, UMAP).

Task 2: Latent Clustering

Apply **GMMs** or **RBMs (as feature extractors)** on the representations from Task 1.

- Cluster the abstracts into k latent groups.
- Sample representative abstracts from each cluster.
- Interpret each cluster: what kinds of research are grouped together?

Task 3: Reconstruction (If Using Autoencoder)

If you used an Autoencoder:

- Reconstruct the abstract vectors and evaluate reconstruction loss.
- Identify poorly reconstructed samples and common traits.

Task 4: Generation / Interpolation (Optional Challenge)

If using a **VAE**:

- Sample new latent vectors and decode them.
- Interpolate between two abstracts and analyze how content evolves.

Deliverables

Submit the following:

1. **Report** (max 5 pages):
    - Dataset and embedding method
    - Description of models used
    - Latent space visualizations
    - Cluster interpretations
    - Optional reconstruction/generation insights
2. **Code notebooks** (Google Colab or Jupyter)
3. **A 10-15 minute presentation showing workflow and results**

Questions

- How did your model combination help uncover latent structure?
- What surprised you about the clusters or reconstructions?
- How could this be extended to new research areas or languages?