Homework 4

May 1, 2025

Objective

The goal of this assignment is to apply unsupervised learning techniques to a real-world text dataset to discover latent structure among documents. You will:

- Convert documents into numerical feature representations
- Apply dimensionality reduction for visualization
- Use clustering algorithms to group similar documents
- Analyze and interpret the results

Dataset

Use the `20 Newsgroups` dataset available via `sklearn.datasets.fetch_20newsgroups`. This dataset contains approximately 18,000 newsgroup posts across 20 different topics.

```
from sklearn.datasets import fetch_20newsgroups
data = fetch_20newsgroups(subset='all', remove=('headers', 'footers', 'quotes'))
documents = data.data
```

Tasks

1. Preprocess the Data

- Convert the raw text documents into TF-IDF features using `TfidfVectorizer`.
- Limit the vocabulary size (e.g., `max_features=1000` or 2000) to reduce dimensionality.
- Remove stop words and apply other basic preprocessing steps.

2. Dimensionality Reduction

- Apply **PCA** and **t-SNE** (separately) to reduce the data to 2D.
- Visualize the documents in a 2D scatter plot.
- Optionally color-code the original 20 labels (for reference only).

3. Clustering

- Apply at least **two clustering algorithms**, such as:
  - K-Means
  - DBSCAN
  - Gaussian Mixture Model
- Visualize the cluster assignments in the 2D PCA or t-SNE space.
- Compare them with the reference class labels using clustering metrics such as B-Cubed precision and recall.

4. Analyze the Clusters

- For each cluster:
  - Identify the **top 10 words** with the highest average TF-IDF scores.
  - Try to **name or describe** the cluster based on its content.

5. Reflective Questions

Include short answers to the following questions in your notebook:

1. Which dimensionality reduction technique produced more meaningful visual separation? Why?
2. Which clustering algorithm matched the original topics better?
3. What challenges did you face when clustering textual data?
4. Could you identify any meaningful topics based on the clusters alone?

Deliverables

Submit a single **Jupyter Notebook** that includes:

- All code
- Visualizations
- Explanatory markdown cells
- Answers to the reflective questions

Optional Extensions (Bonus)

- Try using **word embeddings** (e.g., via `spaCy` or `gensim`) instead of TF-IDF.
- Perform **hierarchical clustering** and plot a dendrogram.
- Try clustering short social media posts (e.g., tweets) if you want a smaller or different dataset.