**Week 4 Lab Exercise Sheet: Advanced Word Clouds, Topic Modeling & Dimensionality Reduction**

**Objective:**

This lab will guide you through **advanced word cloud visualisation** techniques before preparing you for **two assignments**:

1. **Topic Modeling with Wikipedia Articles**
2. **PCA for Digit Classification**

By the end of this lab, you should have a **deeper understanding of NLP techniques** and **dimensionality reduction**, enabling you to work on the assignments effectively.

**🔹 Task 1: Generate a Simple Word Cloud**

Before diving into **advanced word clouds**, ensure you can generate a **basic word cloud** from a text corpus.

✅ **Steps:**

* Load a **text dataset** (e.g., a Wikipedia article or news text).
* Remove **stopwords, punctuation, and unnecessary symbols**.
* Convert the text into **a word cloud image**.

**🔹 Task 2: Advanced Word Cloud Visualisation**

Now, extend your basic word cloud by adding **complex features**:

✅ **Steps:**

1. **Shape-Based Word Clouds**
   * Use **a custom mask** to generate a word cloud in a **specific shape** (e.g., a map of Vietnam, a book, or a speech bubble).
2. **Frequency vs. TF-IDF Word Clouds**
   * Compare a **regular frequency-based word cloud** vs. a **TF-IDF weighted word cloud**.
   * Explain how TF-IDF gives more importance to unique words instead of commonly used ones.
3. **Multi-Coloured & Themed Word Clouds**
   * Use **colour gradients** to differentiate **positive vs. negative** words.
   * Create **different word clouds for different topics** in the same dataset.
4. **Bigrams & Phrase-Based Word Clouds**
   * Extract **bigrams (two-word phrases)** to include words like **"Vietnam War"** instead of separate words **"Vietnam" and "War"**.
5. **Compare Word Clouds Before & After Preprocessing**
   * Generate one word cloud **before** text cleaning (including stopwords, symbols).
   * Generate another word cloud **after** applying **lemmatisation & stopword removal**.
   * Compare the **differences** and discuss why preprocessing improves visualisation.

**🔹 Task 3: Understanding Topic Modeling & Dimensionality Reduction**

Before working on the assignments, ensure you understand **these key concepts**:

**📌 Topic Modeling (For Assignment 1)**

* Topic modeling is used to **discover hidden themes** in large text data.
* **Latent Dirichlet Allocation (LDA)** is a popular method that groups words into topics based on co-occurrence patterns.
* **pyLDAvis** helps visualise the topic distribution and **important words per topic**.

✅ **Preparation Task:**

* Research **how LDA works** and how topics are extracted from documents.
* Understand how the **relevance slider (λ) in pyLDAvis** helps refine word selection for each topic.

**📌 Dimensionality Reduction (For Assignment 2)**

* High-dimensional data (e.g., images or text embeddings) is difficult to visualise.
* **Principal Component Analysis (PCA)** helps reduce dimensions while keeping important information.
* In the **Digits dataset**, PCA can transform **64-dimensional images** into a **2D or 3D representation** for easier classification.

✅ **Preparation Task:**

* Research how **PCA works** and why it is useful for high-dimensional data.
* Understand how **explained variance** helps determine **how much information is retained after reducing dimensions**.
* Read about **common misclassifications** (e.g., **why people confuse 1 and 7 in handwriting recognition**).

**🔹 Assignment Overview (check the lab4\_assignments.docx file)**

This lab consists of **two assignments**:

1️ **Topic Modeling with Wikipedia Articles**

* You will extract Wikipedia text and **apply LDA** to find dominant topics.
* Use **pyLDAvis** to visualise topics and analyse word distributions.

2️ **PCA for Digit Classification**

* You will **apply PCA** to the **Digits dataset** and reduce 64D data into **2D or 3D**.
* Identify which **digits are visually similar** and **why certain numbers are often confused**.

**🚀 Next Steps**

* **Complete the guided exercises** on advanced word clouds.
* **Ensure you understand topic modeling & dimensionality reduction** before starting the assignments.
* **Start working on both assignments** and prepare a report discussing your results.

Good luck! 🚀