📚 Computer Laboratory II (Information Retrieval & Web Mining) Viva Preparation Guide

This guide summarizes the core concepts, practical implementations, and high-probability oral questions for the assignments in the Information Retrieval (IR) and Web Mining Lab (Subject Code: 417526).

📝 Assignment 1: Text Pre-processing (Stop Word Removal, Stemming)

Questions & Answers

- **What is the goal of Text Pre-processing in NLP?**
  - To transform raw, unstructured, and noisy text into a **clean and consistent format** that can be accurately analyzed and fed into a model for learning.
- **What is Tokenization?**
  - The process of breaking up a text (like a sentence) into fundamental units called **tokens** (usually individual words or punctuation marks).
- **What are Stop Words, and why are they removed?**
  - They are **frequently occurring words** (like 'is', 'the', 'and') that are often redundant for specific NLP applications (like document classification or spam filtering) and are removed to **reduce dimensionality** and improve signal-to-noise ratio.
- **Explain Stemming.**
  - It's the process of converting all words to their **base or root form (stem)**, often by crudely chopping off suffixes (e.g., "learning" becomes "learn").
- **How does Lemmatization differ from Stemming?**
  - **Lemmatization** is a more advanced technique that converts words to their true **dictionary form (lemma)**, using contextual and linguistic knowledge (Parts of Speech tagging), making it more accurate than simple stemming.

🗄️ Assignment 2: Document Retrieval using Inverted Files

Questions & Answers

- **What is an Inverted Index (Inverted File)?**
  - It is a core **index data structure** in Information Retrieval that stores a mapping from **content (words or tokens)** to their **locations (documents or lines)** within a set of documents.
- **Why are Inverted Files essential for searching?**
  - They allow for extremely **fast lookup** of documents that contain a specific query term, as you go directly from the word to the document list, rather than scanning every document.

- **What is the core structure of an Inverted Index?**
  - It functions like a HashMap or Dictionary, where the **key** is the term/word and the **value** is a list of documents (or locations) where that term appears.
- **Before creating the index, what pre-processing steps are mandatory?**
  - **Lowercasing** (to maintain uniformity) and **Tokenization** (to isolate words), often followed by **Stop Word removal**.

Assignment 3: Constructing a Bayesian Network

Questions & Answers

- **What is a Bayesian Network?**
  - It is a **Directed Acyclic Graph (DAG)** where each **node** represents a unique random variable, and each **edge** represents a **conditional dependency** between those variables.
- **What are the two major components of a Bayesian Network?**
  - 1. The **Directed Acyclic Graph (DAG)**, showing the dependencies.
  - 2. A set of **Conditional Probability Distributions (CPDs)**, defined for each node based on its parent nodes.
- **In the context of the Heart Disease diagnosis example, what is the goal?**
  - The goal is to calculate the **posterior conditional probability distribution** of an unobserved **Cause** (like Heart Disease presence) given the observed **Evidence** (like cp, $trestbps$, $chol$).
- **What kind of problems are Bayesian Networks best suited for?**
  - Problems involving **uncertainty** and **causal relationships**, such as **diagnosis**, prediction, and reasoning under incomplete information.

Assignment 4: E-mail Spam Filtering using Text Classification

Questions & Answers

- **Is Spam Filtering a supervised or unsupervised learning problem?**
  - It is a **supervised learning** problem because the dataset (corpus) requires emails to be **labeled** as either 'spam' or 'ham' (non-spam).
- **Why is pre-processing crucial for spam filtering?**
  - Text must be cleaned via **Tokenization, Stemming/Lemmatization, and Stop Word removal** to reduce complexity and ensure the model focuses on the content words that distinguish spam from legitimate mail.
- **Why would the k-NN algorithm be used for this problem?**
  - It classifies a new email based on the class (spam or ham) of its **K-Nearest Neighbors** in the feature space, assuming that similar emails are likely to have the same label.

- **What is a major evaluation metric besides Accuracy for a spam filter?**
  - **Precision ($TP/(TP+FP)$)** is critical. A high precision score means the filter minimizes **False Positives (FP)** (i.e., legitimate emails wrongly marked as spam), which is highly important for user satisfaction.
- **What is the Porter Stemmer's Algorithm?**
  - One of the most famous algorithms used for stemming, which uses a set of rules to remove common suffixes from English words.

📊 Assignment 5: Agglomerative Hierarchical Clustering

Questions & Answers

- **What is Agglomerative Hierarchical Clustering (AHC)?**
  - It is a **bottom-up** clustering approach where each data point starts as its own cluster, and then pairs of clusters are **merged** iteratively until all points belong to a single cluster or the stopping criterion is met.
- **What is a Dendrogram, and what is its role?**
  - A tree-like diagram that records the sequence of merging (or splitting) operations. It is used to visually **determine the optimal number of clusters** by identifying the largest vertical distance that does not intersect a horizontal line.
- **What is the Silhouette Score used for?**
  - It is a **mathematical technique** used to determine the optimal number of clusters. It measures how similar an object is to its own cluster compared to other clusters, with a higher score indicating better-defined clusters.
- **What must be done to the data before clustering (after cleaning)?**
  - The data must be **Scaled** (StandardScaler) and potentially **Normalized** to ensure all features contribute equally to the distance calculation, and **Dimensionality Reduction (PCA)** may be applied to simplify visualization.

Assignment 6: Page Rank Algorithm

Questions & Answers

- **What is the primary purpose of the Page Rank Algorithm?**
  - To determine the **relative importance or relevance** of a webpage within a set of hyperlinked documents (the web graph).
- **How is the web modeled for Page Rank?**
  - As a **directed graph** where nodes represent web pages and edges represent hyperlinks between them.

- **Explain the concept of a "Random Walk" in Page Rank.**
  - It simulates a user randomly jumping from one page to another. The **stationary distribution** of this random walk (which is π) determines the Page Rank score; the higher the probability of landing on a page, the more important it is.
- **What are "Spider Traps" and "Dead Ends"?**
  - **Spider Traps** are cycles of pages where a walker gets stuck (e.g., A links to B, B links to A). **Dead Ends** are pages with no outgoing links.
- **How does the Damping Factor (β) solve the issues of traps and dead ends?**
  - It introduces the concept of **teleportation**. With probability β (typically 0.85), the walker follows a link, and with probability $(1 - \beta)$, the walker **jumps randomly** to any page in the graph, ensuring the algorithm converges.