Project Synopsis

ON

TOPWORDS

INDEX & TABLES

1. DESCRIPTION OF PROJECT

INTRODUCTION:

Word frequency is word counting technique in which a list of words and their frequency is generated, where the frequency is the occurrences in a given composition. It is used commonly in computational linguistics. Word frequency has many applications in diverse fields. Within pedagogy, it allows teaching to cover high-frequency vocabulary before the low-frequency ones, enabling the development of better class curriculum. But let's talk about a field that any developer might find compelling, which are analytics, for a publisher that requires analytics to suggest vocabulary improvements, it detects trends in word usage, and determine payment for their writers. Payment is determined by many publishers based on the number and complexity of the wording.

Topwords project is based on this technique which requires a PDF file and the number of desired words as input and gives the number of words with their frequencies in the document as output.

This is also a quick way to tell what some document is focusing on, what its main topic is. It lists the unique words mentioned in the document, and then check how many times each word has been mentioned (frequency). This way would give you an indication of what the document is mainly about. But that wouldn't work easily manually, so we need some automated process. Yes, an automated process will make this much easier.

METHOD:

Regular expression is a sequence of characters that define a search pattern, mainly for use in pattern matching with strings, or string matching, i.e. "find and replace"-like operations. The concept arose in the 1950s, when the American mathematician Stephen Kleene formalized the description of a regular language, and came into common use with the Unix text processing utilities ed, an editor, and grep, a filter.

2. MODULES

1. **Login :**

   In this module, Admin enters the User id and password is checked and only valid user id and password will get entry into search zone. This is a security feature to avoid entry of unauthorized users.

2. **Browse:**

   Through this Admin can browse to a specific directory and select the desired file. This module simplifies job of Admin by manually typing directory path.

3. **Search:**

   Through this Admin can find the desired number of words and their frequencies in the selected PDF file

4. **Reset:**

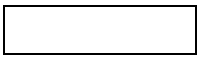
   This modules allows Admin to reset all the input entries and the output textbox to blank so they can select some other file or set the required number of words whose frequencies are to be known.

3. DATA FLOW DIAGRAM

DFD

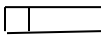
The Data flow Diagram shows the flow of data. It is generally made of symbols given below:

(1) A **square** shows the Entity: -



(2) A **Circle** shows the Process: -

(3) An **open Ended Rectangle** shows the data store: --



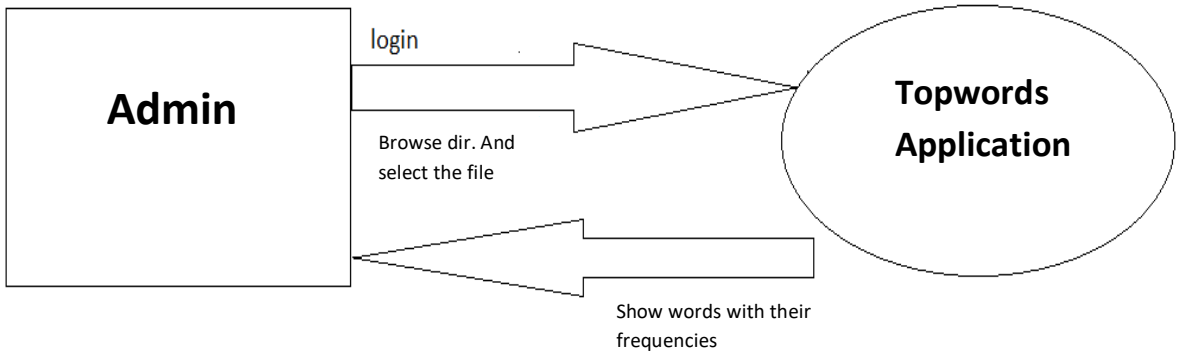
(4) An **arrow** shows the data flow:-

The DFD can be up to several levels. The 0 level DFD states the flow of data in the system as seen from the outward in each module.

The first level DFD show more detail, about the single process of the 0 level DFD

The second level DFD can show even more details and so on.

Context Level DFD



4. E-R Diagram

Definition:

An entity-relationship (ER) diagram is a specialized graphic that illustrates the interrelationships between entities in a database. ER diagrams often use symbols to represent three different types of information. Boxes are commonly used to represent entities. Diamonds are normally used to represent relationships and ovals are used to represent attributes.

Entity Relationship (ER) diagram:

This diagramming technique is used to visually present a database schema or data model and was original proposed by Chen in the 1970s. There are many different data modeling notations; some are very similar to UML class diagrams (with the exception of operations). However, the notation the used here is slightly different, as proposed by Elmasri, et al.
The database schema for this system is shown in figure. The table object has been left out of the diagram because the table management feature set had been dropped from the requirements before this stage of the design process.

Some important database design decisions are as follows:
_To store the total price of an order with the order rather than calculating it on the fly when looking at past orders. This is because the price of menu items could change at any time, so the total price at the time of ordering must be stored so that the total price is not incorrectly calculated in future.

_ Similar to the previous point, the order receipt is stored as a hard-copy and not regenerated when reviewing past orders because things such as the restaurant name or VAT percentage are subject to change. Receipts stored need to be exactly the same as the customer copy in case of dispute.

Note: In this project we have not used any database table since it is related to analysis and we read data from PDF files which are unstructured so there is no need of ERD.

5. Language/Libraries/Tools

Front End :

Python tkinter

Back End :

Python and Data Science

Libraries :

- ✓ tkinter
- ✓ docxpy
- ✓ os
- ✓ pyinstaller
- ✓ PyPDF2
- ✓ re(regular expression)
- ✓ PDFminer
- ✓ importlib

Other S/W :

- ✓ Python3.x
- ✓ IDLE
- ✓ Anaconda(jupyter lab)