INTELLIGENT DOCUMENT FINDER

Introduction

How easy do you find it to remember the exact location of a presentation that you created last year? Not very easy, right? We deal with hundreds of documents daily and forget about them some time down the line. But what if we want that old presentation again for our your work, but unfortunately you do not remember the name or content of that document to retrieve it from the large storage of your computer. In such cases, we make use of a document finder, which can search for the document of our need based on a query input. This will not only help in faster access to the document, but will also help in grouping similar documents together and in analysing them.

Problem Statement

We are looking for an Intelligent Document Finder tool that can provide easy and intelligent searches among the document files. The required document type includes presentations, pdf, doc and txt files. The main idea behind this problem statement is combining human tagging with an automated semantic search for efficient document finding. The tool is supposed to have manual as well as auto tagging capabilities. Once the documents are tagged, the user will enter a few queries in the search page of the tool to look for the most relevant documents.

Explanation:

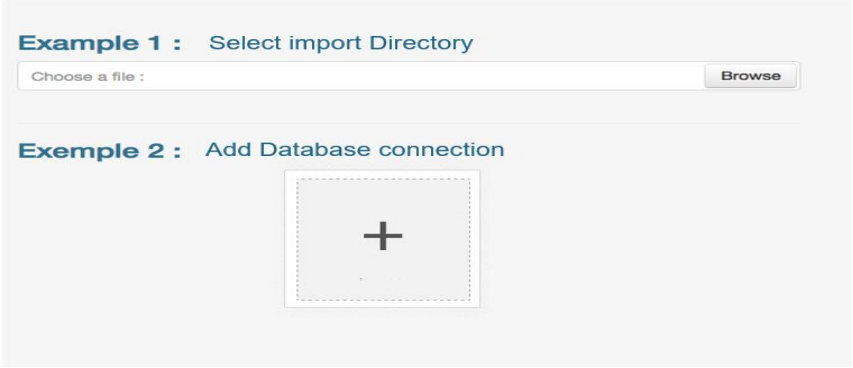
Teams are expected to come up with a **website application** with following features implemented:

- **Multiple document type support**(.pptx, .pdf, .txt, .docx)
- Support for data import from **SQL databases** and **storage directories**
- The tool must Provide **automatic tagging** based on the content present inside the particular document. Automatic tagging must be document specific and meaningful enough to help in query search.
- There must be an interface to provide **manual tags** for each document by *highlighting* the contents or entering it as *free text*.
- **Semantic search** across all the available documents based on the user's query. The result should contain a list of documents against the entered query with the most relevant document on the top. The tool should be intelligent enough to understand that semantically **White house** is similar to **Rashtrapati Bhavan.**
- **Abstractive summarization** of the available files to give the user an overview of the contents present in that file
- It should be an **OS independent web application** with support for **multiple browsers**. i.e. Chrome, Firefox and IE

Sample web application interface:

Note: Attached sample images are for demo only, actual design can be different from it.

1. Database Integration (Refer Image 1):
    - Locate data import path in the input page of the tool



2. Features of Search: (Refer Image 2)
    - Enter Query keywords as free text in the search page
    - Select search type (search by tags/document context search)
3. Result: (Refer Image 2)
    - A list showing details of the matching files to the entered query. The ranking should be done on the basis of relevance to the entered query with the top ones being the most relevant.
    - Abstractive summary of all the documents in the list.

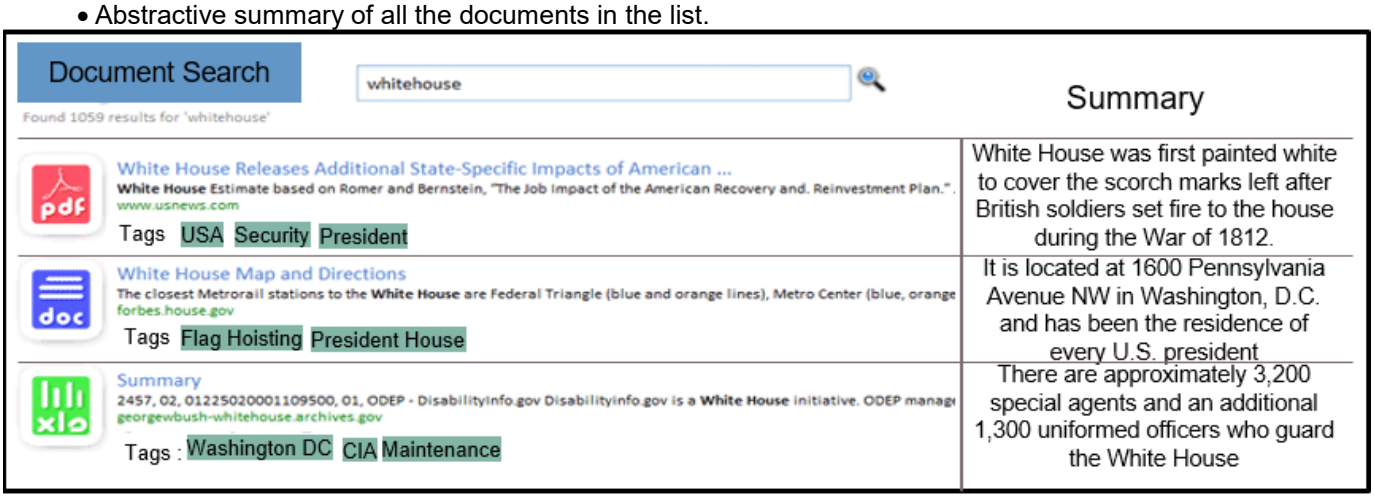
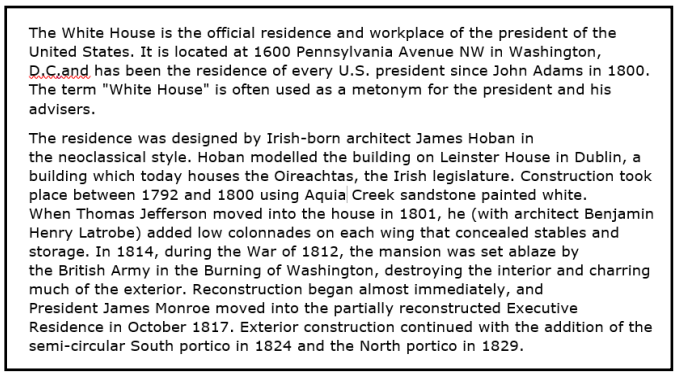


Image 2

Reference

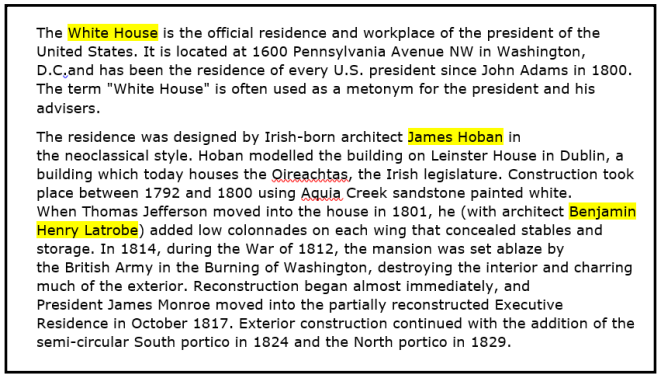
In the reference images below a pdf file is displayed and few of the words are highlighted by the user, which work as tags for that document. Once the document is tagged, a user can search by entering a query. User entered tags can also be used along with the tags generated by the model for better search results. The search algorithm should be working on semantic search and not only on keyword search.



The White House is the official residence and workplace of the president of the United States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C.and has been the residence of every U.S. president since John Adams in 1800. The term "White House" is often used as a metonym for the president and his advisers.

The residence was designed by Irish-born architect James Hoban in the neoclassical style. Hoban modelled the building on Leinster House in Dublin, a building which today houses the Oireachtas, the Irish legislature. Construction took place between 1792 and 1800 using Aquia Creek sandstone painted white. When Thomas Jefferson moved into the house in 1801, he (with architect Benjamin Henry Latrobe) added low colonnades on each wing that concealed stables and storage. In 1814, during the War of 1812, the mansion was set ablaze by the British Army in the Burning of Washington, destroying the interior and charring much of the exterior. Reconstruction began almost immediately, and President James Monroe moved into the partially reconstructed Executive Residence in October 1817. Exterior construction continued with the addition of the semi-circular South portico in 1824 and the North portico in 1829.

1. Open one of the document files using the tool



The White House is the official residence and workplace of the president of the United States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C.and has been the residence of every U.S. president since John Adams in 1800. The term "White House" is often used as a metonym for the president and his advisers.

The residence was designed by Irish-born architect James Hoban in the neoclassical style. Hoban modelled the building on Leinster House in Dublin, a building which today houses the Oireachtas, the Irish legislature. Construction took place between 1792 and 1800 using Aquia Creek sandstone painted white. When Thomas Jefferson moved into the house in 1801, he (with architect Benjamin Henry Latrobe) added low colonnades on each wing that concealed stables and storage. In 1814, during the War of 1812, the mansion was set ablaze by the British Army in the Burning of Washington, destroying the interior and charring much of the exterior. Reconstruction began almost immediately, and President James Monroe moved into the partially reconstructed Executive Residence in October 1817. Exterior construction continued with the addition of the semi-circular South portico in 1824 and the North portico in 1829.

2. Highlight the tags for the document or add it as free text

Deliverables

Your deliverable should be a working prototype solution for the problem chosen. Teams may either submit their work via their GitHub repository links or via uploading zipped folders, along with instructions for the creation of an execution environment and running the solution. The solutions will be executed and judged for their accuracy and experience. Besides, you need to submit a write-up covering:

- Final solution design
- Algorithm/Models actually applied
- Tech stack used
- Accuracy attained
- Any deviations

Dataset

Participants are free to use any open-source data that relates to the problem statement. Participants must be able to present the dataset used for training and testing during the evaluation process, if requested by the authorities. The dataset must be a folder containing all the documents and then the tool must be able to view all the documents for tagging by the user. Once annotated, the tool must be able to search for a relevant document based on the query input.

Evaluation Metrics

The judgement will be based on the following criteria:

- Usability and scope for large scale implementation
- Functionality
- Completeness of the idea/prototype

Note: Code & report to be shared by contestants. Only participants who have shared their code will be eligible for prizes