

Information Retrieval

Sessional - 2

A. Mustafi

August 18, 2023

1. Write a program to read the file *black-beauty.txt* from the *classics* corpus and extract tokens that represent a year range. A year range is usually indicated by the template `<dddd>-<dddd>`.
2. For the same text file how many URLs can you extract. Assume an URL has the basic template `http[s]://hostname.extension`.
3. Write a script to create a corpus containing all the Inaugural Presidential Addresses delivered by Presidents of the United States. The link to all such monologues can be found at <https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/inaugural-addresses>