

Online Submission Deadline: 23rd March 2021

Crawler Implementation, TF-IDF, Index Construction and Compression

[2.5 x 4]

- Upload your code and result as a single PDF file in VTOP [Mandatory] and MS Team Assignment [optional] on or before the deadline.
- No other form of submission will be acceptable.
- If you failed to upload in VTOP on or before the deadline, but successfully uploaded in MS Team Assignment, then 2 marks of penalty will be imposed on the secured marks.
- If you fail to upload your assignment in both VTOP and MS Team Assignment, then your assignment will not be evaluated and ZERO (0) mark will be awarded.
- File should contain
 - Question
 - Code
 - Result / Output screen

1. Write a python program to
 - a) show the implementation of a concurrent breadth-first crawler (No. of threads = 4, breadth = 5 and depth = 3).
 - b) Develop the crawler program to handle various challenges (such as Parsing, Stemming, Lemmatization, Link Extraction, Canonicalization, Spider Trap etc.) faced by crawler while implementing.
 - c) Based on the contents retrieved, prepare one inverted index file (with proper representation) using pandas.
2. Write a python program to show the implementation of Golomb Encoding-decoding technique.
 - a) Encode x=50, 74, with b=11 and b=16.
 - b) Decode the Golomb encoded sequence 1111111110010001101 with b = 10.
3. Write a python program to extract the contents (excluding any tags) from two websites

https://en.wikipedia.org/wiki/Artificial_intelligence

https://en.wikipedia.org/wiki/Machine_learning

Save the content in two separate files. Construct a trie based on the content retrieved in using HashMap / B-Tree / Dictionary. Write a program to show the implementation of **Predictive Typing** and **Auto-Correct** using the trie prepared.

Assessment - 2

4. Write a program to extract the contents (excluding any tags) from the following six websites

https://en.wikipedia.org/wiki/Web_mining

https://en.wikipedia.org/wiki/Data_mining

https://en.wikipedia.org/wiki/Artificial_intelligence

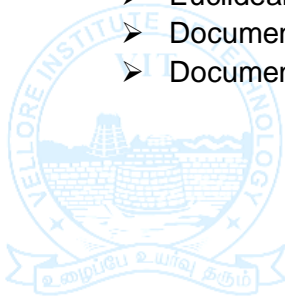
https://en.wikipedia.org/wiki/Machine_learning

https://en.wikipedia.org/wiki/Natural_language_processing

https://en.wikipedia.org/wiki/Text_mining

Save the content retrieved excluding stopwords in six separate files. Considering a vector space model, do the following operations according to the query "Mining of large text data" and represent the result in appropriate format using pandas.

- Bag-of-Words (Document set)
- TF (Document set)
- IDF (Document set)
- TF-IDF (Document set)
- TF-IDF (Query)
- Normalized (Query)
- Normalized - TF-IDF (Document set)
- Cosine Similarity
- Euclidean Distance
- Document Ranking (Display Order)
- Document Similarity (Among Documents)



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)