Project_1: Implementation and Experimental Evaluation of Basic DHTs

Professors: S. Sioutas, A. Komninos, G. Vonitsanos (Postdoc Researcher@CEID)

Goal: The major task is the implementation and experimental evaluation of a variety of DHTs in a programming language of your preference (we suggest Python, C++ or Java). You could use artificial synthetic-data sets or real-data sets in order to evaluate the performance of the following fundamental operations (Queries): Build, Insert key, Delete key, Update key, Lookup (key), Node Join, Node Leave.

You could also use a virtual or real distributed environment using threads and sockets. Dockers and Containers using k8s is also a realistic elastic distributed cloud-based environment suitable for the above implementations.

You can download real datasets from the following URLs:

Find Open Datasets and Machine Learning Projects | Kaggle
20 Free Datasets for Data Science Projects | Built In
https://freegisdata.rtwilson.com/
https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

DHTs (Chord && Pastry): Develop Chord and Pastry decentralized (p2p) infrastructures based on DHT methods for storing a set of **(key, value)** pairs, where **value** is a set of attributes related to a specific **key.** Evaluate the performance of the following basic operations: insert key, delete key, node join, node leave and lookup (or exact match) queries. Compare the number of hops each p2p protocol requires. Plot the performance comparison between Chord and Pastry for all the above operations.

You could use the the Coffee Reviews Dataset (Coffee Reviews Dataset) from Kaggle. This dataset organizes global reviews of coffee between 2017 and 2022 based on factors like blend name, type of roast, price and geographical origin of coffee beans. It is pre-processed and cleaned, and can be used for pandas, data engineering, analysis and feature engineering practice. The original version of the dataset comes with 12 features, while the simplified version has 9 features. f.e. we would like to detect the N-top most similar Reviews (documents) conducted during 2019 up to 2021, took review-rating more than 94, it's price per 100g (100g_USD) is between 4$ and 10$, **and the country origin (loc country) is USA**, where N is a user defined parameter (f.e. N=3). For the basic distributed lookup query, consider as **key** the "**loc country**" attribute. Having Located the correct peer, then you could use local centralized multidimensional data structure and LSH mechanisms to support all the other (except for key) attribute's constraints defined by the query. Towards to this direction, obviously, we must equip each peer with the appropriate local centralized multi-dimensional indexing and LSH mechanisms.

Background Knowledge: Data Structures, Multi-Dimensional Data Structures, Algorithms and Complexity, Databases, Object Oriented Programming (C++, JAVA), Functional Programming (Python, Scala).

References:

1. https://en.wikipedia.org/wiki/Chord_(peer-to-peer)
2. https://en.wikipedia.org/wiki/Pastry_(DHT)
3. https://eclass.upatras.gr/courses/CEID1411

(***) Deliverables: Zip or Rar file with executable files. Deadline: ~ 1st WEEK of February, 2024.