Katana Graph

Graph AI Preprocessing

27th June 2022

Company Background

Katana Graph is a startup company focusing on building a Graph Intelligence Platform for enterprise customers. More information can be found at the website (https://katanagraph.com/).

OVERVIEW

Preprocessing features is an essential part of a machine learning workflow. A good set of initial features can ensure and accelerate convergence. While there is a vast array of preprocessing operations in the machine learning literature, most approaches are non-distributed, i.e. they consider the entire dataset to be present at the time of computation. However, such a solution will not be scalable to very large datasets. Therefore, it is essential that a set of distributed preprocessing algorithms are developed. These functions should behave like their non-distributed counterparts but will have mutually exclusive portions of the dataset. The function should communicate in a distributed environment and exchange learning parameters or data to produce an efficient solution.

GOALS

In this project, a team of 4-5 students will learn critical components of computer science and machine learning: linear algebra, parallel programming and model pretraining. The students will be required to implement methods ranging from Distributed Principal component analysis, clustering and transformer models to featurize raw datasets. These methods will be graph agnostic. These methods can be integrated into Graph AI preprocessing library of Katana Graph.

Some of the algorithms that are intended for distributed implementation are the following:

- Decompositions
  - Principal Component Analysis
  - Latent Dirichlet Allocation (LDA) for topic models
- Clustering
  - K-Means Clustering
- Pre-trained Embedding Generators
  - Sentence Transformers

REQUIREMENTS

- An understanding of Python programming and data structures.
- Preliminary understanding of parallel processing. Knowledge of the OpenMPI package is recommended.
- Comfortable with matrix operations and linear algebra like eigenvectors, singular value decomposition, etc. is highly recommended.
- Preliminary understanding of Neural networks and implementing them.
- Knowledge of Pytorch, transformers and using off the shelf methods as part of a pipeline are recommended.

Intellectual Property Ownership

- Katana Graph owns the intellectual property based on the entire project