COSE474-2024F: Final Project Proposal
"Classifying the concept of CS terminology"

장동윤(Dongyoon Chang)

1. Introduction (Motivation, Problem Definition)

When I recently study CS, So many concepts of computer science and technology collide with each other, which makes me confusing. I feel the need to cluster all the concepts with related one.

Even writing similar ones together at noting app like Notion can be the solution for it, but keep noting all the terminologies and searching for what I want makes lots of fatigue. So I was looking for CS concepts_clustering(classifying) AI model, which can map concepts into bigger one.

2. Goal

The goal is classifying the concept of CS terminology into larger concept. At first, I choose several larger concepts, which can properly represent the vast CS_concepts. Then, I input some terms which is wanted to be clustered by bigger one.

For example, when I want to cluster Java's DI(Dependency Injection) concept, I suggest some bigger concepts, which can be computer_architecture, OS, network, DB, data_structure, algorithm, software_engineering, Back_End tech(OOP), Front_End_tech. Then, DI might be clustered to OOP, because its definition is about Spring, which is Java's framework. It controls the dependency of Java's objects, which makes developer more easier. When we use Hibernate as input term, it will be clustered to DB.

I'm gonna do prompt tuning. First, I will batch several datas based on my initial knowledge or some given datsets. Then, my model is gonna solve the coding question that I made, in order to improve the model

3. Pre-trained Model

As I achieve this goal, I'm going to use pre-trained model which is for NLP, labeling datas for text. In my search, I will use BERT or T5 for my project. Actually, T5 is much proper for mine, because we can easily do prompt tuning by it.

4. Datasets

When I'm searching at Hugging_Face, there are no proper dataset for coding terminology, so I'm gonna make it on my own.

5. State-of-the-art methods and baselines

I am gonna compare my result with GPT's answer, which is one of the State-of-the-art for text clustering.