CUHK RA Recruitment Technical Task Sheet
for Prof. Ling Cen (2025)

Ling Cen

Instruction

The problem sheet contains two programming tasks with some specifications and several sub-tasks for each. There are no limitations to the programming language, but Python is recommended. You are given one week to complete the tasks, and submit your code scripts and a short report before the deadline. Several submissions may be allowed, but the latest one will be counted only.

Academic Integrity Policy. Each candidates needs to finish the tasks independently.

1 Topic Model

An **earnings conference call** is a scheduled teleconference or webcast where a publicly traded company discusses its financial performance, typically including presentations from company executives, financial results analysis, future outlook, and a Q&A session with analysts and investors. In this task, a dataset of the questions extracted from Q&A session, raning from 2000 to 2022, is provided. You are required to train topic model(s) to mine hot topics over the years.

The dataset for this task is provided in `content_samples.jl`. Each sample includes multiple attributes, but only the following attributes are related to the task directly.

`Name` : Name of analyst
`Position` : Position of analyst
`Content` : The question asked by the analyst
`time` : The time when the analyst asked the question

Task 1.1. Encode `Content` into embeddings using large language models (LLMs).

Task 1.2. Use more than two methods to identify most top 20 frequently discussed hot topics of the questions of each 4 years ($[2000, 2003], [2004, 2007], \ldots, [2020, 2022]$).

Task 1.3. Use one method in Task 1.2. to identify the analysts who discussed these hot topics.

2 Name Match

Company name matching aims to identify and link different company names in databases whereas name on query table may different from name on candidates table, causing by variations, abbreviations, or other factors, of one company name. The dataset for this task includes two files namely as the following. Note that both tables have only one column.

`QueryNameSample.csv` and `CandidateNameSample.csv`

Task. For each company name in the query table, please search an identical, or the most similar name from the candidate table.