PSET 3
MEcon
6102

Dr. Ye Luo

April 2024

This work is designed to perform a comparison between nlp task and LLM task. The data contains a list of earning call transcripts. Take the Q-A (question and answer) sessions of the transcripts and analyze on each dialogue between a financial analyst and a senior manager. Our goal is to extract certain information from each dialogue.

1 Natural Language processing using Stanford's nlp package nltk.

It will take you about 30-60 minutes to do all of the followings. (a) Extract the data that contains a collection of earning call transcripts from the link

https://www.dropbox.com/s/sxx4n4d5p7xtyk6/01Earnings%20Call%20S%26P%20PDF.rar?dl=0

Download the 3 python scripts. Make sure that they are in the same folder.

(b) Install PyPDF2 in the anaconda prompt(or other environment of python) using commands such as pip install PyPDF2.

(c) Similarly, install nltk using pip install nltk

(d) Download glove.6B.50d file from the webite https://nlp.stanford.edu/projects/glove/, and put it in the same folder. You may need to download the entire glove package. This is a 50 dimensional representation of words, a (small) language model. As you can see, there are models with other dimensions.

(f) Install stopwords by command in the python environment:

import nltk

nltk.download('stopwords')

(g) There is a file of analyst list. Download it and put it in the same folder. You may open this file to examine it.

Try to perform the followings:

Step 1. Run analyst list.py. Briefly explain what this script does.

Step 2. Run pdf read.py. Briefly explain what this script does.

Step 3. Run earningcalls.py. Briefly explain what this script does. Read the Stanford NLP website and answer the following question: How do you calculate the similarity measure between question and answer?

Step 4. Compile a table of Analyst, Firm of the Analyst, and paragraph-wise QA-similarity measure such as paragraph similarity measure. Report the percentage of missing data points from the nlp scripts being provided (i.e., some of the dialogues are not well processed or not recognized by the nltk package and the scripts).

2 LLM Extraction

You should log in your HKU chatgpt account. Choose 25 Question-Answer dialogues (copy paste them into a dataframe, for example).

Write a prompt for processing these 25 dataframe to extract:

(1) Analyst name.

(2) Firm name of the analyst.

(3) QA paragraph-wise similarity measure.

Compare the results extracted by NLP scripts and the results extracted by LLM. Briefly comment on the comparison.