Natural Language Generation from Structured Financial Data

Team #12: Huang XinYi, Gunjan Agarwal, Christie Wong, Xin Haohong, Ho Jing Kai

COMPANY BACKGROUND

- German semiconductor manufacturer
- **Line of business:** Business Intelligence Analytics (BIA)
- **Primary Client:** Finance Department
- **Role:** Providing service to clients by using conventional BI and AI/ML analytics

OVERVIEW

Problem Statement

Manually interpret data → Finance department cannot generate report **timely** + Extensive **cost** to company

Project Objectives

Design **NLG model** for data analysis → **Shorter** analytic process, More **accurate** summaries + **Automated** process, Rendering manual labour redundant in data processing

METHODOLOGY

Studying NLG Models

- Research and study suitable NLG models to use
- Run NLG models
- Understand how the model works & its limitations
- Choose desired models that fulfills project objective

Using Python, PowerBI, Tableau, Excel

- Install necessary python packages
- Identify data type, components and functions of model

Data Analysis

- Conduct Exploratory Data Analysis (EDA)
- Clean and process data
- Determine the text content to be generated E.g mean, median, trends

Manually input the generated data analysis into the NLG model

Tailor the style of generated text summaries to be informative

Improve on the limitations of NLG models

RESULTS & ANALYSIS

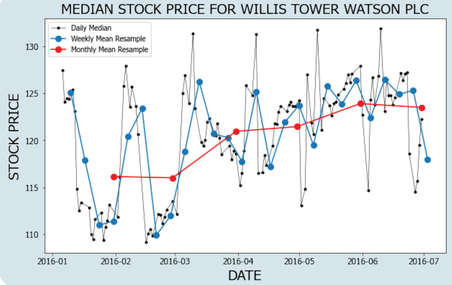
Exploratory Data Analysis

Step 1: Understanding the data - ***Stock prices of companies on the NYSE***

Date	Symbol	Open	Close	Low	High	Volume
29/12/2016	AIV	44.4	44.7	44.3	44.8	653200
29/12/2016	FTV	54	54	53.6	54.1	479500
30/12/2016	WLTW	122.6	122.3	121.4	123.6	466400
30/12/2016	A	45.8	45.6	45.4	45.8	1216100

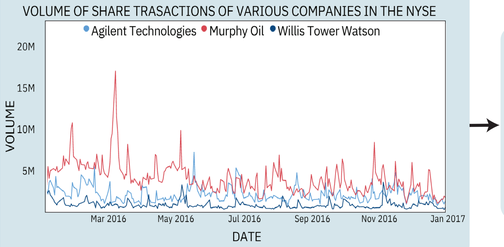
Symbol: Alphabetic code for company
Open/Close: Stock price at time of market opening/ closing
Low/High: Lowest/ highest recorded stock price of the day
Volume: No. of transactions in shares

Step 2: Data Analysis



Weekly/Monthly Mean Resample: Group data into weeks/ months and compute the mean

Useful information for text summary: What is the daily median, weekly/ monthly mean trends?



Volume of share transactions: Plot the trend of the volumes of shares transacted for different companies for comparison

Useful information for text summary: Which company has the lowest volume of shares transacted on a particular day?

NLG Models

1 GPT-2: Generative Pre-trained Transformer 2

GPT-2 is a transformers model pre-trained on a very large corpus of English data in a self-supervised way. The model was pretrained on raw texts only, with no human labeling.

START → **Feeding Data Into Model** → **GPT-2**

"Extractive Question Answering is the task of extracting an answer from a text given a question. An example of a question answering dataset is the SQuAD dataset, which is entirely based on that task."

Selection of GPT-2 Task: Extraction Question Answering

- **Iterate** over the questions. Build a sequence from the text and the current question, with the correct model-specific token type ids and attention masks.
- **Compute** the softmax of the result to get probabilities over the tokens.
- **Fetch** the tokens, convert those tokens to a string.

Ask questions using the "question_answer" command

Question: "What is extractive question answering?"

Output: Answer, Accuracy score

Data answer : 'the task of extracting an answer from a text given a question', score: 0.6177

2 T5: Text-To-Text Transfer Transformer

- T5 is a unified framework that converts every language problem into a **text-to-text format.**
- **Model Structure:** Encoder-Decoder
- **Attention mask mechanism:** "fully-visible"

Rank	Heat	Lane	Name	Nationality	Time
1	6	4	Matt Grevers	United States	52.92
2	6	2	Cheng Feiyi	China	53.22
3	4	4	Nick Thoman	United States	53.48
4	5	4	Camille Lacourt	France	53.51
5	6	5	Ryosuke Irie	Japan	53.56

Text summary generated from structured table

Lacourt was dropped to a fourth-place time in 53.51.

Solution Overview

T5 converts structured data to a text summary, which will be fed into the Extraction Question Answering task empowered by GPT-2, allowing the user to get specific answers from the text summary.

Data: Prices.csv

Date	Symbol	Open	Close	Low	High	Volume
3/1/2016	MSFT	53.79	51.64	51.29	54.07	66883600
3/1/2016	FCX	4.18	3.74	3.7	4.28	64602100
3/1/2016	GE	28.91	28.24	28.2	29.05	55717700
3/1/2016	WMB	16.48	13.61	12.77	16.54	52053600
3/1/2016	WLTW	114.27	115.51	112.65	115.57	694100

*This is a limited representation of the prices.csv data that we are using.

→ **T5** →

Text Summary generated by T5 Model:
The **daily median** on 1/3/2016 for WLTW is 141.11
The **weekly mean** from 1/3/2016 to 8/3/2016 for WLTW is 116.08
The **monthly mean** from 1/3/2016 to 31/3/2016 for WLTW is 118.51
MTD has the **lowest volume** of 236400 transacted on 1/3/2016.
MRO has the **highest volume** of 273996600 transcated on 1/3/2016.

→ **GPT-2** →

Extraction Question Answering by GPT-2 Model:

Question input:
What is the daily median for 1/3/2016 for the company WLTW?

Output Answer: '141.11', score: 0.9742

ASSUMPTIONS & LIMITATIONS

- Infineon Technologies' financial data is assumed to be **structured similarly to prices.csv**.
- Data must be **formatted specifically for the NLG model** to read, as the data cannot be cleaned automatically.
- GPT-2 model only accepts strings instead of dataframes, resulting in **input compatibility issues**.
- Both models are **not trained with data from prices.csv**, leading to inaccurate answers.

CONCLUSION

- Running the **T5 and GPT-2 models sequentially** on the financial data would yield favourable results.
- However, the **T5 model is more relevant** to the project objectives and the interests of Infineon Technologies.
- More research is required to **understand the data training process** of NLG models.
- With adequate competency, the **T5 model can be modified** to be compatible with the desired financial data.