ScrapingGenAI

- James Do, Heewoon Bae, Julius Colby
- CS 4624 Multimedia, Hypertext, and Information Access
  - Dr. Fox
- Virginia Tech, Blacksburg 24061, 4/30/2024
Outline

• Project Design
• Implementations
  • MarketBeat
  • Reddit
• Challenges
• Future Work
• Acknowledgements
• References
### Project Design

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Data Processing/Analysis</th>
<th>Data Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Create MarketBeat Scraper to scrape earning calls</td>
<td>• Develop a parser to analyze and process data of both MarketBeat and Reddit</td>
<td>• Utilize Matplotlib and Pandas libraries to create visualizations</td>
</tr>
<tr>
<td>• Create Reddit Scraper to scrape posts and comments</td>
<td>• Utilize a Sentiment Intensity Analyzer to better process the collected data</td>
<td>• Visualize processed data through various graphs and charts to effectively discern trends and perform numerical analyses of specific keywords</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Used Tkinter Python library for user GUI</td>
</tr>
</tbody>
</table>
MarketBeat Implementation

- Scraped an online earning call database (MarketBeat) for links to company earning call transcripts
  - Earning call transcripts:
    - Quarterly reports about the financial state of a company
- Parsed individual web pages to pull and store the transcripts
- Applied a keyword search to find relevant sections of the transcript and fed them into a natural language processor to decide positivity
- Created a visualization showcasing positivity and relativity
MarketBeat Scraper

Scraped root of MarketBeat.com to gather links to separate earning call transcripts

- Used BeautifulSoup
- Fed each link to transcript into parser
MarketBeat Parser

Parsed each company transcript to provide specified output

- Used BeautifulSoup
- Requested output: Stock Ticker, Company Name, Date of Transcript, and the Transcript itself
- Outputs CSV and JSON

```
"Tesla": {
  "Company": "Tesla",
  "Year": "2023",
  "Quarter": "Q1",
  "Date": "04/23/23 1:30 PM ET",
  "Transcript": "My name is Martin Viehe, VP of Investor Relations, and I’m joined to ![Operator Instructions] But before we jump into Q&A, Elon has some o
  "Thanks, Martin. To recap, in Q1, we navigated several unforeseen challenges. We actually know it will, so significantly faster than the car business. These new vehicles, including more affordable models will use aspects of R&D Version 12, which is the pure AI-based self-driving, I’ll say. To make it more accessible, we've reduced the subscription price to $4,000. So in conclusion, we're super excited about our autonomy roadmap. I want to thank you very much, and Vaibhav has some comments as well.
  "Thanks. It’s important to acknowledge what Elon said, from our auto survey, the impact of pricing actions was largely offset by reductions in penmanship fees. In fact, if we exclude Cybertruck and Fremont Model 3 ramp costs, the order demand on the supply side, we’ve undertaken a variety of initiatives, including the introduction of new software. And accordingly, this business will begin contributing significantly on the operating expense front, we saw a sequential increase from one to another. As we prepare the company for the next phase of growth, we had to make some key initiatives. In conclusion, the future is extremely bright and the journey to get there is diaspora.
  "Okay. Let's start with investor Q&A. The first question is, what is the Tesla’s future? Sure. 4680 production increased about 10x, 20x over -- from Q4 reach to 100.
  "Thank you. The second question is on Optimus. So what is the current state of Optimus? We are able to do simple factory tasks or at least, I should say, up to "because if you've got a sentient humanoid robots that is able to navigate "I mean this, perhaps, is a point that is worth emphasizing Tesla's "Thank you. The third question is, what is the current assessment of "I can start. There are a handful of states that already have adopted"
```
MarketBeat Analyzer

Prompts user to input search keywords and returns sections of the transcript that included them

• Returns relevant keywords to output file
• Runs sentiment analyzer on section and returns a positivity rating

JSON output of analyzer
<table>
<thead>
<tr>
<th>Ticker</th>
<th>Company</th>
<th>Year</th>
<th>Quarter</th>
<th>Date</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSLA</td>
<td>Tesla</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 5:30 PM E</td>
<td>My name is Martin V. Operator instruction for Tesla.</td>
</tr>
<tr>
<td>V</td>
<td>Visa</td>
<td>2024</td>
<td>Q2</td>
<td>04/23/24 6:00 PM E</td>
<td>Welcome to Visa’s Pal! I would love to hear from you.</td>
</tr>
<tr>
<td>STX</td>
<td>Seagate Technology</td>
<td>2024</td>
<td>Q3</td>
<td>04/23/24 5:00 PM E</td>
<td>Welcome to the Sea! I would love to hear from you.</td>
</tr>
<tr>
<td>ENPH</td>
<td>Enphase Energy</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 4:30 PM E</td>
<td>Good day, and welcome Good afternoon and</td>
</tr>
<tr>
<td>TXN</td>
<td>Texas Instruments</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 4:30 PM E</td>
<td>Welcome to the Texas Test! This call will include</td>
</tr>
<tr>
<td>AIRE</td>
<td>Alexandria Real Estate</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 4:00 PM E</td>
<td>Good day, and welcome Would you like to hear from a</td>
</tr>
<tr>
<td>LMT</td>
<td>Lockheed Martin</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 11:00 AM</td>
<td>Good day and welcome Now, at this time for</td>
</tr>
<tr>
<td>SNWY</td>
<td>Shinrin Williams</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 11:00 AM</td>
<td>Good morning, That’s conference call</td>
</tr>
<tr>
<td>GL</td>
<td>Globe Life</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 11:00 AM</td>
<td>Hello, and welcome And I’d love to turn th</td>
</tr>
<tr>
<td>MSCI</td>
<td>MSCI</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 10:00 AM</td>
<td>Good day, ladies and gentlemen. Thank you, Roger.</td>
</tr>
<tr>
<td>FOX</td>
<td>FreesportMcMotR</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 10:00 AM</td>
<td>Ladies and gentlemen. Thank you, Roger.</td>
</tr>
<tr>
<td>NUE</td>
<td>NuCor</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 10:00 AM</td>
<td>Good morning, and I would love to hear from you.</td>
</tr>
<tr>
<td>HAL</td>
<td>Halliburton</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Good day, and I would love to hear from you.</td>
</tr>
<tr>
<td>PM</td>
<td>Philip Morris Internet</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Good day, and I would love to hear from you.</td>
</tr>
<tr>
<td>IVZ</td>
<td>Intrexco</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Welcome to Intrexco! Now I’ll take your call th</td>
</tr>
<tr>
<td>PKQ</td>
<td>Packaging Co. of Art</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Good morning, everyone. At this time, I’d like to</td>
</tr>
<tr>
<td>WBB</td>
<td>W. R. Berkley</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Good day and welcome Would you like to hear from</td>
</tr>
<tr>
<td>AMP</td>
<td>Aeroparts Financial</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Welcome to the Q1 call now. I will now turn the ea</td>
</tr>
<tr>
<td>PNR</td>
<td>Pinnacle</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Hello and welcome. Thank you, and welcome</td>
</tr>
<tr>
<td>NEE</td>
<td>NexEra Energy</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:00 AM E</td>
<td>Good morning, and I would love to hear from you.</td>
</tr>
<tr>
<td>DGX</td>
<td>Quest Diagnostics</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Welcome to the Quest Diagnostics call.</td>
</tr>
<tr>
<td>RTX</td>
<td>RTX</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Good day, ladies and gentlemen. On the call today are</td>
</tr>
<tr>
<td>KMB</td>
<td>Kimberly-Clark</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Good morning, every It’s your morning call, please</td>
</tr>
<tr>
<td>GM</td>
<td>General Motors</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Good morning, and I’d like a reminder. This is</td>
</tr>
<tr>
<td>UPS</td>
<td>United Parcel Servic</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Good morning, everyone. It’s your call, please</td>
</tr>
<tr>
<td>PHM</td>
<td>Phillips Group</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:30 AM E</td>
<td>Thank you for standing. Thank you. I would love to</td>
</tr>
<tr>
<td>PEP</td>
<td>Pepsico</td>
<td>2024</td>
<td>Q1</td>
<td>04/23/24 9:15 AM E</td>
<td>Good morning, and it’s your call, please.</td>
</tr>
</tbody>
</table>
| LKQ    | LKQ             | 2024 | Q1      | 04/23/24 8:00 AM E | Good morning, everyone. I will now hand you o
MarketBeat Visualizer

Takes data returned by the analyzer and formats a graph showcasing which companies are relevant

• Uses Matplotlib
• Shows positivity rating
Reddit Implementation

• Scraped Reddit app to collect posts and comments of certain subreddits
• Utilized a Reddit API to be able to scrape posts and comments
• Implemented a keyword search through the posts and comments for users to specify what to scrape
• Added Sentiment Intensity Analyzer to determine the positivity score of posts and comments
• Created visualization that shows the positivity of posts/comments and through a monthly basis and analyze trends on certain keywords
Reddit Scraper

• Scraped Reddit to obtain posts and comments that include one of the keywords from the user
• Utilized Reddit’s API to be able to pull content from the app

```bash
[jamesdo@willow CS4624]$ python redditScraper_json.py
Enter the subreddit name: artificial
Enter keywords separated by comma: travel
Enter the start date in YYYY-MM-DD format (or leave blank for no start date): 2023-01-01
Enter the end date in YYYY-MM-DD format (or leave blank for no end date): 2023-12-31
Enter the file path to save the search results (e.g., output.json): test15.json
[jamesdo@willow CS4624]$ Ø
```
Reddit Parser

Parses each post and their comments to get the data of the positivity score and the keywords used when scraping

- Utilizes the NLTK library for the positivity score
- Outputs a JSON file
Reddit Analyzer

User will be prompted to put in a subreddit, specific keywords and timeframe of how long the scraper should scrape

• Output will be the total number of comments scraped per month, keywords comments, and average positivity score
Reddit Visualization

Takes in Reddit scraper user input format to then create a pie and line graph. The pie graph shows the total number of comments with pieces being the months. The line graph shows the change in the number of comments that contain the keywords.

• Combined web scraper and data visualizer into a single script for user
• Used Tkinter Python library for user GUI
• Used Matplotliblib and Pandas for data visualization
Reddit Visualization
Challenges

• Reddit API issues
• User input issues
• Reddit API rate limits
• Ensuring data consistency
Future Work

• Implement a scraper to pull transcripts from individual sectors
• Implement a scraper to pull transcripts from specific companies and their history
• Implement a front end to the Reddit scraper
• Improve upon the functionality of the Reddit scraper
Acknowledgements

Florian Zach

Edward Fox
References


References

  https://pypi.org/project/beautifulsoup4/ [accessed May 4, 2024]
  https://pandas.pydata.org/ [access May 4, 2024]
  https://matplotlib.org/ [accessed May 4, 2024]
  https://www.nltk.org/ [accessed May 4, 2024]
  https://medium.com/@sujathamudadal1213/explain-pattern-python-library-5bb4c530956b [accessed May 4, 2024]