# NLP Asset Analysis

## Dependencies & Imports

In [ ]:
```
%%capture
%pip install transformers
%pip install sentencepiece
%pip install torch torchvision
```

In [ ]:
```
from transformers import PegasusTokenizer, PegasusForConditionalGeneration, logging # type: ignore
from bs4 import BeautifulSoup # type: ignore
import requests # type: ignore
logging.set_verbosity_error()
```

## Summarization Model

In [ ]:
```
model_name = 'human-centered-summarization/financial-summarization-pegasus'
tokenizer = PegasusTokenizer.from_pretrained(model_name)
model = PegasusForConditionalGeneration.from_pretrained(model_name)
```

## News & Sentiment Pipeline

In [ ]:
```
tickers = ['GOOG', 'TSLA', 'BTC']
```

### Search for Stock News

In [ ]:
```
def get_stock_news_urls(ticker):
    search_url = 'https://www.google.com/search?q=yahoo+finance+{}+news&tbm=nws'.format(ticker)
    r = requests.get(search_url)
    soup = BeautifulSoup(r.text, 'html.parser')
    atags = soup.find_all('a')
    hrefs = [link['href'] for link in atags]
    return hrefs
```

In [ ]:
```
raw_urls = {ticker : get_stock_news_urls(ticker) for ticker in tickers};
```

### Filter URLs

In [ ]:
```
import re
```

In [ ]:
```
exclude_list = ['maps', 'policies', 'preferences', 'accounts', 'support', 'search']
```

In [ ]:
```
def filter_urls(urls, exclude_list):
    val = []
    for url in urls:
        if 'https://' in url and 'finance' in url and not any(exclude_word in url for exclude_word in exclude_list):
            res = re.findall(r'(https?://\S+)', url)[0].split('&')[0]
            val.append(res)
    return list(set(val))[:5]
```

In [ ]:
```
cleaned_urls = {ticker : filter_urls(raw_urls[ticker], exclude_list) for ticker in tickers}
cleaned_urls
```

Out[ ]:
```
{'GOOG': ['https://finance.yahoo.com/news/amazon-earnings-195332083.html',
  'https://finance.yahoo.com/news/alphabet-earnings-july-25-153933519.html',
  'https://finance.yahoo.com/news/alphabet-stock-rises-4-after-google-rolls-out-new-bard-features-international-expansion-070139824.html',
  'https://finance.yahoo.com/news/google-parent-alphabets-stock-top-175837293.html',
  'https://finance.yahoo.com/news/want-job-working-ai-ll-180044780.html'],
 'TSLA': ['https://finance.yahoo.com/news/tesla-model-y-model-3-probed-by-nhtsa-on-loss-of-steering-complaints-174245647.html',
  'https://finance.yahoo.com/news/tesla-stock-tumbles-nearly-10-on-margin-concerns-musk-warning-on-q3-production-200240714.html',
  'https://finance.yahoo.com/news/lucid-board-member-on-ev-competition-were-not-targeting-tesla-133735484.html',
  'https://finance.yahoo.com/news/nasdaq-sinks-as-tesla-netflix-plunge-dow-gains-for-9th-day-stock-market-news-today-200246194.html',
  'https://finance.yahoo.com/news/10-best-ev-battery-autonomous-222242368.html'],
 'BTC': ['https://finance.yahoo.com/news/binance-boosts-first-digitals-stablecoin-155401419.html',
  'https://finance.yahoo.com/news/sai-tech-announces-host-upcoming-120000199.html',
  'https://uk.finance.yahoo.com/news/bitcoin-price-crypto-binance-fraud-charge-us-092356621.html',
  'https://finance.yahoo.com/news/bear-day-microstrategy-mstr-102000522.html',
  'https://finance.yahoo.com/news/bitcoin-stays-above-us-30-015141435.html']}
```

### Scrape URLs

In [ ]:
```
def process(urls):
    articles = []
    for url in urls:
        r = requests.get(url)
        soup = BeautifulSoup(r.text, 'html.parser')
        paragraphs = soup.find_all('p')
        text = [paragraph.text for paragraph in paragraphs]
        words = ' '.join(text).split(' ')[:350]
        article = ' '.join(words)
        articles.append(article)
    return articles
```

In [ ]:
```
articles = {ticker : process(cleaned_urls[ticker]) for ticker in tickers}
```

### Summarize

In [ ]:
```
def summarize(articles):
    summaries = []
    for article in articles:
        input_ids = tokenizer.encode(article, return_tensors='pt')
        output = model.generate(input_ids, max_length=100, num_beams=5, early_stopping=True)
        summary = tokenizer.decode(output[0], skip_special_tokens=True)
        summaries.append(summary)
    return  summaries
```

In [ ]:
```
summaries = {ticker : summarize(articles[ticker]) for ticker in tickers}
summaries
```

Out[ ]:
```
{'GOOG': ['Second-quarter profit beat came despite slowdown in sales of Amazon Web Services. Revenue outlook was a striking beat, coming in between $138 billion and $143 billion',
  'Google parent company reported revenue of $74.6 billion. YouTube ad revenues also topped estimates',
  'Google's chatbot is now available in more than 40 languages. New features include audio responses, Pinned conversations',
  "AFP sues Musk's X social media platform, accusing it of neglecting. Regulatory Hurdles: Agence France-Presse (AFP) has initiated legal proceedings against Elon Musk",
  'How to make money with AI and the skills needed. Here are the types of jobs that you can find in AI'],
 'TSLA': ['NHTSA says investigation covers an estimated 280,000 Model 3 and Model Y cars. Tesla may not be able to fix issue via software update',
  'Margins came in below expectations, but CEO says demand for new truck is off the hook.',
  'Lucid says EV market 'will explode,' not ' deliberately targeting Tesla.' Shares down more than 20% this year.',
  'Tesla, Netflix results failed to excite investors. Johnson & Johnson, American Airlines also reported results',
  'IEA expects a significant increase in electric vehicle sales compared to 2022. India, Thailand, and Indonesia experienced remarkable growth in 2022'],
 'BTC': ['Exchange to waive fees for BTC/FDUSD, maker fees until further notice.',
  'SAI's new U.S. R&D Center and OCEC Computing Heat Recycle Center to be unveiled in Marietta, Ohio.',
  'DOJ is considering bringing fraud charges against Bitstamp, reports Semafor. Bitcoin is wavering above the $29,000 level as trading volumes remain low',
  'Founder and Chairman Michael Saylor says company is all-in on crypto. But Wall Street analysts need more visibility on core business',
  'Ether edged lower along with most other top 10 cryptocurrencies. JPMorgan report claims recent flurry of Bitcoin ETF applications may not be game-changer'}
```

## Sentiment Analysis

In [ ]:
```
from transformers import pipeline # type: ignore
sentiment = pipeline('sentiment-analysis')
```

In [ ]:
```
scores = {ticker : sentiment(summaries[ticker]) for ticker in tickers}
scores
```

Out[ ]:
```
{'GOOG': [{'label': 'POSITIVE', 'score': 0.9606388807296753},
  {'label': 'POSITIVE', 'score': 0.9964740872383110},
  {'label': 'POSITIVE', 'score': 0.9792576432220888},
  {'label': 'NEGATIVE', 'score': 0.9980740547180176},
  {'label': 'POSITIVE', 'score': 0.9550225734710693}],
 'TSLA': [{'label': 'NEGATIVE', 'score': 0.9995212554931641},
  {'label': 'NEGATIVE', 'score': 0.9878749251365662},
  {'label': 'NEGATIVE', 'score': 0.9985577464103699},
  {'label': 'NEGATIVE', 'score': 0.9997245669364929},
  {'label': 'POSITIVE', 'score': 0.9981102943420411}],
 'BTC': [{'label': 'NEGATIVE', 'score': 0.9841852188110352},
  {'label': 'POSITIVE', 'score': 0.9958518743515015},
  {'label': 'NEGATIVE', 'score': 0.9990792274475098},
  {'label': 'NEGATIVE', 'score': 0.9965921640396118},
  {'label': 'NEGATIVE', 'score': 0.9975816011428833}]}
```

## Exporting to CSV

In [ ]:
```
def create_output_array(summaries, scores, urls):
    output = [['Ticker', 'Summary', 'Label', 'Confidence', 'URL']]
    for ticker in tickers:
        for i in range(len(summaries[ticker])):
            output_this = [
                ticker,
                summaries[ticker][i],
                scores[ticker][i]['label'],
                scores[ticker][i]['score'],
                urls[ticker][i]
            ]
            output.append(output_this)
    return output
```

In [ ]:
```
output = create_output_array(summaries, scores, cleaned_urls)
output
```

Out[ ]:
```
[['Ticker', 'Summary', 'Label', 'Confidence', 'URL'],
 ['GOOG',
  'Second-quarter profit beat came despite slowdown in sales of Amazon Web Services. Revenue outlook was a striking beat, coming in between $138 billion and $143 billion',
  'POSITIVE',
  0.9606388807296753,
  'https://finance.yahoo.com/news/amazon-earnings-195332083.html'],
 ['GOOG',
  'Google parent company reported revenue of $74.6 billion. YouTube ad revenues also topped estimates',
  'POSITIVE',
  0.9964740872383110,
  'https://finance.yahoo.com/news/alphabet-earnings-july-25-153933519.html'],
 ['GOOG',
  'Google's chatbot is now available in more than 40 languages. New features include audio responses, Pinned conversations',
  'POSITIVE',
  0.9792576432220888,
  'https://finance.yahoo.com/news/alphabet-stock-rises-4-after-google-rolls-out-new-bard-features-international-expansion-070139824.html'],
 ['GOOG',
  "AFP sues Musk's X social media platform, accusing it of neglecting. Regulatory Hurdles: Agence France-Presse (AFP) has initiated legal proceedings against Elon Musk",
  'NEGATIVE',
  0.9980740547180176,
  'https://finance.yahoo.com/news/google-parent-alphabets-stock-top-175837293.html'],
 ['GOOG',
  'How to make money with AI and the skills needed. Here are the types of jobs that you can find in AI',
  'POSITIVE',
  0.9550225734710693,
  'https://finance.yahoo.com/news/want-job-working-ai-ll-180044780.html'],
 ['TSLA',
  'NHTSA says investigation covers an estimated 280,000 Model 3 and Model Y cars. Tesla may not be able to fix issue via software update',
  'NEGATIVE',
  0.9995212554931641,
  'https://finance.yahoo.com/news/tesla-model-y-model-3-probed-by-nhtsa-on-loss-of-steering-complaints-174245647.html'],
 ['TSLA',
  'Margins came in below expectations, but CEO says demand for new truck is off the hook.',
  'NEGATIVE',
  0.9878749251365662,
  'https://finance.yahoo.com/news/tesla-stock-tumbles-nearly-10-on-margin-concerns-musk-warning-on-q3-production-200240714.html'],
 ['TSLA',
  'Lucid says EV market 'will explode,' not ' deliberately targeting Tesla.' Shares down more than 20% this year.',
  'NEGATIVE',
  0.9985577464103699,
  'https://finance.yahoo.com/news/lucid-board-member-on-ev-competition-were-not-targeting-tesla-133735484.html'],
 ['TSLA',
  'Tesla, Netflix results failed to excite investors. Johnson & Johnson, American Airlines also reported results',
  'NEGATIVE',
  0.9997245669364929,
  'https://finance.yahoo.com/news/nasdaq-sinks-as-tesla-netflix-plunge-dow-gains-for-9th-day-stock-market-news-today-200246194.html'],
 ['TSLA',
  'IEA expects a significant increase in electric vehicle sales compared to 2022. India, Thailand, and Indonesia experienced remarkable growth in 2022',
  'POSITIVE',
  0.9981102943420411,
  'https://finance.yahoo.com/news/10-best-ev-battery-autonomous-222242368.html'],
 ['BTC',
  'Exchange to waive fees for BTC/FDUSD, maker fees until further notice.',
  'NEGATIVE',
  0.9841852188110352,
  'https://finance.yahoo.com/news/binance-boosts-first-digitals-stablecoin-155401419.html'],
 ['BTC',
  'SAI's new U.S. R&D Center and OCEC Computing Heat Recycle Center to be unveiled in Marietta, Ohio.',
  'POSITIVE',
  0.9958518743515015,
  'https://finance.yahoo.com/news/sai-tech-announces-host-upcoming-120000199.html'],
 ['BTC',
  'DOJ is considering bringing fraud charges against Bitstamp, reports Semafor. Bitcoin is wavering above the $29,000 level as trading volumes remain low',
  'NEGATIVE',
  0.9990792274475098,
  'https://uk.finance.yahoo.com/news/bitcoin-price-crypto-binance-fraud-charge-us-092356621.html'],
 ['BTC',
  'Founder and Chairman Michael Saylor says company is all-in on crypto. But Wall Street analysts need more visibility on core business',
  'NEGATIVE',
  0.9965921640396118,
  'https://finance.yahoo.com/news/bear-day-microstrategy-mstr-102000522.html'],
 ['BTC',
  'Ether edged lower along with most other top 10 cryptocurrencies. JPMorgan report claims recent flurry of Bitcoin ETF applications may not be game-changer',
  'NEGATIVE',
  0.9975816011428833,
  'https://finance.yahoo.com/news/bitcoin-stays-above-us-30-015141435.html']]
```

In [ ]:
```
import csv
with open('summaries.csv', mode='w', newline='') as f:
    csv_writer = csv.writer(f, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
    csv_writer.writerows(output)
```