Problem Identification:
Statement: Can we identify correlations in community activity and fluctuations in the stock price leading up to the Jan ‘21 short squeeze of $GME?

Context: In Jan of 2021 a grassroots group of individuals on r/Wallstreetbets acted on information that GameStop’s publicly held stock options were at a short rate of 140%. Through cooperative action stock prices increase from $17.25 to over $500 costing financial institutions Billions of dollars. Posts about $GME can be traced back before 2016 with increasing frequency before Jan ‘21.
Criteria for success:

- An LDA model that dynamically identifies key topics at either the post or comment level and reveals a correlation between topic frequency and stock price.

Scope of solution space:

- Collection of all comments from the millions of posts related to $GME from Jan ‘21 back to Dec ‘16.
- An LDA model that does Dynamic Topic Modeling.
- A Power BI reporting tool that visualizes topic frequencies alongside $GME stock prices.

Constraints:

- The PSAW API does not have accurate historical records for the subreddit.
- I collected and stored the wrong data in my database.
- The model doesn’t point to any meaningful information in the data.

Stakeholders:

- Manager
- CFO
- CEO
- CTO

Key data sources:

- CSV containing open and close prices as well as diffs for $GME
- Two tables in a locally hosted database with comment and post information collected using the PSAW API (see below)

Data sources outline:

```
- TABLE posts (
-    id TEXT NOT NULL,
-    dt TIMESTAMP WITHOUT TIME ZONE NOT NULL,
-    title TEXT NOT NULL,
-    author TEXT NOT NULL,
-    mentions TEXT,
-    body TEXT,
-    link TEXT,
-    PRIMARY KEY (id, dt)
- );
```

```
- TABLE comments (
-    id TEXT NOT NULL,
-    post_id TEXT NOT NULL,
-    parent_id TEXT,
-    dt TIMESTAMP WITHOUT TIME ZONE NOT NULL,
-    is_op BOOLEAN NOT NULL,
-    author TEXT NOT NULL,
-    body TEXT,
-    PRIMARY KEY (id, dt),
-    CONSTRAINT fkpost_id FOREIGN KEY (post_id) REFERENCES posts(id)
-
- );
```

-