General Info:

* Determine who will be the presenter on your proposal video.
* Random Decide on the recording package to use (We used PowerPoint, but I needed a non-web version for some of the editing features)
* **The project has three main components which you need to satisfy: 1) an interesting and large dataset 2) the application and development of a non-trivial algorithm for analysis and 3) the visualization of the results**

**Idea Topic**

1. What are you trying to do? Articulate your objectives using absolutely no jargon.

Analyzing with Financial Data over Time

* Sentiment Analysis of Data over time
* Topic Modelling of various dominant topics
* Summarization of big financial documents

1. How is it done today; what are the limits of current practice?

* Analysts go through the whole documents which needs specialized knowledge
* Time consuming to read through every sentence for people interested in investing in any company
* 10K financial reports easier for non-financial experts

1. What's new in your approach? Why will it be successful?
   * Add a time component to it?
   * Sentiment analysis not done extensively on 10-K reports
   * Current Usage/ Later Usage- Can add features to stock prediction models to
2. Who cares?
   * Covered in 2
3. If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
   * Survey
4. What are the risks and payoffs?
   * Some sources of Data needs to scrape through API – might need some research to do
   * Find out how to make embedding (Processing)
5. How much will it cost?
   * Data Storage: For storing the corpus on AWS s3 (Storage), it is $0.021 per GB. AWS Athena(Query) $5 per TB data scanned and $2/hr AWS EC2 (48 vCPU, 192 GiB Memory) instance for compute,
6. How long will it take?

For a group comprising six members, we are expecting the project to be completed in one-and-a-half to two months and will be dedicating at least 6-8 hours of time per week per member in the first few stages which translates to 200-300 hours of work in total.

1. What are the midterm and final "exams" to check for success?

Similar Project Video: <https://www.youtube.com/embed/KgzmUvG1BAE>

**Data Collection**

**Data Source1**: Reddit Data

Data collection and preprocessing: For data collection, we can use **Pushshift API** to scrape the Reddit posts and top 10 comments for each post. Then we performed text cleaning for both the posts and the comments.

1. You need to specify the subreddit, start date, and end date, data\_collection\_updated.py script can be used to collect comments and the URL content if it exists.

Link to Pushshift: <https://medium.com/mcd-unison/using-pushshift-api-for-data-analysis-on-reddit-b08d339c48b8>

Data Source 2(Financial Documents - For 10-K/10-Q/8-K): SEC Edgar data

<https://medium.com/@oshojha/useful-sentiment-analysis-mining-sec-filings-part-1-358942fc98ed>

Data Souce 3 (News Data) – stock market events, financial news head

2. Data Cleaning It will mostly includes removing whitespaces, special characters, hyperlinks, stop words, punctuations, doing lemmatization.

**Modelling** –

Embeddings from Text data – Converting text to numeric vectors (Tokenize)

Possible methods -

Sentiment Analysis: fastText , Naïve Bayes Support Vector Machine (NBSVM) , Neural Networks , BERT

**Visualisation:**

People have used React, Tableau, Plotly Dash

A similar project which used