Final Project Proposal

DS 710

For your final project, you will collect and analyze data from Twitter to answer a question that interests you.

1. What is your question? You might be interested in a brand, an organization, a meme, a trend, a news story, or a place. What do you want to learn about? Do you have predictions or hypotheses about what you will find?

My question relates to how people tweet in the morning versus at night. I want to learn if there is more volume in the morning or at night. I am also interested in knowing the most popular words for both.

I predict that the morning tweets will be more optimistic. I don’t know about whether the volume will be greater or smaller in the morning compared to the evening.

1. What is your audience? Who is interested in the answer to your question?

The obvious answer to this question would be Twitter. This could help them learn about the usage of their product.

Beyond that, this could add in our psychology understanding. Do we start the day fresh? What along the way changes that?

1. What information will you gather? What features will the tweets you collect have?

I will gather the tweets from random days as far back as I can go. I will try to restrict the US. This may be challenging so this requirement could change. The most important feature of the tweet will be the timestamp. I am hoping this is in both the user’s local time as well as UTC time.

1. How will you collect the data? Will you use the REST or Streaming APIs, or a combination of both?

I may use the REST API to look back as far as I can. I also may set up and let the Streaming API run for a day.

1. How will you analyze the data? What features are you looking for? Are there specific hypotheses you plan to test?

I will be using frequency counts for the two groups. I may do some sentiment analysis on the most common words for both night and day. I plan to test whether the volume between the two is different as well as if the most common words are different.