Análisis de Datos con Twitter

Javier Castro M.

Universidad De Chile

June 11, 2021

Objetivos

Motivación: El principio del proyecto es entender cómo los datos impactan en el resultado de nuestro modelo. O bien, probar que los posibles sesgos que presenten nuestros datos se traspasaran a los resultados.

Primera idea: Crear un generador de texto que entrene con tweets creados en Chile durante cierto tiempo.

Twint: Twitter Intelligence Tool

Desde el **github** de **twint**:

"Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API."

Típicamente el uso de `twint` sigue la siguiente estructura:

```
c = twint.Config()
c.Hide_output = True
c.Geo = "-33.45776872061534,-70.66448325142338,10km"
c.Since = '2019-10-16'
c.Until = '2019-10-22'
c.Store_csv = True
c.Output = 'data/santiago_estallido.csv'
twint.run.Search(c)
```

Figure: Uso típico de `twint`.

Exploración de los datos

Se obtuvieron varios set de datos, estos consisten en archivos `csv` tales que cada "fila" corresponde a un tweet. Se buscaron solamente tweets en Santiago:

1. Tweets durante el estallido social.
2. Tweets durante el 8 de marzo.
3. Tweets durante las votaciones pasadas.
4. Tweets que contienen la palabra "covid" durante el 2020.

Obs: Los tweets que contienen videos o imágenes vienen acompañados de un link que direcciona al mismo tweet.

Twitter en el estallido social

Figure: Wordcloud de twitter el 18 de octubre del 2019 con coordenadas en Beaucheff y un radio de $10km$.

Twitter el 8 de marzo

Figure: Wordcloud de twitter el 8 de marzo del 2020 con coordenadas en Beaucheff y un radio de $10km$.

Twitter en las votaciones pasadas

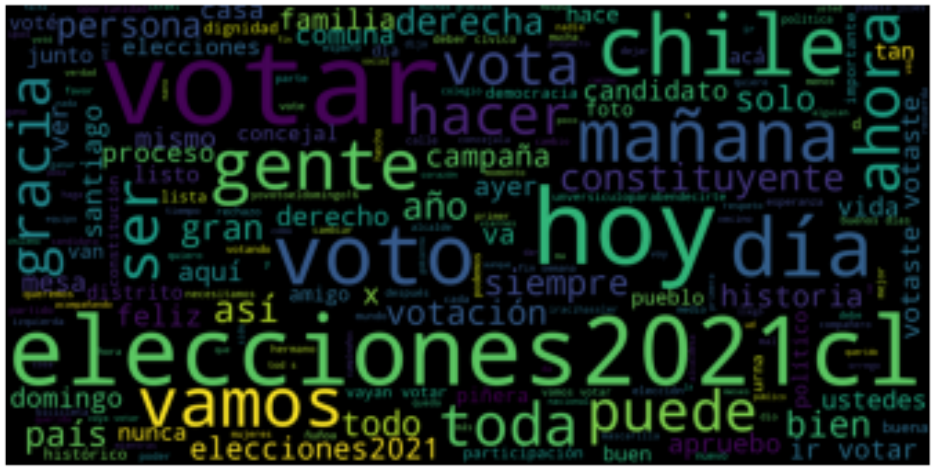

Figure: Wordcloud de twitter durante las votaciones del 15 y 16 de mayo del 2021 usando `.Near = "Santiago, Chile"`.

Tweets: "covid"

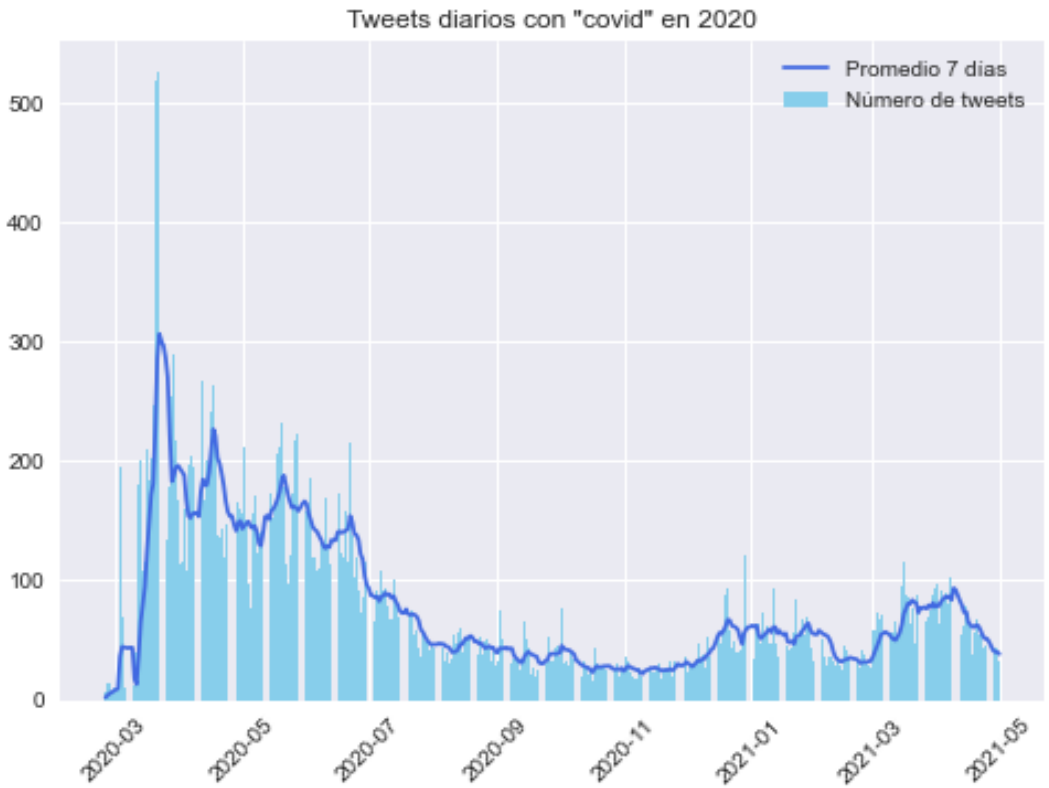

Figure: Cantidad de tweets diarios que contienen la palabra "covid".

Obs: El peak que se observa en el gráfico corresponde al 21 de marzo de 2020.

¿Con qué sigo?

1. Naive Bayes: para palabras c_i,

$$\mathbb{P}(c_{n+1}|c_1,...,c_n) \propto \mathbb{P}(c_{n+1}) \prod_{i=1}^{n} \mathbb{P}(c_i|c_{n+1})$$

2. Topic modelling, algo como **LDA** (*Latent Dirichlet Allocation*). Los tópicos se ven como distribuciones sobre palabras.
3. Generador de texto con redes neuronales tipo LSTM.
4. Otros?

Fin

Muchas Gracias!