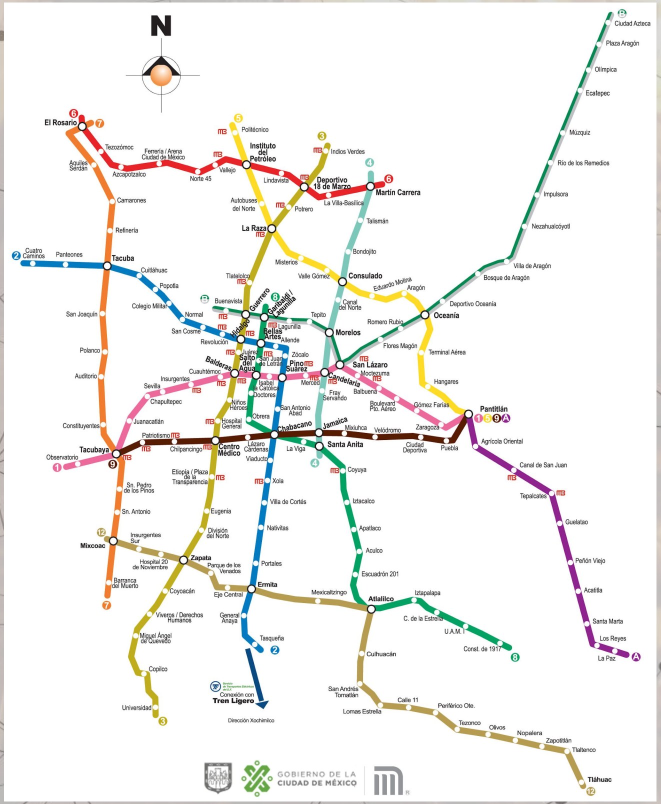
PANDATAS RELOADED



MEX-IT SMARTER!

A STORY OF COMMUTERS, TWEETS AND PACKED SUBWAY WAGONS

A BIT OF CONTEXT

The Mexico City subway system is one of the largest in the world (2nd largest in North America, after NYC).

- 12 lines in total
- Massive undertaking in lacustrine soil
- Celebrating 50 years of operations in 2019
    - Transports some 1.6 billion riders per year

OBVIOUS PROBLEMS...

And **FAR FROM PERFECT!**

NEED A RIDE?

MAIN OBJECTIVE

To better understand users' concerns and opinions through their comments on 🐦 ***Twitter.*** Although problems in the network are obvious, we want to discover in which direction passengers' opinions are geared towards.

With a thorough analysis of these messages, we can also determine the metro's most pressing problems and most problematic lines.'

ROADMAP

01 DATA EXTRACTION

CONNECTING TO TWITTER'S API AND PERMISSIONS

02 DATA CLEANING & STRUCTURING

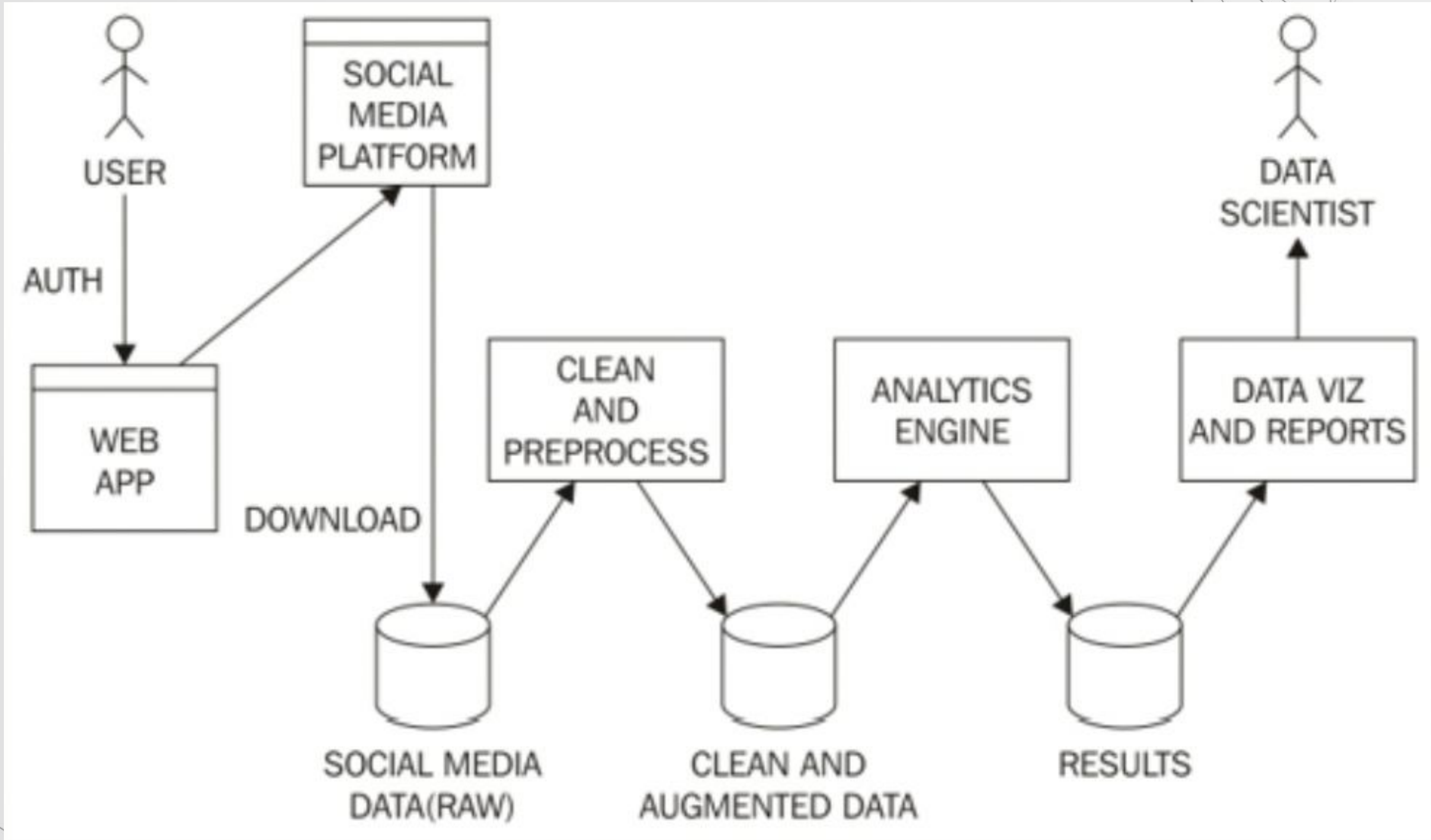
REMOVING SPECIAL CHARACTERS, LINKS, AND UNNECESSARY WORDS (STOP WORD)

03 TRAINING DATA | SENTIMENT ANALYSIS

WE CLASSIFIED THE DATA AND TRAINED OUR MODEL **TO DETERMINE 'POLARITY' OF TWEETS.**

04 VISUALIZATION

DETERMINE THE MOST RELEVANT INFORMATION TO DISPLAY IN A VISUAL INTERFACE


USER
SOCIAL MEDIA PLATFORM
DATA SCIENTIST
AUTH
WEB APP
DOWNLOAD
CLEAN AND PREPROCESS
ANALYTICS ENGINE
DATA VIZ AND REPORTS
SOCIAL MEDIA DATA(RAW)
CLEAN AND AUGMENTED DATA
RESULTS

HEAR THEM BIRDS SINGIN'

Use of Twitter's API

APPLICATION | APPROVAL

Developers need to specify use and motivation for access to retrieve large amounts of tweets.

API KEYS

We had to create an API key and user's access tokens.

APP BEARER TOKENS

This is an authentication and step and a more secure entry point for developers on Twitter.

ACCESS DESIRED API

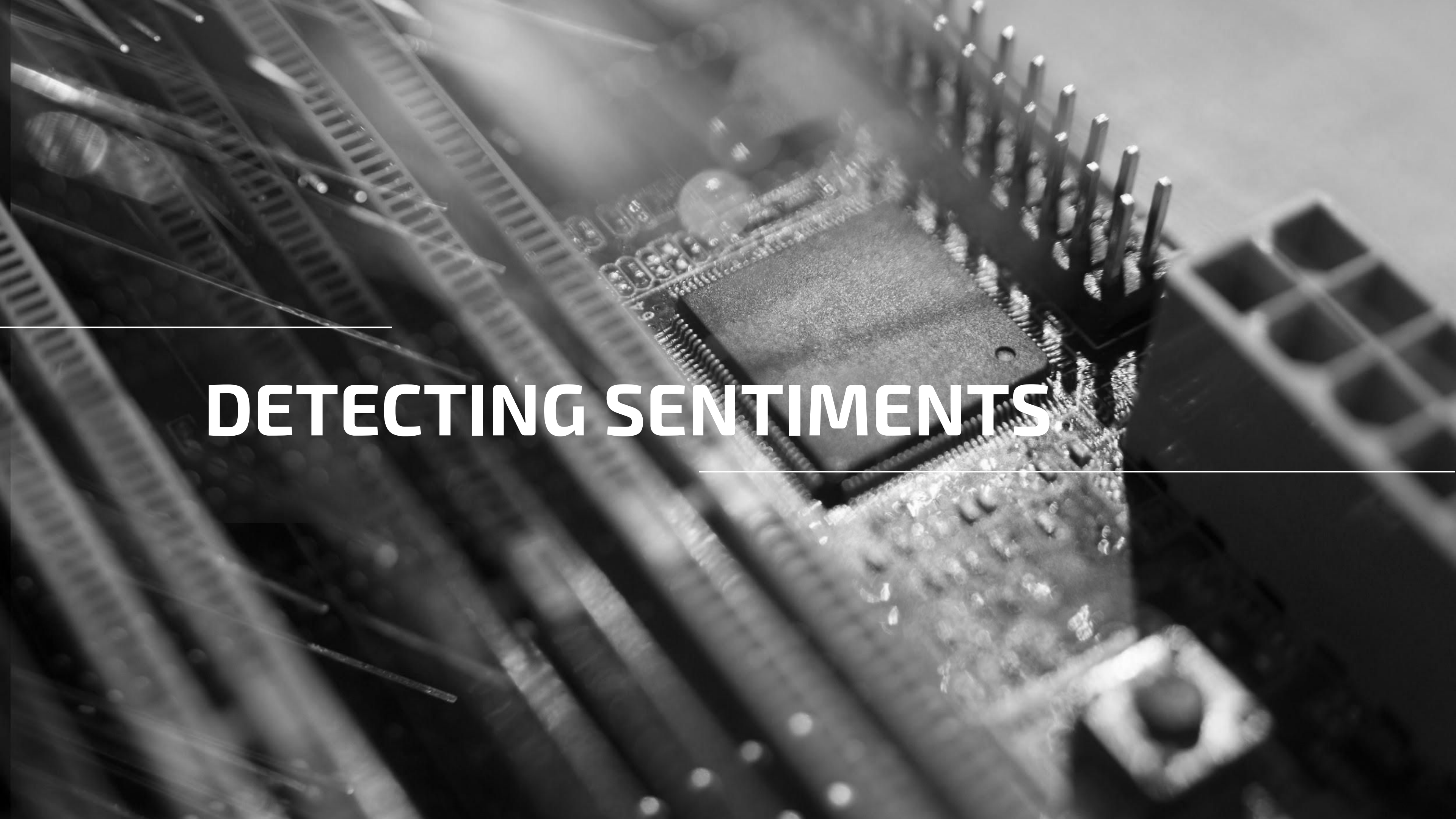
WE CHOSE THE "DEVELOPERS" API.

START REQUESTS

Our team tapped into tweets from the Mexico City area which mentioned key words related to the Mexico City Subway.

PROCESS DATA!

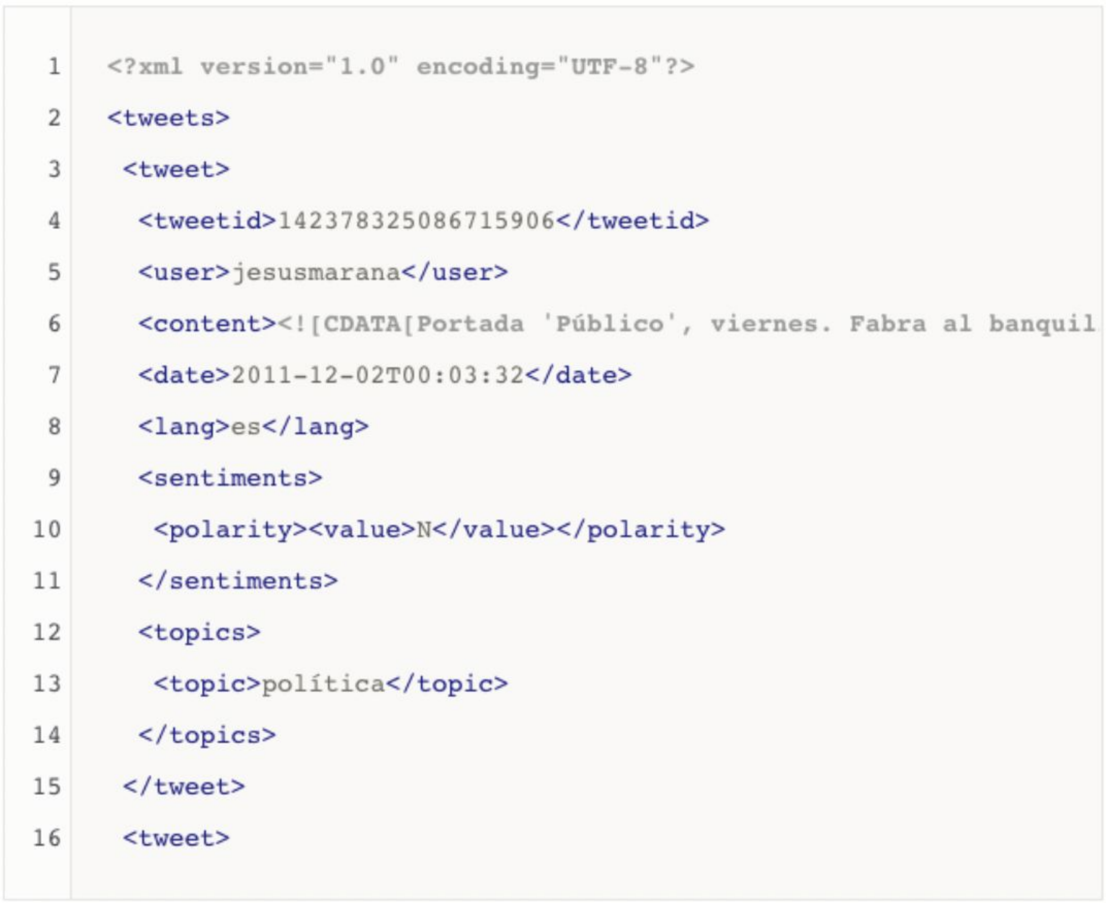
Once localized, our team downloaded all of these messages into a series of CSV files for analysis with **TWEEPY library in Python.**



DETECTING SENTIMENTS

TRAINING (IN SPANISH)

Gathered and parsed pre-tagged tweets from Sentiment Analysis in Spanish Workshop



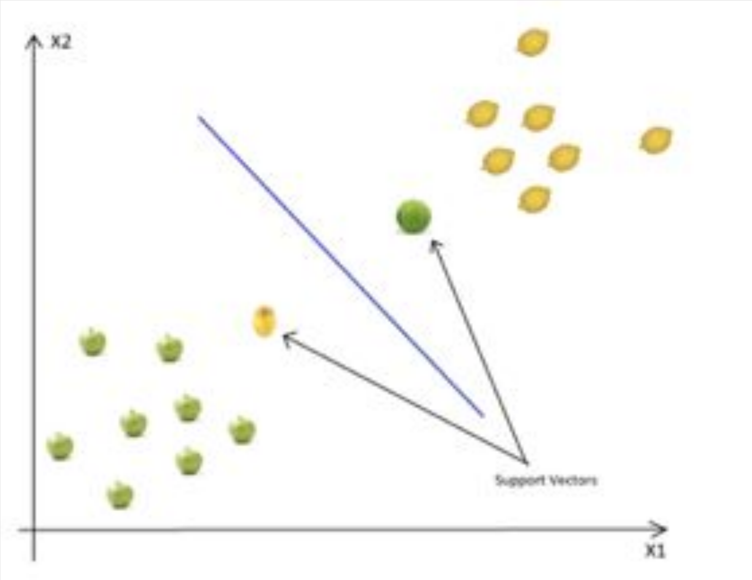
```xml
<?xml version="1.0" encoding="UTF-8"?>
<tweets>
 <tweet>
  <tweetid>142378325086715906</tweetid>
  <user>jesusmarana</user>
  <content><![CDATA[Portada 'Público', viernes. Fabra al banquil
  <date>2011-12-02T00:03:32</date>
  <lang>es</lang>
  <sentiments>
   <polarity><value>N</value></polarity>
  </sentiments>
  <topics>
   <topic>política</topic>
  </topics>
 </tweet>
 <tweet>
```

DETECTING SENTIMENT

1 BINARIZE LEVEL OF SENTIMENT
(P+, P, NEU, N, N+) TO 0 & 1

2 PROCESSING TEXT: VECTORIZE, TOKENIZE, STOPWORD-REMOVAL, AND STEMMED THE TEXT

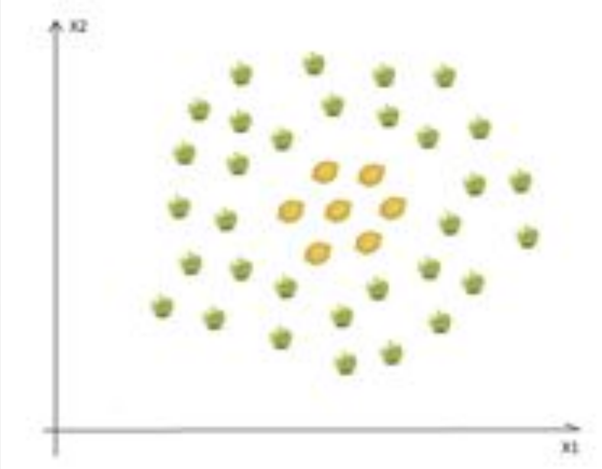
3 CHOOSING AN “ML"CLASSIFIER”: Linear Support Vector Classifier

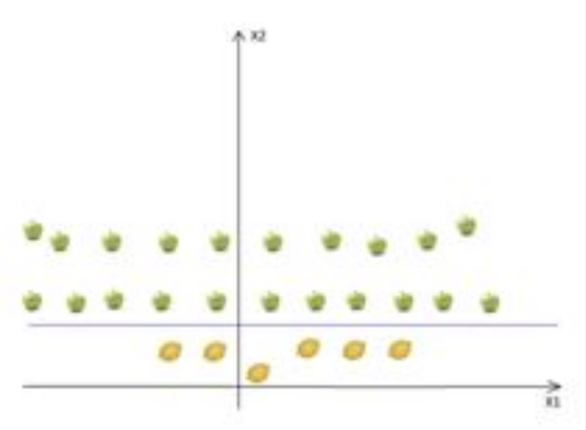


DETECTING SENTIMENT

4 TUNING PARAMETERS (W. *GridSearch CV*)

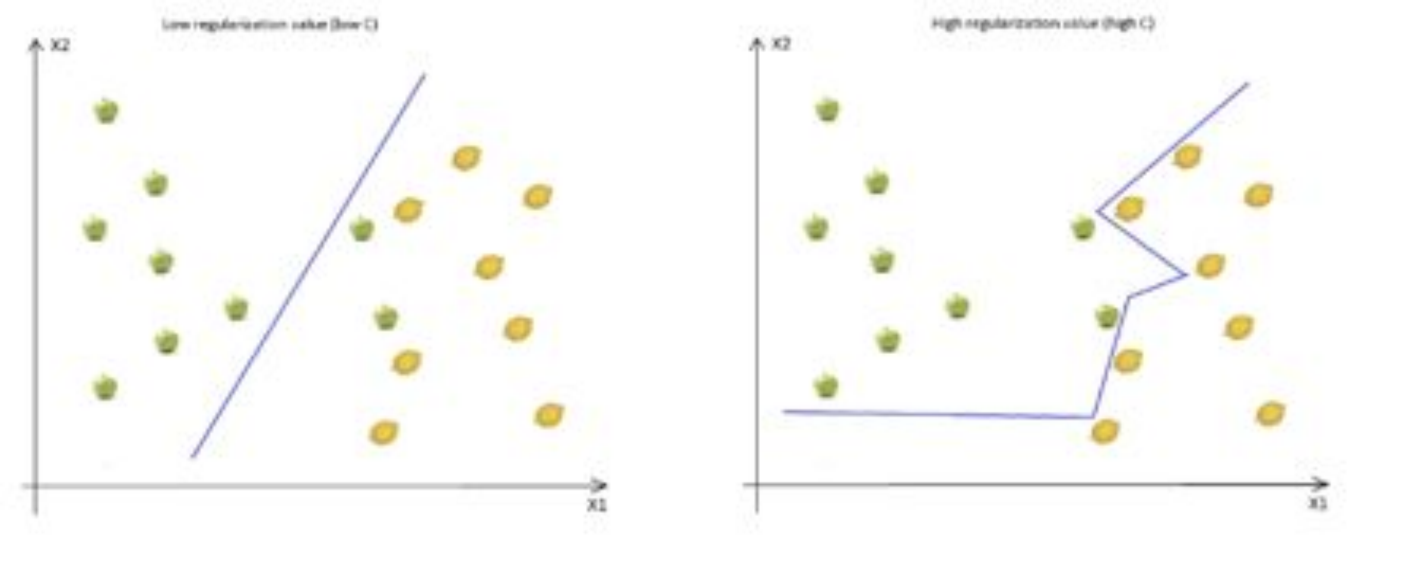
kernel: transforms the problem in order to be able to use linear algebra



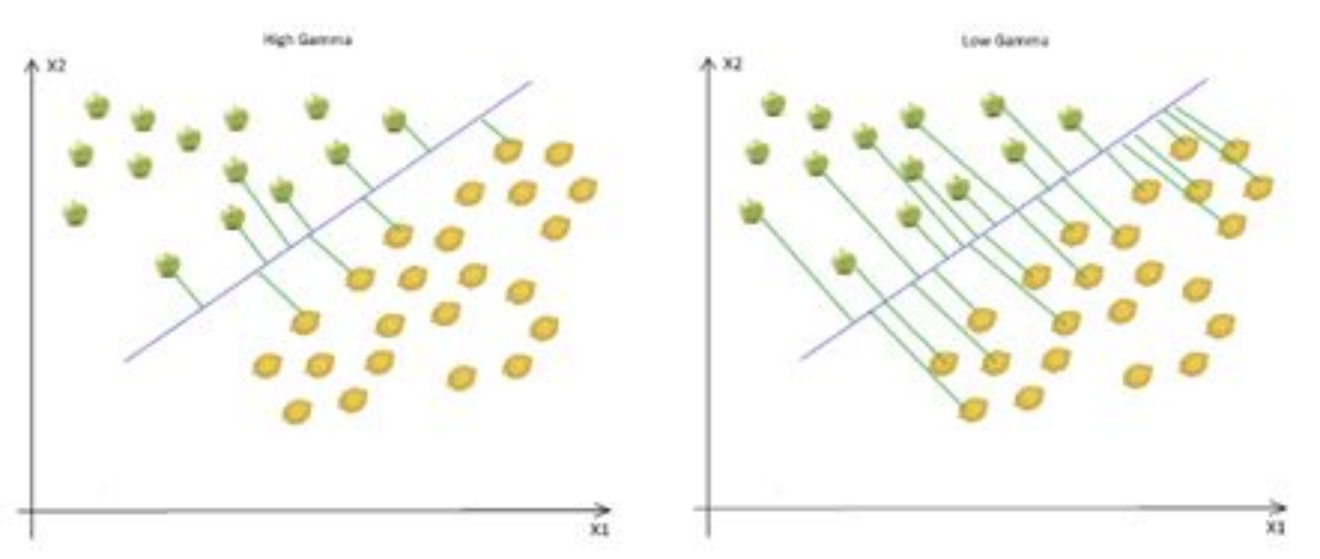


DETECTING SENTIMENT

Regularization: improve accuracy without overfitting

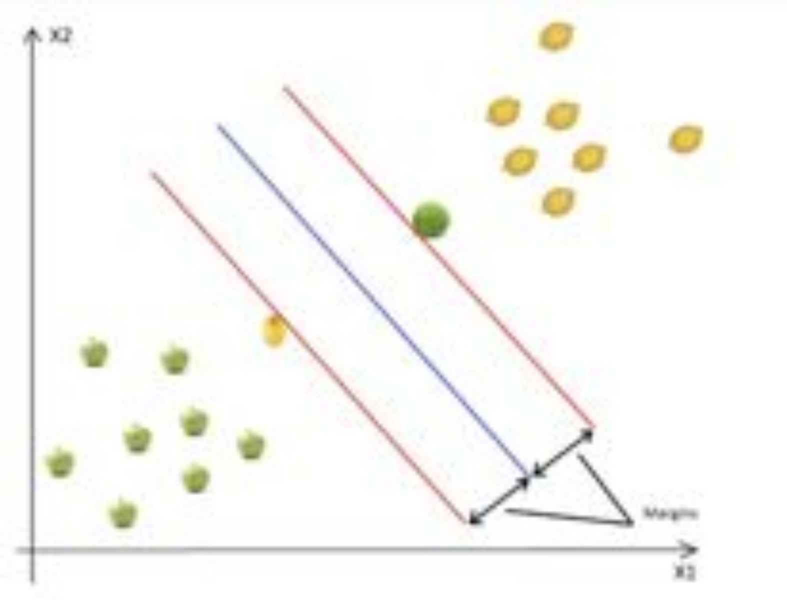

Low regularization value (low C)
X2
X1
High regularization value (high C)
X2
X1

DETECTING SENTIMENT



Gamma: defines how far are the farthest points considered

DETECTING SENTIMENT

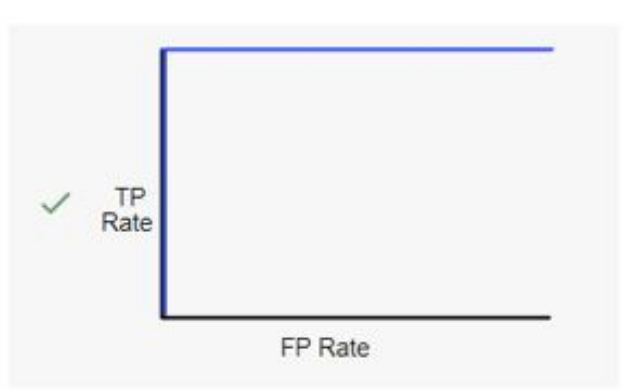


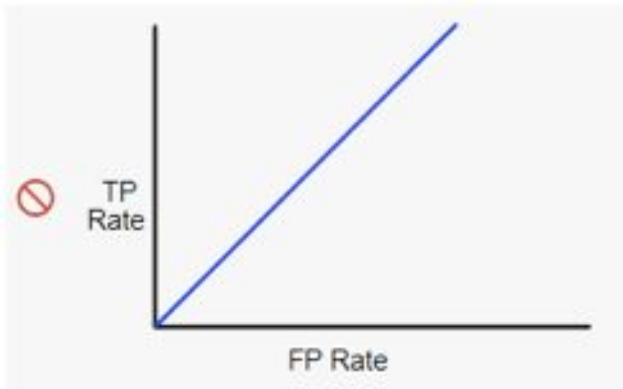
Margin: defines where the separation is the larger for both classes

DETECTING SENTIMENT

7 MODEL EVALUATION: AUC

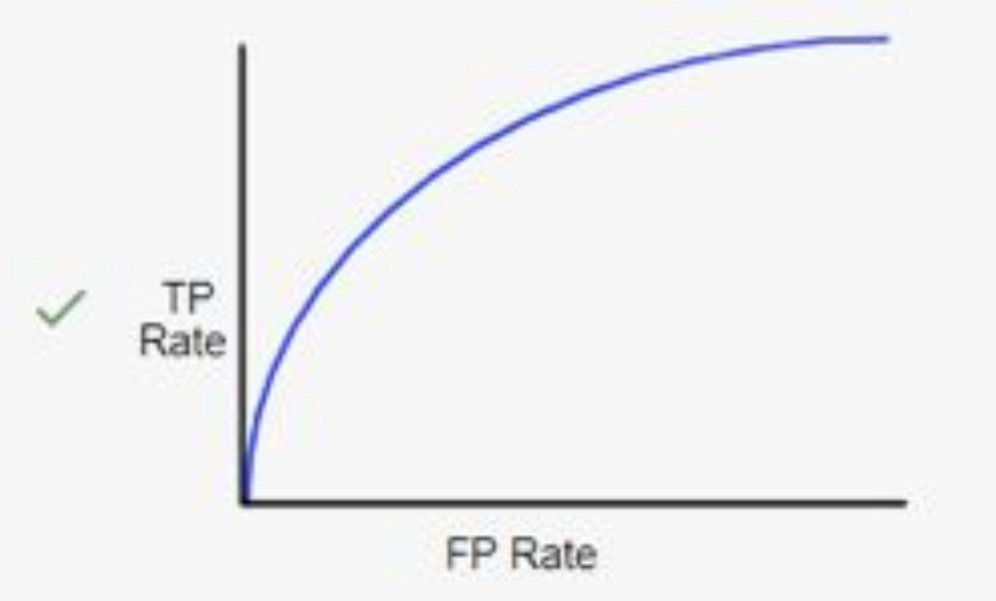
Best possible: AUC = 1 predicts true positives always



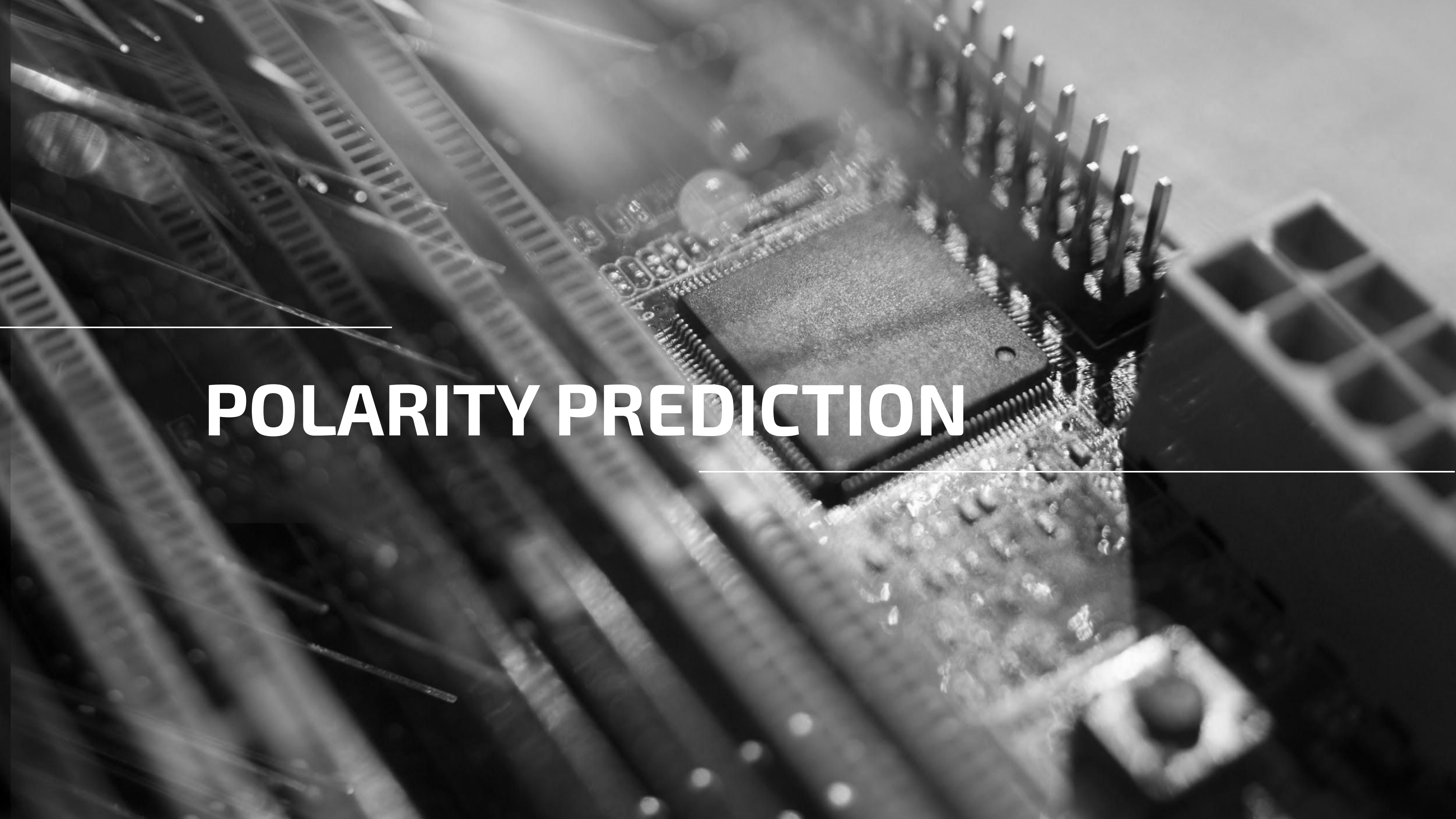


AUC = 0.5 predicts true positives only 50% of the times

OUR MODEL



AUC = 0.79 (our model) predicts true positives most of the times



POLARITY PREDICTION

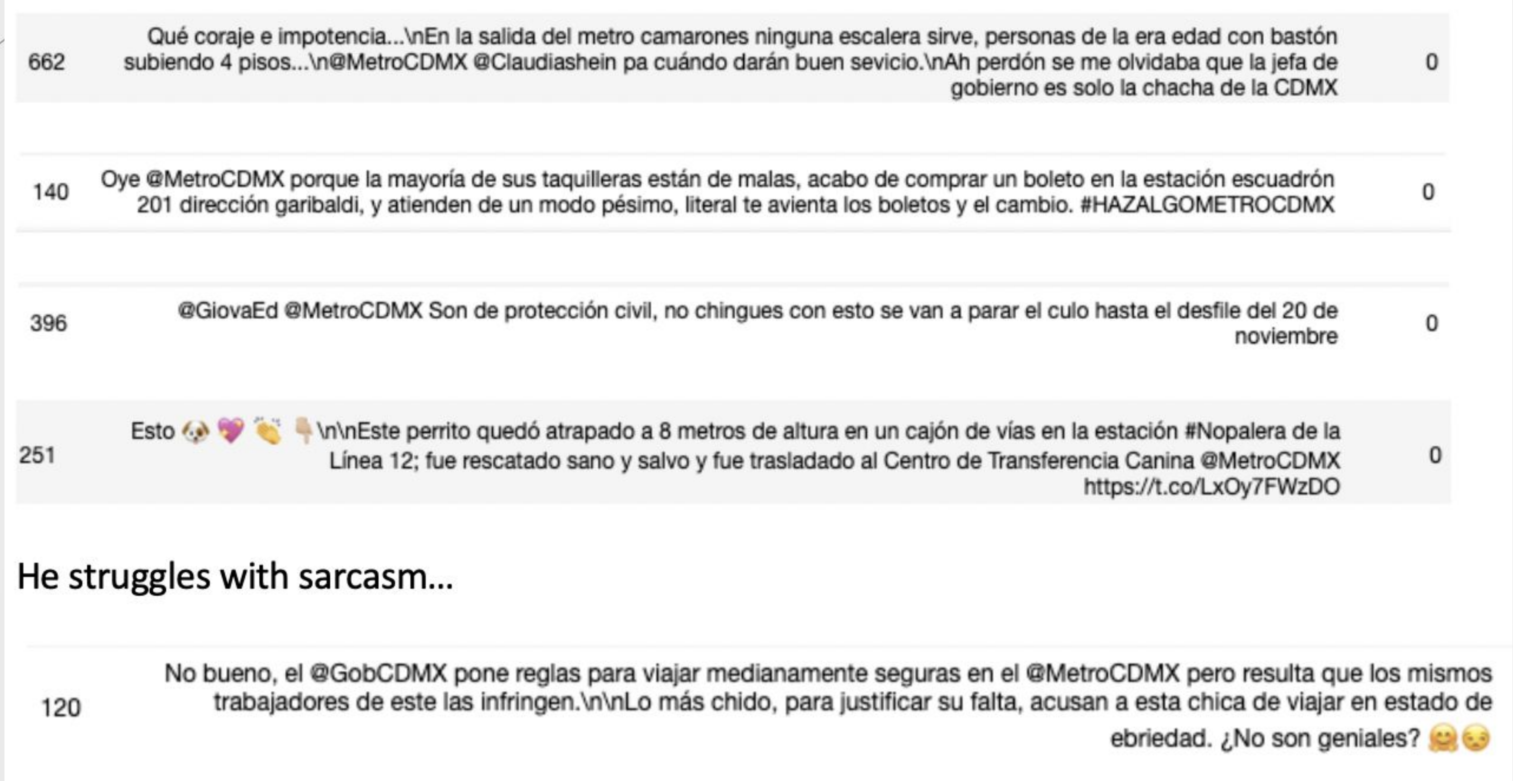
MEET SHELDON:

BAD WITH SARCASM

We decided to name our scikit-learn algorithm 'Sheldon', based on the Big Bang Theory's main character, since it cannot recognize sarcasm or irony in a sentence.

Most tweets mentioning CDMX's Subway System were positive given the algorithm's inability to interpret sarcastic messages

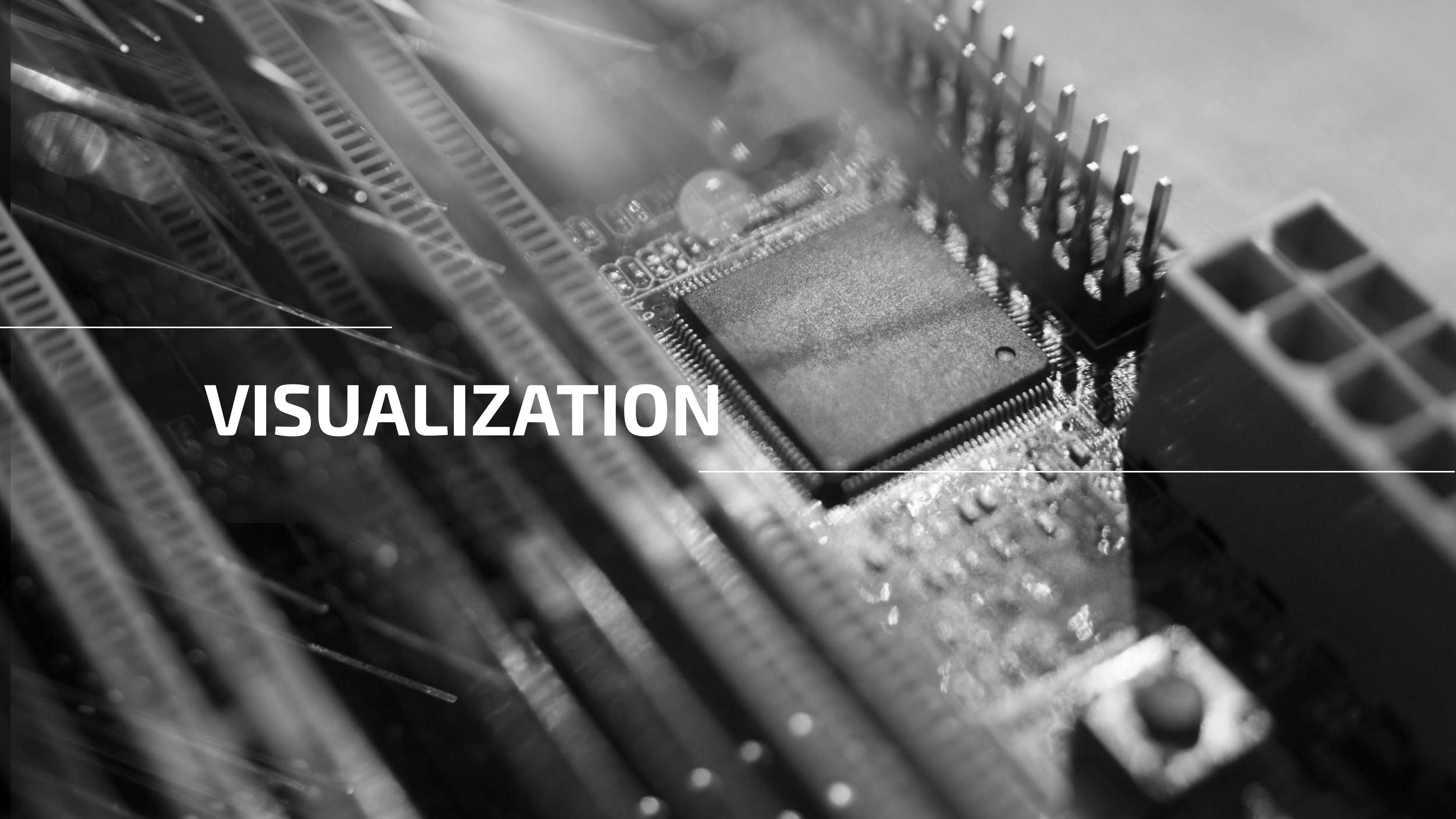
MEET SHELDON:



662	Qué coraje e impotencia...\nEn la salida del metro camarones ninguna escalera sirve, personas de la era edad con bastón subiendo 4 pisos...\n@MetroCDMX @Claudiashein pa cuándo darán buen sevicio.\nAh perdón se me olvidaba que la jefa de gobierno es solo la chacha de la CDMX	0
140	Oye @MetroCDMX porque la mayoría de sus taquilleras están de malas, acabo de comprar un boleto en la estación escuadrón 201 dirección garibaldi, y atienden de un modo pésimo, literal te avienta los boletos y el cambio. #HAZALGOMETROCDMX	0
396	@GiovaEd @MetroCDMX Son de protección civil, no chingues con esto se van a parar el culo hasta el desfile del 20 de noviembre	0
251	Esto 🐶 💖 👏 👇🏻\n\nEste perrito quedó atrapado a 8 metros de altura en un cajón de vías en la estación #Nopalera de la Línea 12; fue rescatado sano y salvo y fue trasladado al Centro de Transferencia Canina @MetroCDMX https://t.co/LxOy7FWzDO	0

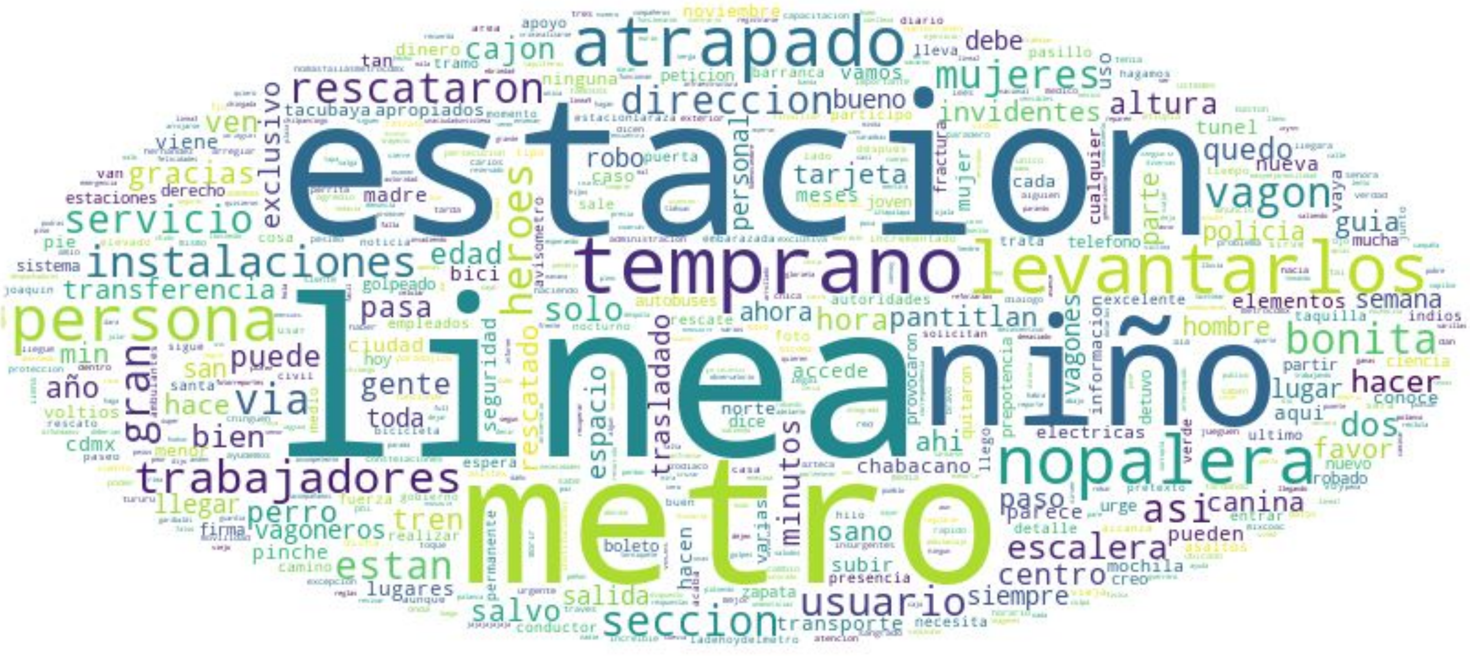
He struggles with sarcasm...

120	No bueno, el @GobCDMX pone reglas para viajar medianamente seguras en el @MetroCDMX pero resulta que los mismos trabajadores de este las infringen.\n\nLo más chido, para justificar su falta, acusan a esta chica de viajar en estado de ebriedad. ¿No son geniales? 🤗😧



VISUALIZATION

World Cloud: Metro CDMX

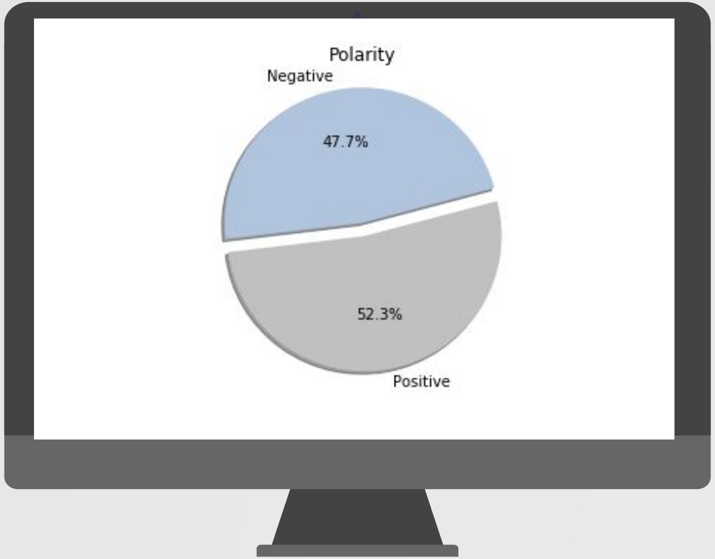


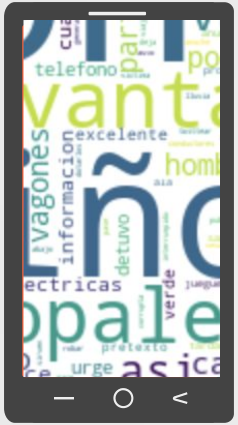
Tweet example: "Pues que los niños se levanten más temprano o también quiere que los vayamos a levantarlos?"

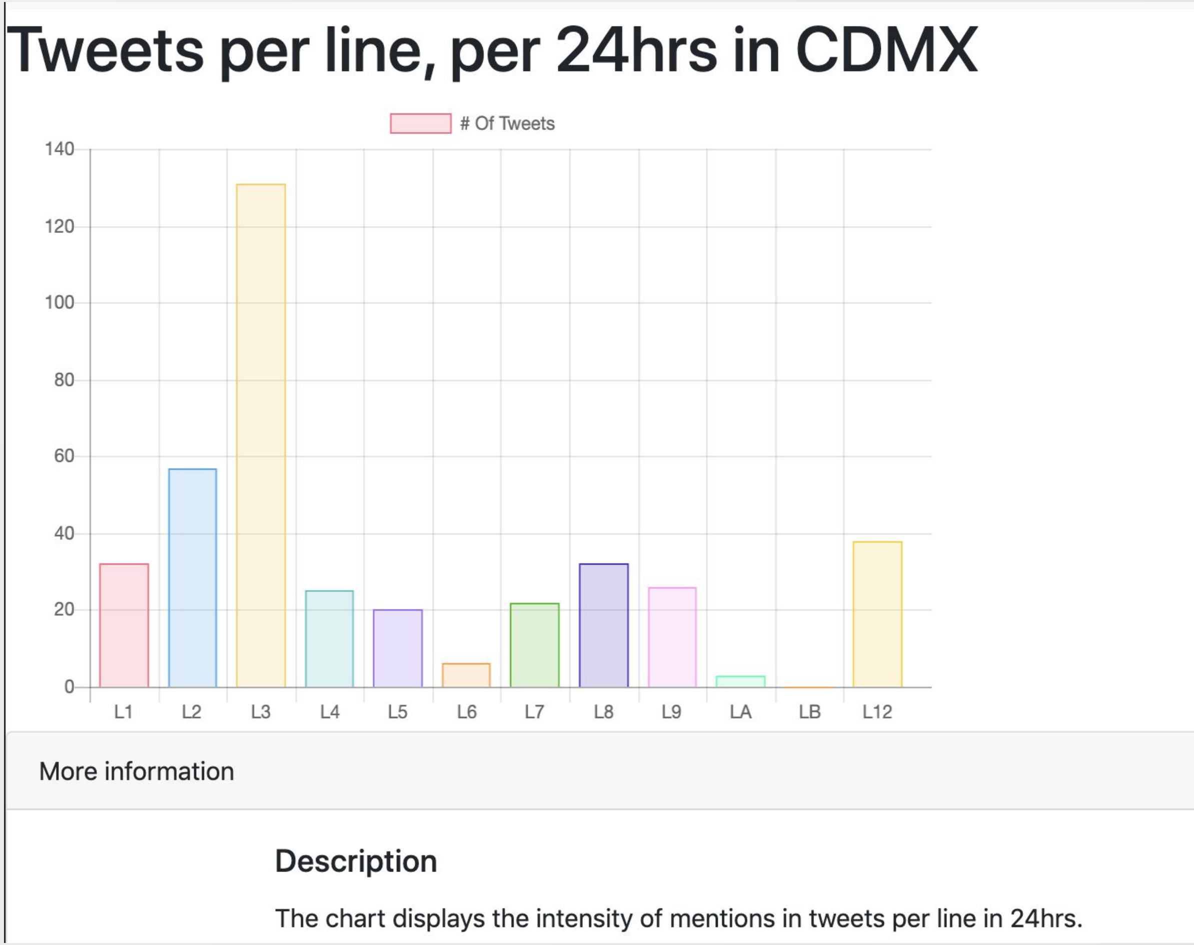
Python library: World Cloud

OVERALL SENTIMENT IN CDMX METRO SYSTEM 06-07 NOV 2019

est
lir
me
ciones
jadores
tacubaya
apropiados
madre
edad
bici
heroes
pasa
gente
toda
ciudad
hoy
seguridad
rescatado
espacio
tren
fuerza
realizar
golpeado
empleados
noticia
avisometro
espera
bicicleta







Tweets per line, per 24hrs in CDMX

More information

Description

The chart displays the intensity of mentions in tweets per line in 24hrs.

SOME CONSIDERATIONS:

1 TWITTER WON’T GIVE YOU ALL OF THE AVAILABLE DATA (GOTTA PAY FOR THAT!)

2 SEMANTICS IN MEXICAN SPANISH ARE PRETTY CONFUSING (I.E. ‘MADRE’)

3 SARCASM AND IRONY ARE A MAJOR ISSUE FOR MACHINE LEARNING (SHELDON)

FUTURE OF MEX-IT SMARTER:

1 STUDY ALL THE MOBILITY OPTIONS IN MEXICO (ECOBICI,METRO BUS, TRAFFIC JAM, MICROBUSES, ETC.)

2 APPLY THE RESULTS OF THE INFORMATION OBTAINED FOR *SMART CITIES* TO MAKE THESE MORE EFFICIENT.

3 TRAIN A SPECIFIC MODEL FOR MEXICAN SPANISH

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.