UNIVERSIDAD SANTO TOMÁS

PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

SECCIONAL TUNJA

VIGILADA MINEDUCACIÓN - SNIES 1732

Acreditación Institucional Internacional

OTORGADA POR EL IAC CINDA ACUERDO 55 DEL 8 DE MAYO–VIGENCIA 5 AÑOS

ACREDITACIÓN INSTITUCIONAL DE ALTA CALIDAD MULTICAMPUS

Vigencia por seis años

QS STARS RATED FOR EXCELLENCE

ISO 9001 icontec SC4289-1

CERTIFIED IQNet MANAGEMENT SYSTEM

UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
SECCIONAL TUNJA
VIGILADA MINEDUCACIÓN - SNIES 1732

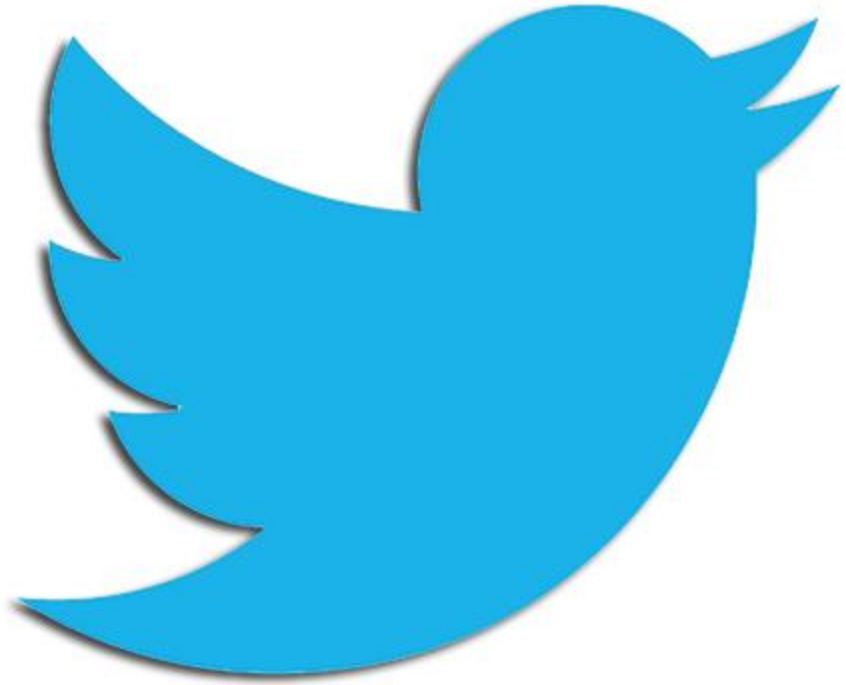
Faculty: Systems engineer
Course: Deep Learning
Topic: sentiment analysis

Professor: Luis Fernando Castellanos Guarin
Email: Luis.castellanosg@usantoto.edu.co
Phone: 3214582098

CONTENIDO

1. Configurar una cuenta de Twitter si aún no tienes una:
   - *Desde su cuenta de Twitter, solicitar el acceso de desarrollador*
   - *Crear una aplicación en Twitter que generará las credenciales de API*
   - *Configurar la aplicación*
2. Acceso a la API de Twitter desde Python:
   - Enviar un tweet
   - Buscar Tweets en Twitter de forma personalizada
   - Quitar/mantener retweets
   - Convertir los Tweets en un dataframe de panda
3. Entendiendo una opinión:
   - Tareas para analizar sentimientos
   - Niveles de análisis de sentimientos
4. Clasificación de textos/documentos
5. Fases de preparación del corpus.

En esta semana exploraremos el **análisis de los datos de redes sociales** a los que se accede desde **Twitter** mediante Python. Utilizaremos la API **RESTful de Twitter** para acceder a los datos sobre los usuarios de Twitter y sobre qué se tuitea.

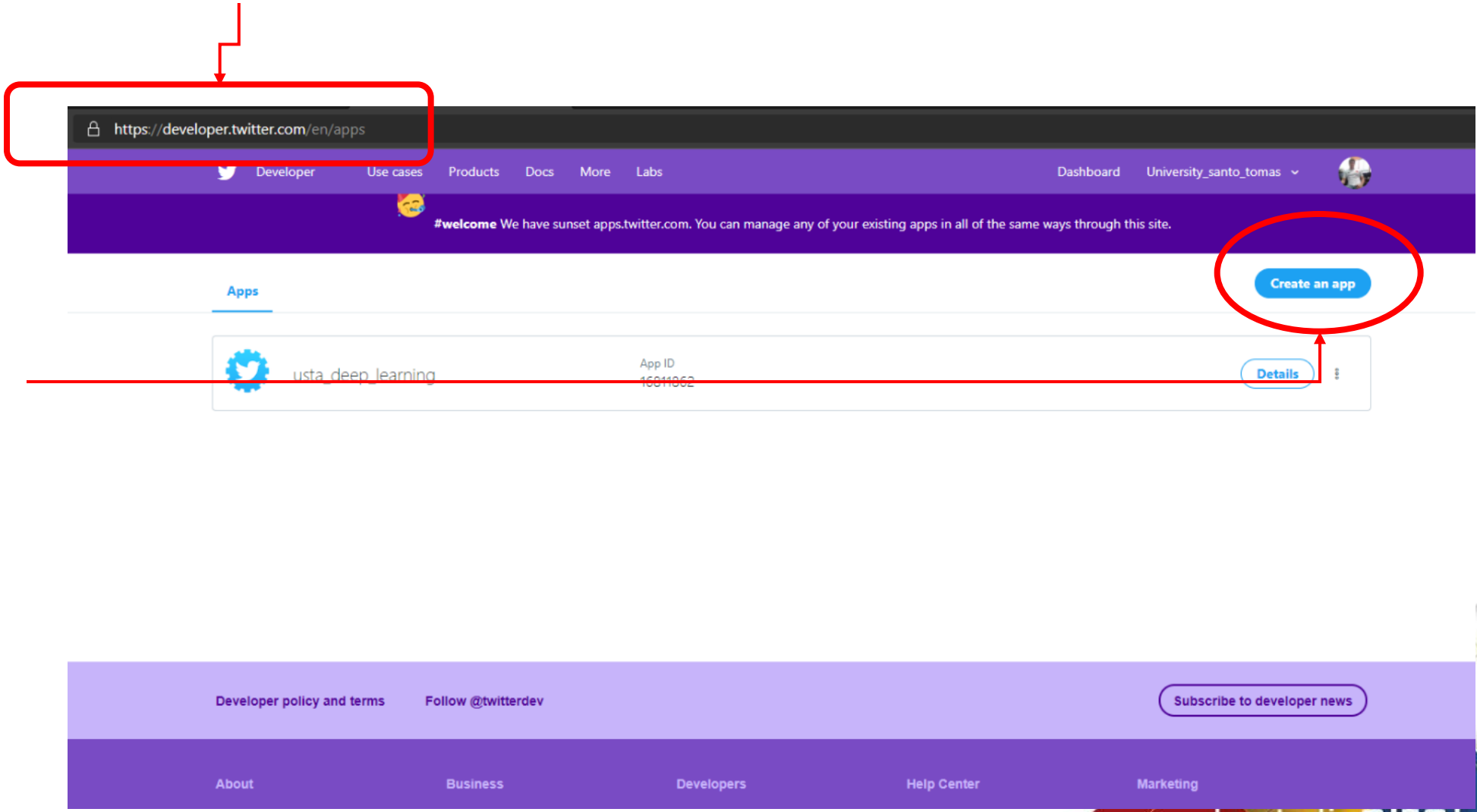


https://www.earthdatascience.org/courses/earth-analytics-python/using-apis-natural-language-processing-twitter/get-and-use-twitter-data-in-python/

1. Configura una cuenta de Twitter

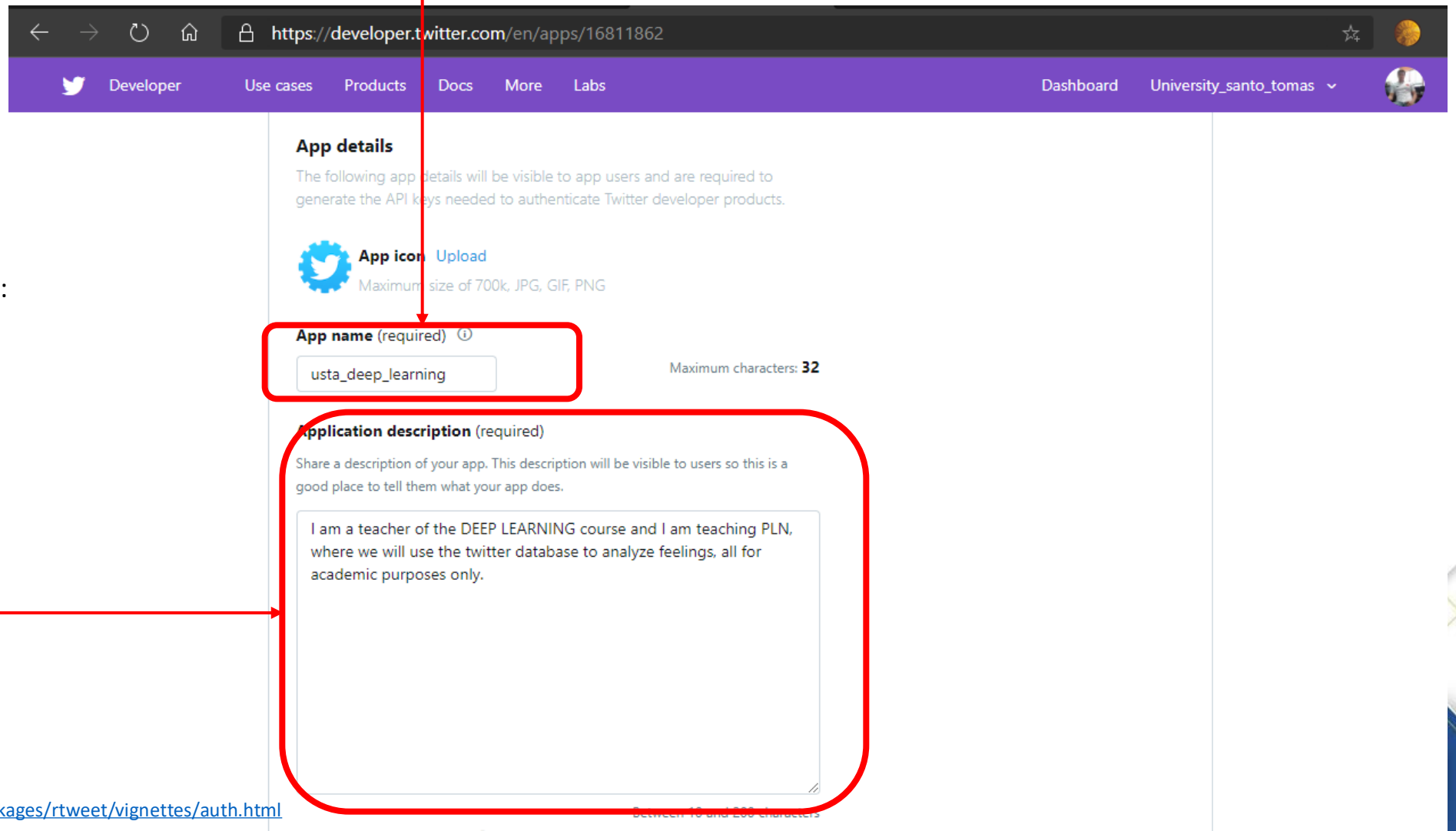
1. Ingresamos a la pagina: https://developer.twitter.com/en/apps

2. Crear una aplicación en Twitter que pueda usar para acceder a tweets.



1. Configura una cuenta de Twitter

3. App Name: Cómo se llamará su aplicación



4. Application Description: Cómo se describirá su aplicación a sus usuarios, es recomendable explicar que este aplicación es solo con fines academicos

https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html

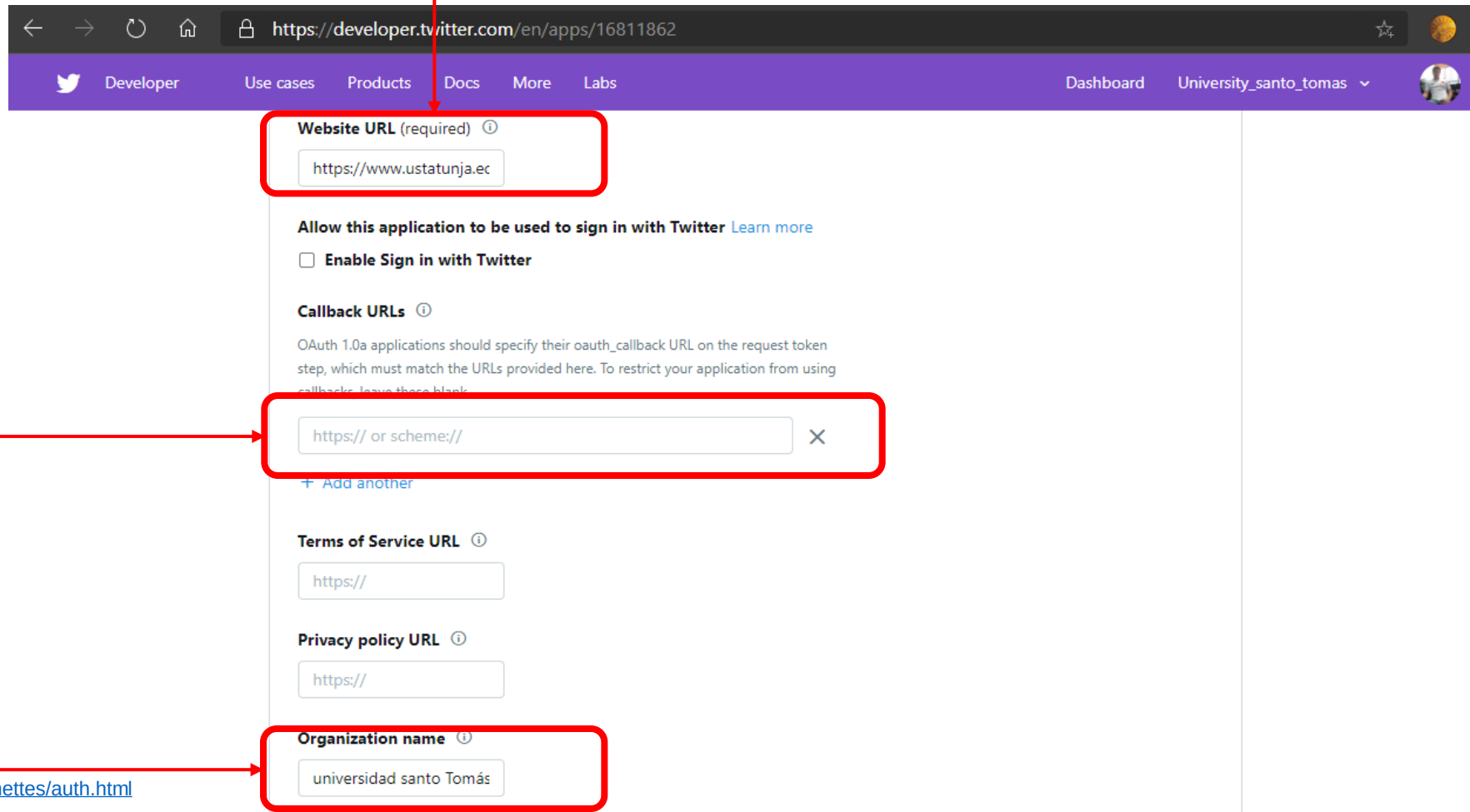
1. Configura una cuenta de Twitter

Website URLs: Sitio web asociado con la aplicación: recomiendo usar la URL de tu perfil de Twitter

Callback URLs: introduzca exactamente lo siguiente: http://127.0.0.1:1410 (no es vital se puede dejar vacío)

Organization name: colocar el nombre de la universidad, donde evidenciamos que es con fines académicos

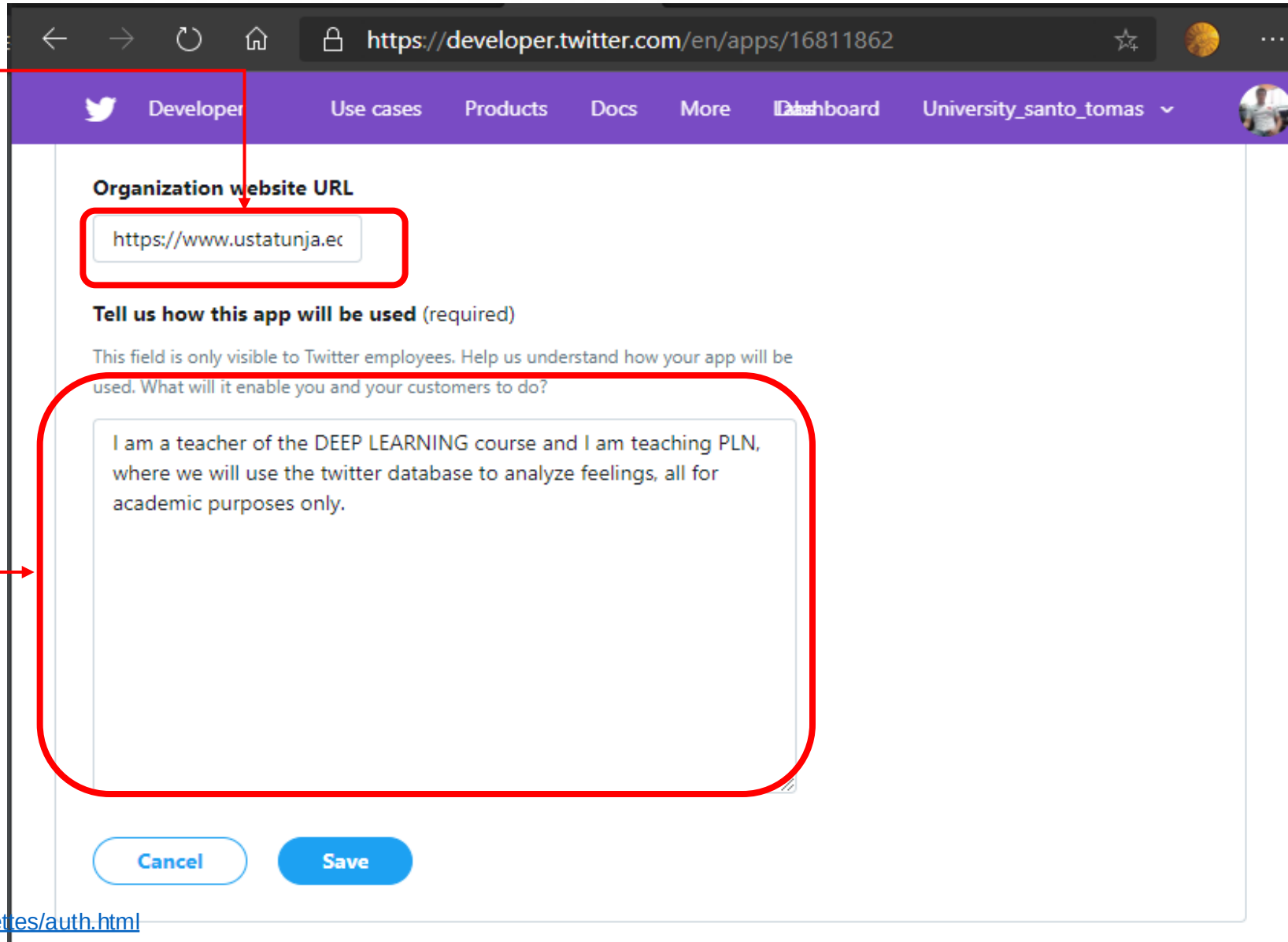
https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html



1. Configura una cuenta de Twitter

Organization Website URLs:
Sitio web de la USTA Tunja

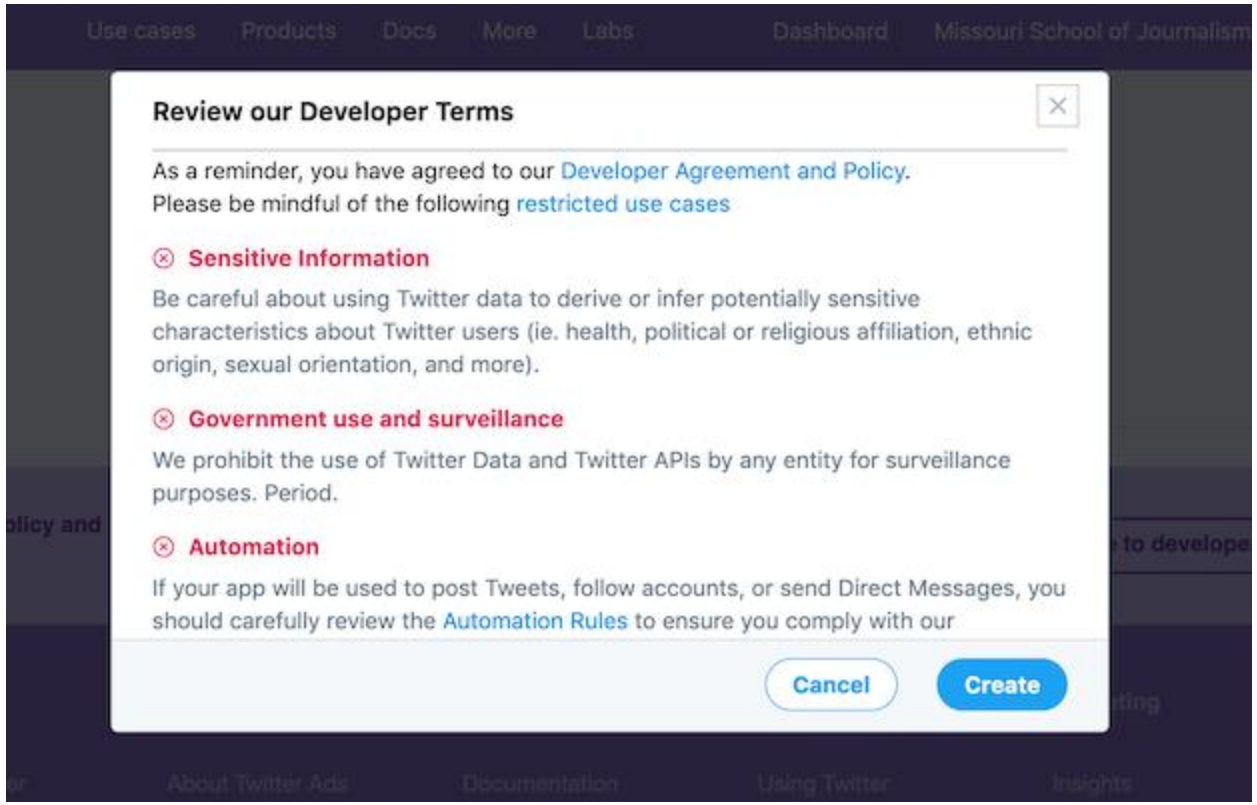
Tell us how this app will be used:
especificar que usted es un estudiante de ingeniería de sistemas y que la información que se va a obtener es para fines netamente académicos.



https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html

1. Configura una cuenta de Twitter

- Cuando haya completado los campos de formulario requeridos, haga clic en el botón azul en la parte inferior **Create**
- Lea e indique si acepta los términos del **desarrollador**



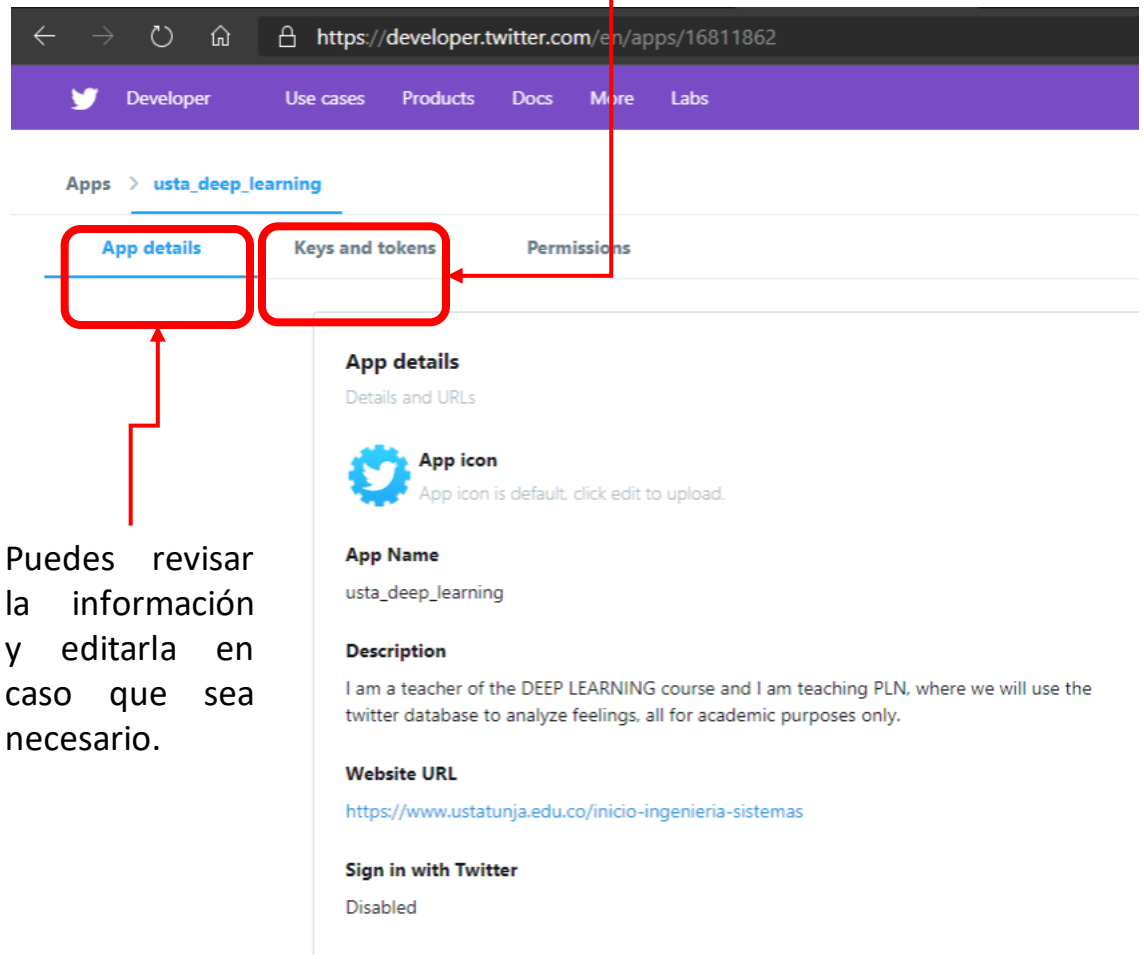
El proceso de aprobación puede durar varios minutos e incluso horas….tenga paciencia….

https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html

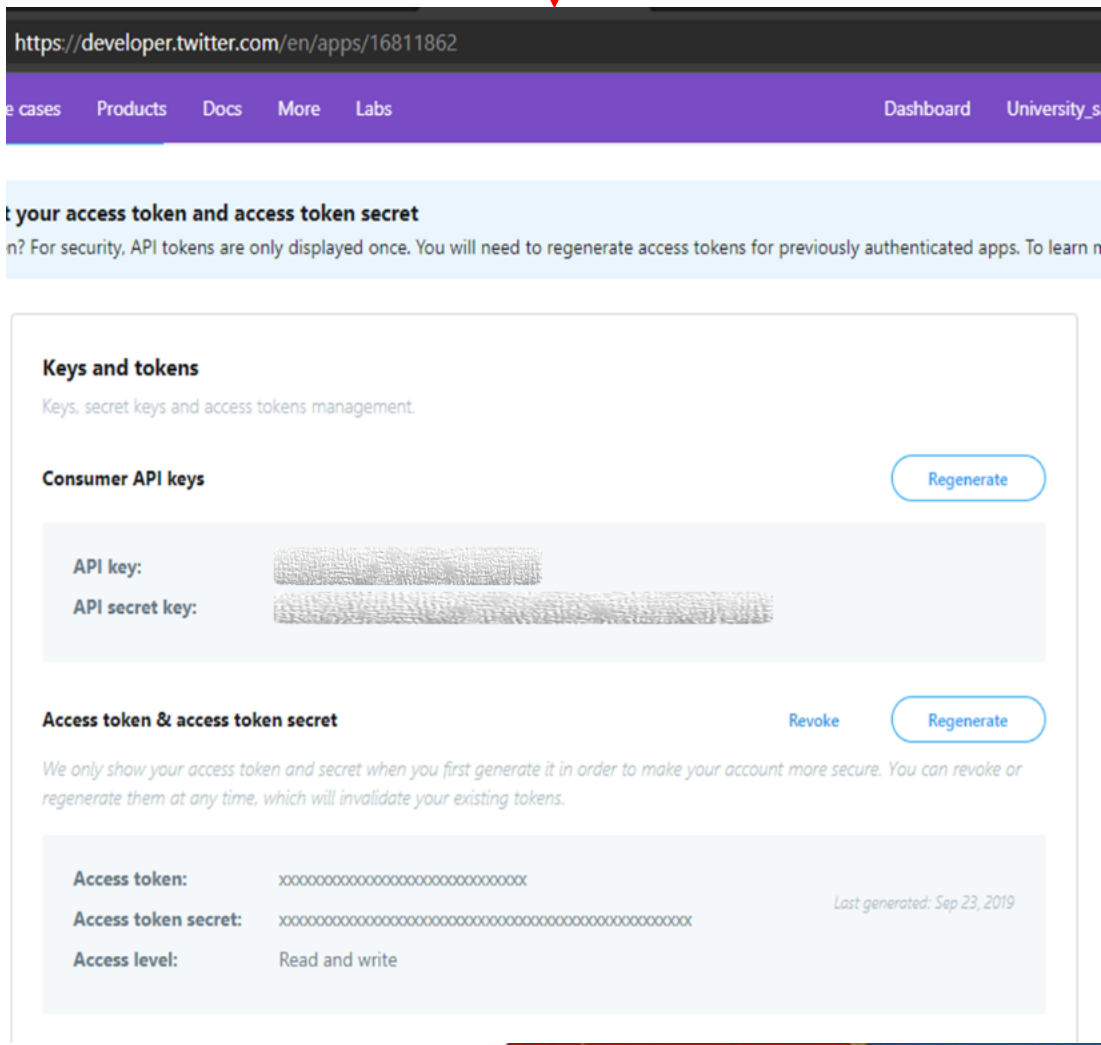
¡Siempre hacia lo alto!

1. Configura una cuenta de Twitter

De aquí tomaremos las credenciales para autenticarnos desde la api en PYTHON.



Puedes revisar la información y editarla en caso que sea necesario.



https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html

Una vez que hayamos hecho todo lo anterior.

¡estamos listo para comenzar a consumir la API de Twitter y obtener o generar tweets¡

Para lo siguiente debemos crear un cuaderno jupyter en Google colaboratory denominado **“feelings_on_twitter”**

2. Acceso a la API de Twitter desde Python

Librerías

Accediendo a la API de Twitter en Python

Una vez que se haya configurado la aplicación de Twitter, estará listo para acceder a los tweets Python. Comience importando las bibliotecas necesarias de Python.

```
import os
import tweepy as tw
import pandas as pd
```

Para acceder a la API de Twitter, necesitará 4 ***keys*** de la página de su aplicación de Twitter. Estas ***keys*** se encuentran en la configuración de la aplicación de Twitter en la ***Keys and Access Tokens***.

```
consumer_key= 'yourkeyhere'
consumer_secret= 'yourkeyhere'
access_token= 'yourkeyhere'
access_token_secret= 'yourkeyhere’

auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

2. Acceso a la API de Twitter desde Python

Enviar un tweet

Enviando Tweets a la API de Twitter desde Python

Puede enviar tweets mediante el acceso a la API. Ten en cuenta que tu tweet debe tener 280 caracteres o menos.

```
# Postear un  tweet desde Python

api.update_status(“#USTATUNJA, subiendo mi primer tweet desde PYTHON")
```

2. Acceso a la API de Twitter desde Python

Buscar tweets

Buscar Twitter para Tweets

¡Ahora estamos listos para buscar en Twitter tweets recientes! Comience por encontrar tweets recientes que usen el **#ClimateChange** hashtag. Utilizará el método .**Cursor** para obtener un objeto que contenga tweets que contengan el hashtag ***#ClimateChange***

```
# Definir el termino de la busqueda y la fecha de inicio
search_words = “#ClimateChange"
date_since = "2018-01-01"
```

Utilizando ***.Cursor()*** buscamos en Twitter tweets que contengan el término de búsqueda #ClimateChange. Puede restringir el número de tweets devueltos especificando un número en el .items()método. ***.items(5)*** y devolverá 5 de los tweets más recientes

```
# Collecional tweets
tweets = tw.Cursor(api.search,
           q=search_words,
           lang="en",
           since=date_since).items(5)
tweets
```

2. Acceso a la API de Twitter desde Python

Visualizando tweets

visualizando Tweets recolectados

```
[tweet.text for tweet in tweets]
```

Eliminando reTweets recolectados

Un retweet es cuando alguien comparte el tweet de otra persona. Es similar a compartir en Facebook.

```
new_search = search_words + " -filter:retweets"
new_search
```

Consulta completa:

```
# Definir el termino de la busqueda y la fecha de inicio
search_words = '#COVID19'
date_since = '2020-01-01'
#para que no tome los tweets que estar retweets
new_search = search_words + " -filter:retweets"
new_search

# Collecional tweets
tweets = tw.Cursor(api.search,
           q=new_search,
           lang="es",
           since=date_since).items(10000)
tweets
```

2. Acceso a la API de Twitter desde Python

tweets a Dataframe

Convertimos los tweets en una Dataframe

Una vez que tenga una lista de elementos con los que deseamos trabajar, podemos crear un marco de datos de pandas que contenga esos datos (Dataframe).

```
data_frame = [[tweet.user.screen_name, tweet.user.location,tweet.text] for tweet in tweets]

tw_dataframe = pd.DataFrame(data= data_frame , columns=["user","location","text"])
tw_dataframe
```

Guardamos la data en un csv.

```
tw_dataframe.to_csv('twitter_covid19_data.csv', index=False, encoding='utf-8')
```

Antes de continuar debemos hacer un pequeño paréntesis sobre

¿que es una opinión (tweet) y como analizarla?

¡Siempre hacia lo alto!

3. Entendiendo una opinión

Opinión de "**Andrea Cifuentes**" 2018-marzo-17:

"Hace tiempo que estaba pensando en cambiar de televisor **[1]**. Me decidí por este televisor Samsung debido a que es muy elegante **[2]**. La calidad de la imagen es increíblemente buena **[3]**. Además, el sonido es magnífico **[4]**. Un problema es que el sistema de navegación es un poco lento **[5]** pero viene traducido a diferentes idiomas, entre ellos el español **[6]**. Mis hijos están encantados con él **[7]**, pero mi marido piensa que es demasiado grande para nuestro salón" **[8]**.

Análisis:

- **[1]** la persona quería cambiar de televisor, pero no expresa ningún tipo de opinión
- **[2] [7]** algunas frases que le dan una apreciación al televisor Samsung en sí mismo.
- **[3][4][5][6][8],** Generan valores a determinados componentes como la calidad de la imagen, el sonido, el sistema de navegación y el tamaño.
- Se observa que hay opiniones tanto positivas como negativas
- También se observa que la opinión no sólo procede de la persona que escribe el comentario si no de sus hijos **[7]** y de su marido **[8].**

3. Entendiendo una opinión

A partir del ejemplo de la diapositiva anterior es posible deducir cuáles serían los componentes que formarían parte de la definición formal de una opinión.

Una opinión se define como una cuádrupla (**O, S, H, T**)

- **O**: el objeto de opinión
- **S** : el sentimiento, positivo o negativo
- **H**: la persona que expresa dicha opinión
- **T** : el momento en el que lo hace.

Infortunadamente con lo anterior no es posible representar determinados elementos que aparecen en el ejemplo, tales como el sonido, la imagen o el sistema de navegación.

¿Como haríamos si la opinión fuera con analogías, ironía, uso de jerarquías, comparativas o una suma de partes o atributos?

3. Entendiendo una opinión

Tareas para analizar sentimientos de una opinión:

1. **Extraer y categorizar entidades:** Debemos encontrar e identificar todas las entidades que contiene y agruparlas en base a su significado común. Cada uno de estos grupos representará a única entidad e_i.

1. **Extraer y categorizar aspectos**: en esta tarea se buscarán y capturarán los aspectos del texto teniendo en cuenta que pueden existir distintas formas de expresarlos. Una vez localizados, se deben extraer y agrupar para, a continuación, asociar cada grupo con su entidad correspondiente. Cada uno de estos grupos de la entidad e_i representa a un único aspecto a_{ij}.

1. **Tarea 3 - Extraer y categorizar a los autores de la opinión**: en este caso, la extracción se hará para cada autor o autores de las opiniones encontradas en todo el texto, tiendo en cuenta de nuevo que un mismo autor h_K puede ser representado en el texto de diferentes maneras.

1. **Extraer el momento temporal**: se trata de detectar el momento t_l en el que la opinión fue emitida.

1. **Clasificar la polaridad a nivel de aspecto**: para cada par de entidad *ei* y aspecto a_{ij} se debe determinar la valoración p_{ijkl} emitida por el autor de la opinión.

Generar la quíntupla de opinión: con todos los elementos identificados en los pasos anteriores, se crearán las quíntuplas que representen las distintas opiniones expresadas por sus autores.

3. Entendiendo una opinión

Tareas para analizar sentimientos de una opinión:

Tarea	Elemento 1	Elemento 2
Extraer y categorizar entidades	El televisor Samsung	el sistema de navegación del televisor
Extraer y categorizar aspectos	la calidad de imagen, el sonido y su tamaño	los idiomas
Extraer y categorizar a los autores de la opinión	Andrea Cifuentes, su marido y sus hijos	
Extraer el momento temporal	2018-marzo-17	
Clasificar la polaridad a nivel de aspecto	la calidad de imagen, el sonido y los idiomas del sistema de navegación tienen valoraciones positivas. valoración positiva del televisor en general por parte de los hijos.	el tamaño del televisor y la velocidad del sistema de navegación obtienen valoraciones negativas

3. Entendiendo una opinión

Tareas para analizar sentimientos de una opinión:

Tarea	Elemento
Generar la quíntupla de opinión	• (Televisor Samsung, calidad imagen, positivo, Andrea Cifuentes, 17 de marzo de 2018) • (Televisor Samsung, sonido, positivo, Andrea Cifuentes, 17 de marzo de 2018) • (Sistema de navegación, velocidad, negativo, Andrea Cifuentes, 17 de marzo de 2018) • (Sistema de navegación, idiomas, positivo, Andrea Cifuentes, 17 de marzo de 2018) • (Televisor Samsung, GENERAL, positivo, Hijos de Andrea Cifuentes, 17 de marzo de 2018) • (Televisor Samsung, tamaño, negativo, Marido de Andrea Cifuentes, 17 de marzo de 2018)

3. Entendiendo una opinión

Niveles de análisis de sentimientos

El análisis de sentimientos de un documento se puede llevar a cabo a tres niveles distintos en base a la granularidad, profundidad y detalle requeridos. Estos niveles son:

- **Análisis a nivel de documento**: en este nivel se analiza el sentimiento global de un documento como un todo indivisible, clasificándolo como positivo, negativo o neutro o usando otro sistema de calificación.

En estos casos, se asume que dicho documento expresa una valoración sobre una única entidad (por ejemplo, un servicio o producto) por lo que no es aplicable en aquellos que hablen sobre varias entidades simultáneamente

- **Análisis a nivel de oración**: en este caso, se divide el documento en oraciones individuales para extraer posteriormente la opinión que contiene cada una de ellas

La opinión de cada oración puede ser, de nuevo, positiva, negativa o neutra o bien tomar un valor en base a cualquier otro tipo de medida.

- **Análisis a nivel de aspecto y entidad**: este es el nivel de análisis con mayor detalle posible, en donde una entidad está formada por distintos elementos o aspectos y sobre cada uno de ellos se expresa una opinión cuya polaridad puede ser distinta en cada caso

Este nivel es el que se corresponde con la quíntupla presentada en el apartado anterior y el que mayor desafío presenta en la actualidad para los investigadores de la materia.

3. Entendiendo una opinión

Niveles de análisis de sentimientos (dificultades)

Algunas de las dificultades son directamente del PLN, pero otros pertenecen de manera exclusiva al campo del análisis de sentimientos, por ejemplo

Palabra	Sentido positivo	Sentido negativo
cabeza	tener la cabeza bien puesta	Perder la cabeza
morir	“Morir de la risa”	Morir por enfermedad
interminable	“Este amor es interminable”	“Que película tan interminable”
No	No hay duda de que es el mejor	la película no es buena
Sarcasmo /ironía	“Me muero, estoy enfermo por ti”	“¡Este televisor es genial! ¡Sólo me ha durado dos meses!”
expresiones coloquiales	“Esta como para chuparse los dedos”	“ese celular cuesta un ojo de la cara”
Ambigüedades	“Juan vio un niño con un telescopio en la ventana” ⍰ Quien tenia el telescopio juan o el niño?	
Emoticones	Si el texto tiene símbolos de carita feliz: :-) :-D ,-)	Si el texto tiene símbolos de carita triste **:-(** o de enojo **:-||**

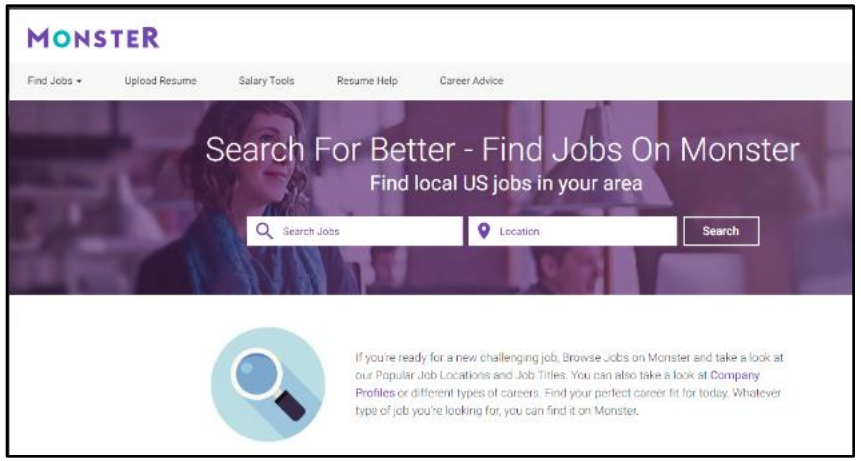
¡Siempre hacia lo alto!

4. Clasificación de textos/documentos

Ejemplos

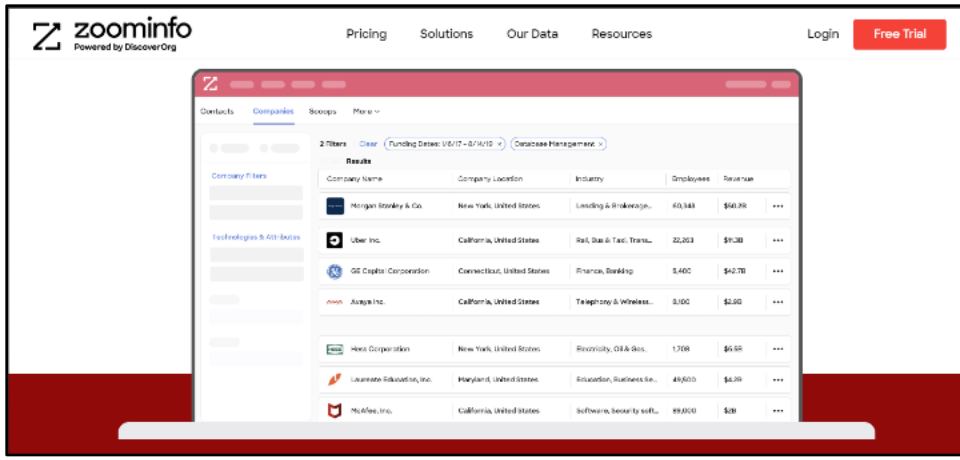
https://www.monster.com/jobs

Extrae ofertas de trabajo de unos 100.000 sitios web de empresas.



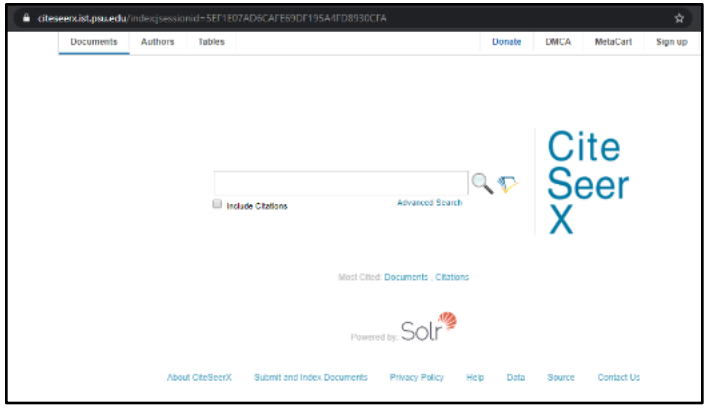
https://www.zoominfo.com/

Extrae información de personas y empresas de las páginas de todo el mundo (>300 millones).



Citeseerx.psu.edu

Extrae información de citas desde la cabecera de artículos científicos.
Puede identificar los artículos semilla en cada campo o calcular índices de impacto.



¡Siempre hacia lo alto!

4. Clasificación de textos/documentos

Para llevar a cabo la clasificación de un documento en base a su sentimiento existen diversos métodos y técnicas, las dos más populares y con mejor calificación en el mundo científico:

- **Clasificación usando aprendizaje no supervisado**, donde se "**trata**" de inferir la polaridad (positiva/negativa) del sentimiento global de un documento a partir de la orientación semántica de las palabras o frases que lo conforman y se divide en dos Métodos:
  - **Basados en diccionarios**(o lexicones, del inglés *lexicon*) hacen uso de listados de palabras y frases previamente etiquetadas con la polaridad de sentimiento que expresan y, en ocasiones, además con su intensidad o la fuerza de dicho sentimiento
  - **Basados en relaciones lingüísticas**, buscan ciertos patrones en los textos que puedan expresar opiniones y sentimientos con mayor probabilidad, extrayendo las palabras que lo forman para luego ser usadas en la categorización del texto global
- **Clasificación usando aprendizaje supervisado**, está basada en el uso de algoritmos de aprendizaje automático, conocidos también como *machine learning*

4. Clasificación de textos/documentos

Como esta materia se llama “Deep Learning” pues obviamente trabajaremos con el mejor:

Clasificación usando aprendizaje supervisado

Y usaremos:

- librerías específicas para este tipo de desarrollos como:
  - **NLTK v3.2.5 o superior:** facilita la realización de múltiples tareas para el procesamiento de lenguajes naturales.
  - **Scikit-Learn v0.19.1 o superior:** librería que implementa diversos algoritmos de aprendizaje automático y ofrece recursos para el análisis y la minería de datos.
  - **Stanford Part-Of-Speech Tagger v3.8 o superior**: utilidad para el análisis sintáctico de textos y que, entre otras cosas, permite clasificar las palabras en base a categoría gramatical.
  - **Pandas v0.22 o superior:** conjunto de utilidades para la manipulación y análisis de datos mediante lenguaje de programación Python

- **Algoritmos y modelos para trabajar en lenguaje Python….pero cual ?**

4. Clasificación de textos/documentos

Métodos de clasificación

Para hacer la clasificación de un documento/texto/tweet en base a su **sentimiento** existen diversos métodos y técnicas los mas usados son:

- Naive Bayes
https://www.youtube.com/watch?v=949tYJgRvRg
- K vecinos más cercanos
https://www.youtube.com/watch?v=FpAu0q2eSHo
- árboles de decisión
https://www.youtube.com/watch?v=gP2X8a3LaTM
- **máquinas de vectores de soporte**
https://www.youtube.com/watch?v=GOaIZqMh5PE
- **Regresión logística**
https://www.youtube.com/watch?v=KN167eUcvrs

Para el análisis de sentimientos el mejor método (por ahora) es maquina de vectores de soporte

pero muy costoso para entrenar así que usaremos regresión logística que es menos costoso y casi tan efectivo

4. Clasificación de textos/documentos

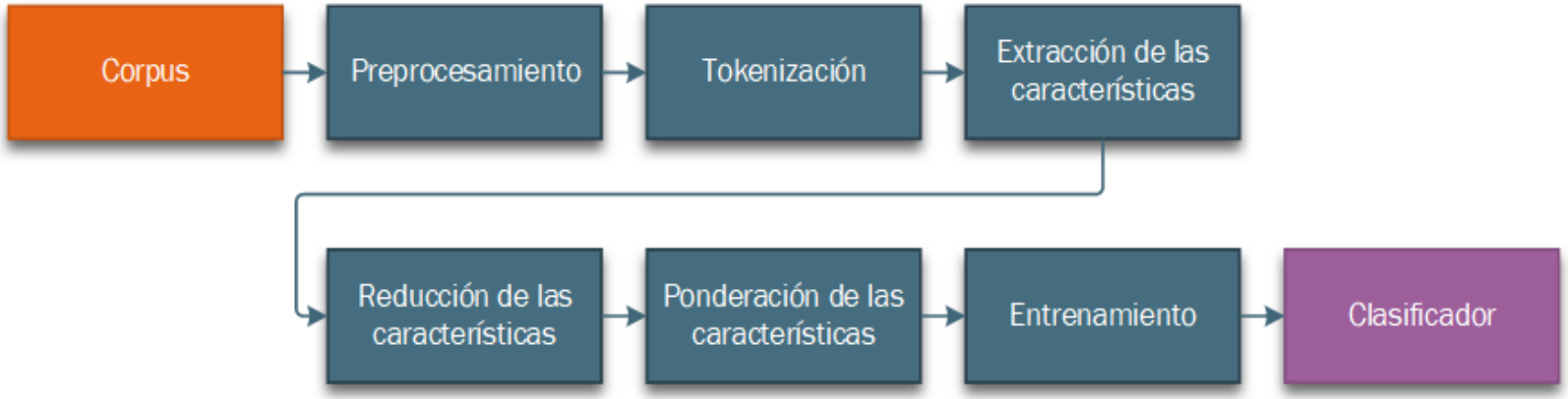
1. **CORPUS**: Obtener los datos del corpus (descargar los tweets) para entrenar los algoritmos.

1. **PREPROCESAMIENTO**: Se limpian y normalizan (mayúsculas, errores ortográficos, etc) los mensajes (tweets), con el objetivo de reducir o eliminar aquellos datos que puedan influir de manera negativa en el resultado final.

1. **TOKENIZACIÓN**: Donde el mensaje (tweet) se divide en unidades más pequeñas o tokens y que habitualmente son las palabras.

1. **CARACTERIZACIÓN**: A partir de los tokens se extraen las características que representen a los mensajes originales y, de manera opcional, se puede aplicar de un método para reducir su número.

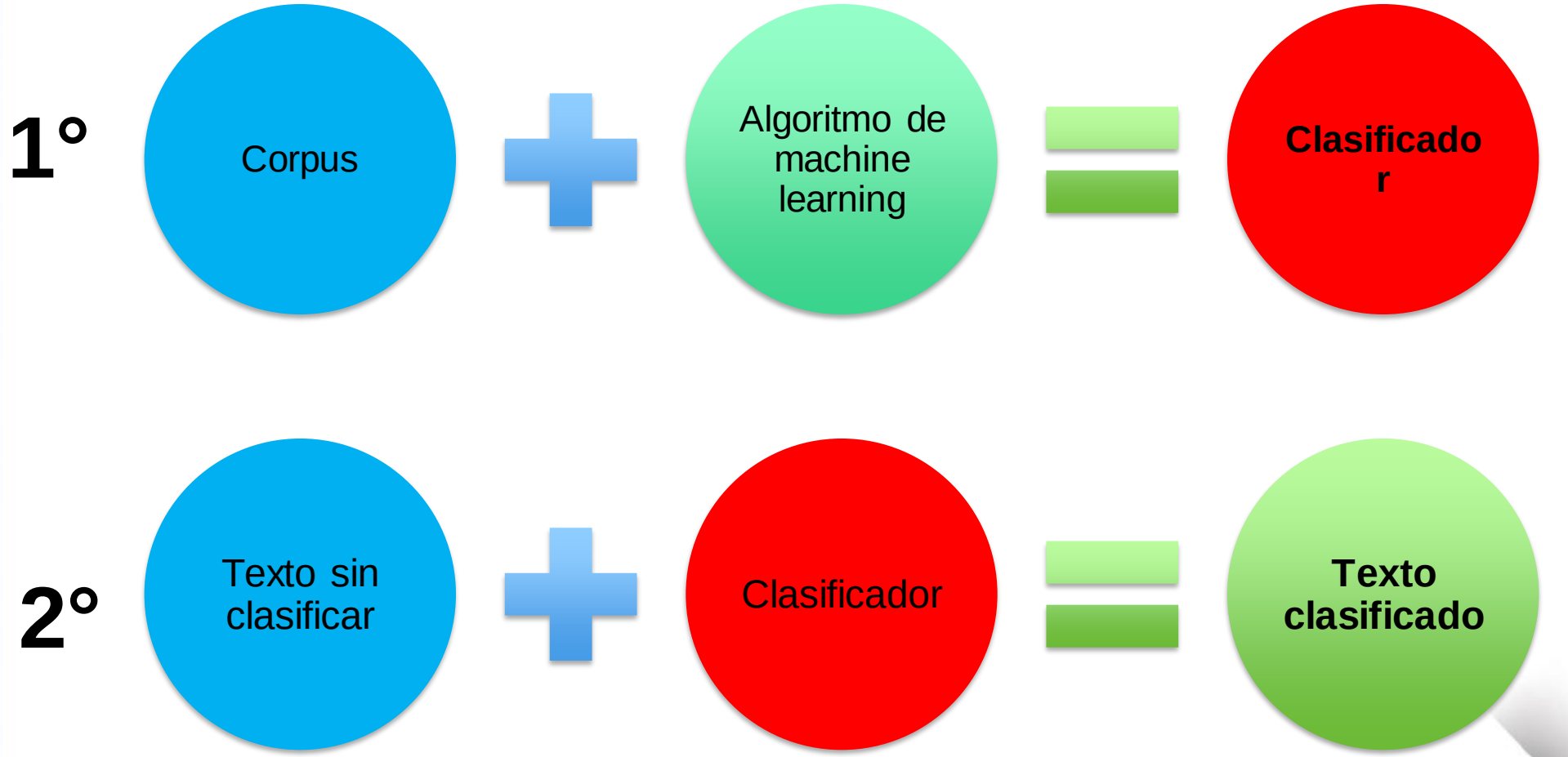
1. Para finalizar, estas características se ponderan en función de la importancia que se les quiera dar y con ellas se entrenan los clasificadores



corpus: Conjunto de textos debidamente etiquetados con su respectivo sentimiento y que sea representativo, aleatorio y equilibrado según el contexto.

4. Clasificación de textos/documentos

Macro proceso



¡Siempre hacia lo alto!

4. Clasificación de textos/documentos

Corpus

Uno de los problemas de los métodos supervisados es la necesidad de contar con un juego de pruebas representativo y previamente etiquetado para entrenar los algoritmos de aprendizaje automático, es decir, un corpus:

Desde el Taller de Análisis de Sentimientos (TASS) que pertenece a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) existen publicadas tres corpus que son los de más uso para nuestro idioma "El español" para analizar tweets:

- **General Corpus**, que incluye 68000 tweets escritos en español por 150 personajes y celebridades conocidas dentro del mundo de la política, economía, comunicación y cultura y que fueron obtenidos entre noviembre de 2011 y marzo de 2012.
-
- **Politics Corpus**, que ofrece 2500 tweets extraídos durante la campaña de Elecciones a las Cortes Generales de España de 2011. Estos mensajes mencionan a los cuatro partidos políticos más relevantes de aquel momento: PP, PSOE, IU y UPyD.
-
- **International TASS Corpus (InterTASS):** contiene 3400 mensajes de Twitter escritos en español y sobre cualquier tipo de tema.

Los mensajes del corpus se encuentran clasificados en cuatro y seis categorías de sentimiento: Muy positivo (P+), Positivo (P), Neutro (NEU), Negativo (N), Muy negativo (N+) y Sin sentimiento (NONE).

Pero para optimizar los análisis de sentimientos reduciremos a cuatro categorías:

Positivo (P), Neutro (NEU), Negativo y Sin sentimiento (NONE).

¿Cual es la diferencia entre un texto Neutro (NEU) y uno Sin sentimiento (NONE)?

4. Clasificación de textos/documentos

Corpus

Diferencia entre mensajes sin sentimiento (NONE) y mensajes neutros (NEU)

sin sentimiento (NONE)	mensajes neutros (NEU)
Los primeros son precisamente eso, tweets en los que no se expresa ninguna idea positiva ni negativa	Poseen un sentimiento "*a medio camino*" entre lo positivo y lo negativo y éste puede ser debido a dos razones: que las palabras usadas sean realmente neutras (AGREEMENT) o bien que contengan palabras tanto positivas como negativas en el mismo mensaje (DISAGREEMENT)
"*Los retos urgentes de las entidades locales frente al coronavirus* https://www.ins.gov.co/"	Ivan Duque: ''*Intentaremos repartir de manera equitativa los costos de esta crisis económica. La primera obligación de un gobernante es ser justo*"
	"Soy y seré del grupo parlamentario durante la legislatura pero discrepo de la exclusión y el reparto. Por eso seré un senador leal pero trabajador."

4. Clasificación de textos/documentos

Corpus

La siguiente tabla muestra el número de tweets de cada clase y para cada una de las colecciones a las que pertenecen

	Positivo (P)	Negativo (N)	Neutro (NEU)	Sin sentimiento (NONE)
General Corpus	25117	18026	1975	22899
Politics Corpus	613	681	933	221
International TASS Corpus	1116	1404	418	475
TOTAL (%)	**26846 (36%)**	**20111 (27%)**	**3326 (5%)**	**23595 (32%)**

73.878 tweets

5. Fases de preparación del corpus

Preproceso

Tokenización

Extracción de características

Reducción de características

Ponderación

Creación del clasificador de línea base.

¡Siempre hacia lo alto!

5.1. Preprocesamiento

Estas son las reglas que vamos a utilizar:

- **Normalización de mayúsculas y minúsculas (todo a minúsculas)**
- **Tratamiento de la duplicidad de caracteres, ejemplo:** “Qué aburrimiento tengo” o “Qué aburrimientooooooo tengooooooooo”.
- **Eliminación de tildes:** en las redes sociales los usuarios no acostumbran a hacer un buen uso de las tildes.
- **Eliminación de números**
- **Eliminación de *retweets***
- **Eliminación de retornos de carro:** algunos mensajes de Twitter contienen saltos de línea y retornos de carro, por lo que el texto aparece escrito en diferentes líneas.
- **Normalización de la onomatopeya de las risas, ejemplo:** “ajaajajjaj”, “jojjojoj”, “jaaaajjj”, “jajaja”, “jeje”, “jijijiji”, “lol” acrónimo de “*Laughing out loud” que es “reírse en voz alta”.*
- **Eliminación de menciones, enlaces y hashtags**
- **Normalización de jerga,** en las redes sociales se utiliza un lenguaje coloquial y muy informal que hace uso intensivo de abreviaturas y secuencias de caracteres sin aparente sentido. Por ejemplo “q” en lugar de “que”, “tb” en lugar de “también”, “dtb” en lugar de “Dios te Bendiga” o el clásico “tqm” para referirse a “te quiero mucho”

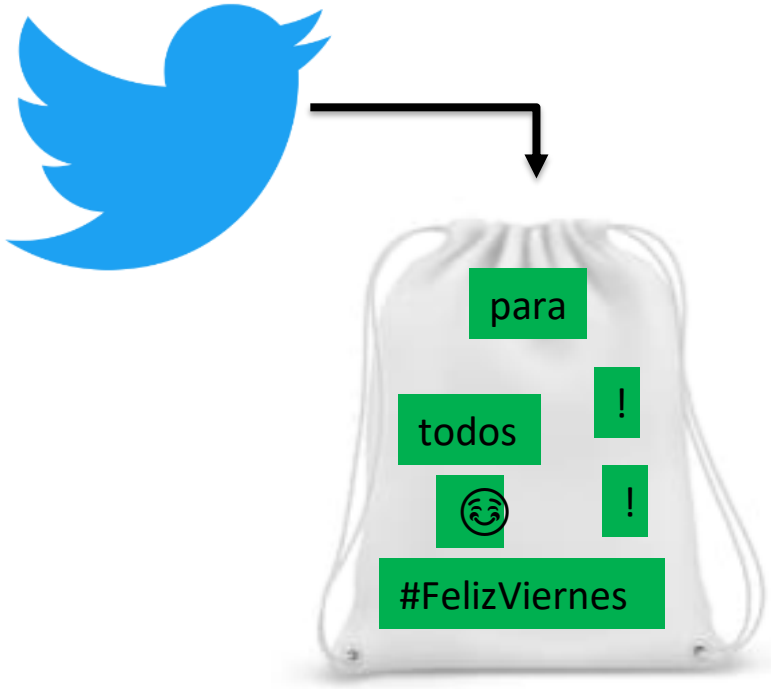
5.2. tokenización

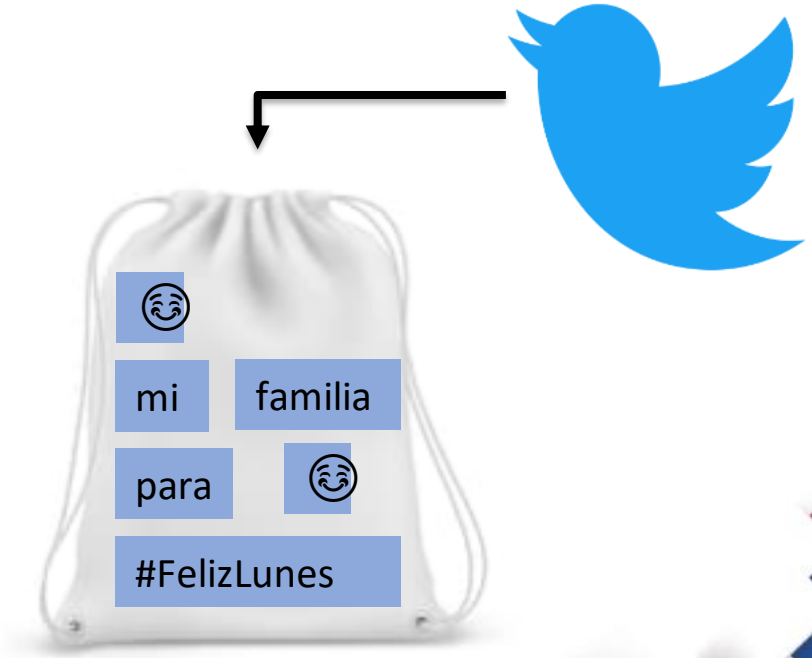
Se tomara cada uno de los mensajes (tweets) y uno por uno se dividirán en unidades más pequeñas denominadas “tokens”:

Preservando los #, emoticones, enlaces, entre otros.

5.3. Extracción de características

Se representara cada mensaje (tweet) por su lista de tokens llamadas “unigramas”.





5.4. Reducción de las características

Esta fase es opcional y su objetivo es disminuir el número de características del corpus mediante la eliminación de determinados *tokens* o de su conversión buscando una misma manera de representarlos. Existe tres técnicas habituales para llevar a cabo esta tarea:

- Eliminación de *stopwords:* : *Eliminar un conjunto de palabras que, aunque son necesarias para construir oraciones con sentido, carecen de información que ayude a determinar la polaridad (positiva/negativa) de los textos en los que se encuentran.*
- Lematización: *Transforma cada palabra en su lema mediante el uso de diccionarios y de un proceso de análisis morfológico. A modo de ejemplo, la lematización convertiría la palabra "guapas" a su lema "guapo"*
- *Stemming:* *Otro método de normalización morfológica pero más agresivo que la lematización. En este caso, una palabra se transforma a su raíz por medio de la supresión de sus sufijos e inflexiones. Siguiendo el ejemplo anterior, las palabras "bellas" , "bellos" se convertiría a su raíz, "bell".*

5.5. Ponderación

Esta fase aplicaremos varias técnicas para ponderar las características extraídas de los textos según su importancia (aplicando pesos), las más conocidas son:

- **Binaria:** *Se califica con 1 si contiene la característica o 0 si no.*
- **Frecuencia absoluta:** *lo mismo que binaria pero se le suma las veces que aparece en el mensaje*.
- **Frecuencia relativa**: *igual a la absoluta pero con normalización de los pesos de las características teniendo presente la tamaño (número de letras) del mensaje.*
- **TF-IDF (Term Frequency – Inverse Document Frecuency)**: *da mayor peso a las características que aparecen en el corpus pero en pocos mensajes.*

5.6. Clasificador de línea base.

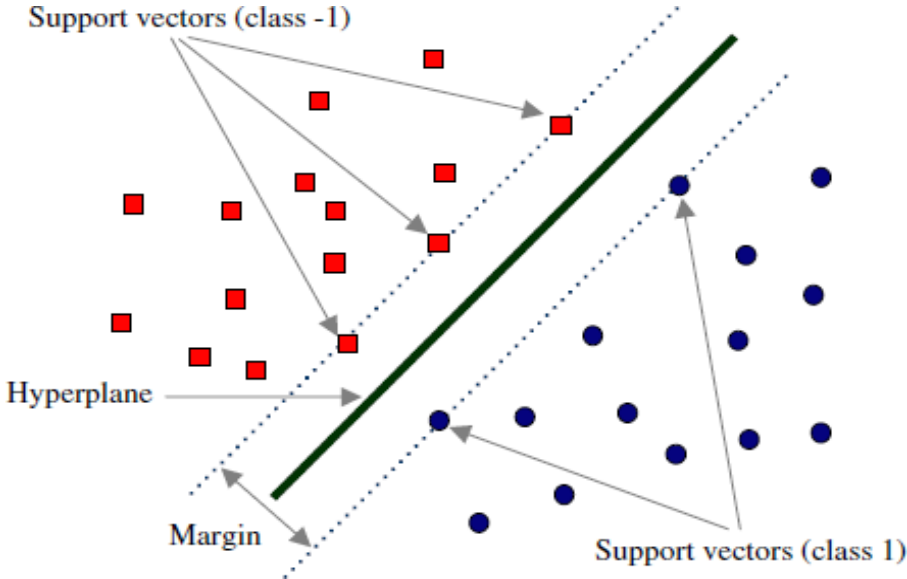
Máquinas de vectores de soporte (Support Vector Machines, SVMs):

tiene las siguientes ventajas:

Ofrece una precisión bastante alta en comparación con otros clasificadores como regresión logística o los arboles de decisión, son más rápidos que las “naive bayes” y usan poca memoria para sus procesos.

Su uso se ve en aplicaciones como :

- *Detección de rostros.*
- *Clasificación de correos electrónicos (spam).*
- *Artículos de noticias (contextos).*
- *Recomendaciones personalizadas en e-commerce*



Desventajas:

No son buenos con grandes cantidades de datos