# proyecto

June 5, 2023

## 1 Proyecto Final

## 2 Caso de uso

Twitch es una plataforma de streaming (videos en directo) que contiene contenido en distintos lenguajes y de distintos tipos. Algunos streamers hacen directos de videojuegos, otros de programacion, etc. Pero ultimamente un conjunto de streamers liderados por Ibai Llanos, han comenzado a hacer directos de un torneo de futbol llamado Kings Legue. Este torneo de futbol es diferente al resto porque tiene algunas reglas peculiares, tales como utilizar cartas comodin que pueden invocar un penal en cualquier momento, sacar jugadores, etc. El objetivo de este proyecto es analizar los comentarios de la jornada 4 de la kings league para entender los topicos que estan presentes en los comentarios del chat.

## 3 Pre- Procesamiento

[ ]:
```
import pandas as pd
# Leer los datos
tabla = pd.read_csv("/home/ereal/Desktop/Master/Text Analytics/Proyecto/
 ↪twitch-chat-1831494742.csv" , error_bad_lines=False)
# Total de comentarios
comentarios_total = len(tabla)
# Eliminar posibles na de la tabla
tabla.dropna(subset=['message'])
```

```
/tmp/ipykernel_12124/4093760913.py:3: FutureWarning: The error_bad_lines
argument has been deprecated and will be removed in a future version. Use
on_bad_lines in the future.


  tabla = pd.read_csv("/home/ereal/Desktop/Master/Text
Analytics/Proyecto/twitch-chat-1831494742.csv" , error_bad_lines=False)
Skipping line 21940: expected 4 fields, saw 5
Skipping line 23849: expected 4 fields, saw 5
Skipping line 26110: expected 4 fields, saw 5
```

[ ]:
```
         time      user_name user_color  \
0           5           d1tp    #000000
1           5    kingjuan200    #FF0000
2           5         moobot    #54BC75
3           6  arnodorian230        NaN
4          10    el_gafitas1    #DAA520
…       …             …        …
48749   23471       xmantekz    #1E90FF
48750   23472   elmadrigamer    #FF7F50
48751   23472     charly_64_    #1E90FF
48752   23472         moobot    #54BC75
48753   23474    miltonjoses        NaN

                                                 message
0                                        vamooooooooos
1                                    OPAAAAAAAAAAAAAAAA
2       Simyo es más fácil que tener a un cactus por …
3                                        por fiiiiiiiin
4                                                     1
…                                                   …
48749                                       Raid Ibai?
48750                                        hasta mñn
48751              @TheGrefg VAMOS QUE SE PUEDE DESCANSA
48752   kleagueQueensLogo Twitter: https://twitter.co…
48753                                     Kuni?? O juan??

[48753 rows x 4 columns]
```

[ ]:
```
# Filtrar comentarios hechos por moobot que se encarga de hacer publicidad en␣
 ↪el chat.
tabla = tabla[(tabla.user_name !="moobot")]
# Obtener la cantidad de comentarios luegos de filtrar
comentarios_filtrado = len(tabla)
# Resumen de la cantidad de comentarios.
print(f"Numero de comentarios total: {comentarios_total}")
print(f"Numero de comentarios despues de filtrar: {comentarios_filtrado}")
print(f"Numero de eliminados {comentarios_total - comentarios_filtrado}")
```

```
Numero de comentarios total: 48754
Numero de comentarios despues de filtrar: 46824
Numero de eliminados 1930
```

[ ]:
```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import re
import pandas as pd
```

```
import numpy as np

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score

import math

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from collections import defaultdict

import seaborn as sns
```

[ ]:
```
import string
from nltk.tokenize import word_tokenize
lemmatizer = nltk.stem.WordNetLemmatizer()

stoplist = stopwords.words('spanish') + list(string.punctuation)
stoplist.append("``")
stoplist.append("''")
stoplist.append("n't")
stoplist.append("'s")
stoplist.append("...")
stoplist.append("--")
stoplist.append("'m")
stoplist.append("'re")
stoplist.append("Q")
stoplist.append("..................")
stoplist.append("n.")
stoplist.append("'ve")
stoplist.append("@")
stoplist.append("!")
stoplist.append("<")
#stoplist.append("kleagueescudo")
#stoplist.append("kleaguelogo")

def lemmatize_text(text):
    st = ""
    text = str.lower(str(text))
    tokens = word_tokenize(text)
    tokens_clean = [token for token in tokens if token not in stoplist]

    for w in tokens_clean:
        st = st + lemmatizer.lemmatize(w) + " "
    return st
```

[ ]:
```
tabla['message'] = tabla['message'].apply(lemmatize_text)
```

[ ]:
```
tabla.head()
```

[ ]:

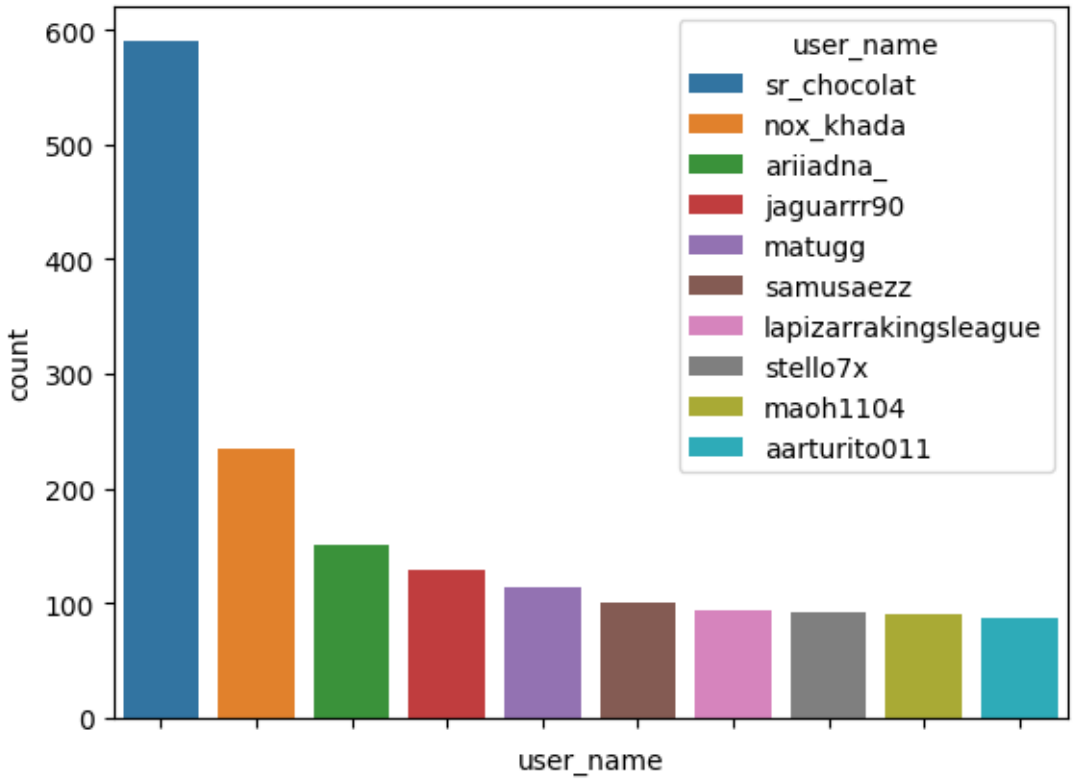
| | time | user_name | user_color | message |
|---|---|---|---|---|
| 0 | 5 | d1tp | #000000 | vamooooooooos |
| 1 | 5 | kingjuan200 | #FF0000 | opaaaaaaaaaaaaaaaa |
| 3 | 6 | arnodorian230 | NaN | fiiiiiiiiin |
| 4 | 10 | el_gafitas1 | #DAA520 | 1 |
| 5 | 13 | jezuoo98 | #DAA520 | kleagueescudo kleagueescudo kleagueescudo |

# 4 Exploracion

## 4.1 Usuarios con mas comentarios

[ ]:
```
usuarios_mas_comentarios = tabla[['user_name','message']].
 ↪groupby("user_name")["message"].count().reset_index(name='count').
 ↪sort_values(['count'], ascending=False) .head(10)
print(usuarios_mas_comentarios)
sns.barplot(data=usuarios_mas_comentarios, x='user_name', y='count',␣
 ↪hue='user_name', dodge=False).set(xticklabels=[]);
```

| | user_name | count |
|---|---|---|
| 13319 | sr_chocolat | 591 |
| 10935 | nox_khada | 235 |
| 1477 | ariiadna_ | 150 |
| 6806 | jaguarrr90 | 128 |
| 9888 | matugg | 114 |
| 12635 | samusaezz | 100 |
| 8614 | lapizarrakingsleague | 93 |
| 13399 | stello7x | 92 |
| 9475 | maoh1104 | 91 |
| 231 | aarturito011 | 87 |




## 5 Topicos

Separando el data set por partido de la jornada 4

[ ]:
```
# XBUYER TEAM vs Rivers (16:00 CET)
partido1 = tabla[(tabla["time"]<=5076)]
print(len(partido1))
```

12236

[ ]:
```
# Jijantes FC vs Los Troncos FC (17:00 CET)
partido2 = tabla[(tabla["time"]>5076) & (tabla["time"]<=8722)]
print(len(partido2))
```

7379

[ ]:
```
# Ultimate Móstoles vs Porcinos FC (18:00 CET)
partido3 = tabla[(tabla["time"]>8722) & (tabla["time"]<=12361)]
print(len(partido3))
```

8484

[ ]:
```
# El Barrio vs 1K FC (19:00 CET)
partido4 = tabla[(tabla["time"]>12361) & (tabla["time"]<=15889)]
print(len(partido4))
```

5706

[ ]:
```
# Rayo de Barcelona vs Saiyans FC (20:00 CET)
partido5 = tabla[(tabla["time"]>15889) & (tabla["time"]<=19628)]
print(len(partido5))
```

7243

[ ]:
```
# Kunisports vs Aniquiladores FC (21:00 CET)
partido6 = tabla[(tabla["time"]>19628) & (tabla["time"]<=23466)]
print(len(partido6))
```

5765

[ ]:
```
# Step 3: Building a Topic Model
from gensim import corpora, models
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import numpy as np

def topic_modeling(messages):
    preprocessed_docs = []
    for doc in messages:
        tokens = word_tokenize(doc.lower())
        preprocessed_docs.append(tokens)


    # Create dictionary and corpus
    dictionary = corpora.Dictionary(preprocessed_docs)
    corpus = [dictionary.doc2bow(doc) for doc in preprocessed_docs]

    # Train the LDA model
    lda_model = models.LdaModel(corpus, num_topics=2, id2word=dictionary,␣
 ↪passes=10)

    # Step 4: Interpretation and Visualization of Results
    # Print the topics
    for topic_id, topic in lda_model.print_topics():
        print(f"Topic ID: {topic_id}\nWords: {topic}\n")

    # Visualization of topics using word clouds
    topics = lda_model.show_topics(num_topics=3, num_words=35, formatted=False)
```

```
    # Generate word clouds for each topic
    for topic_id, words in topics:
        wordcloud = WordCloud(background_color='white').
↪generate_from_frequencies(dict(words))
        plt.figure(figsize=(8, 6))
        plt.imshow(wordcloud, interpolation='bilinear')
        plt.title(f"Topic {topic_id + 1}")
        plt.axis('off')
        plt.show()
```

# 6 Topicos Partido #1

## 6.1 XBUYER TEAM vs Rivers (16:00 CET)

[ ]:
```
topic_modeling(partido1["message"])
```

```
Topic ID: 0
Words: 0.023*"kleagueescudo" + 0.022*"pio" + 0.020*"biblethump" +
0.019*"kleaguepeepoxbuyers" + 0.018*"kleaguepio" + 0.017*"footgoal" +
0.016*"puro" + 0.016*"kleaguexbuyer" + 0.013*"si" + 0.013*"subscribed"

Topic ID: 1
Words: 0.054*"mvp" + 0.053*"lul" + 0.023*"kleaguelogo" + 0.021*"ibairobada" +
0.018*"arbitro" + 0.018*"rivers41pio" + 0.015*"gol" + 0.013*"amarilla" +
0.011*"jorge" + 0.009*"kleagueopa"
```


# 7 Topicos Partido #2

## 7.1 Jijantes FC vs Los Troncos FC (17:00 CET)

[ ]:
```
topic_modeling(partido2["message"])
```

```
Topic ID: 0
Words: 0.113*"amen" + 0.072*"amén" + 0.053*"kleaguejijantes" + 0.048*"lul" +
0.029*"kleagueescudo" + 0.021*"kleaguegerardromero" + 0.020*"dios" +
0.019*"jijantes" + 0.017*"ibairobada" + 0.015*"biblethump"

Topic ID: 1
Words: 0.033*"kleaguelogo" + 0.019*"kleaguepeepojijantes" + 0.014*"kleaguegol" +
0.013*"jijantesfcescudo" + 0.012*"jijantes" + 0.011*"kleagueroja" +
0.010*"kleaguelostroncos" + 0.010*"subscribed" + 0.009*"resultados" +
0.008*"diawara"
```



# 8 Topicos Partido #3

## 8.1 Ultimate Móstoles vs Porcinos FC (18:00 CET)

[ ]:
```
topic_modeling(partido3["message"])
```

Topic ID: 0
Words: 0.031*"ibailando" + 0.028*"biblethump" + 0.023*"3" + 0.020*"seemsgood" + 0.017*"mostoles" + 0.015*"kleaguepeepopoporcinos" + 0.013*"kleagueultimatemostoles" + 0.013*"jijantesfcescudo" + 0.012*"kleaguequeenslogo" + 0.012*"kleaguedjmariio"

Topic ID: 1
Words: 0.071*"lul" + 0.037*"ibaipeeporcino" + 0.026*"porcinos" + 0.024*"kleagueporcinos" + 0.019*"ibaiporcinosfc" + 0.018*"mvp" + 0.014*"league" + 0.012*"resultados" + 0.011*"guti" + 0.010*"cichero"

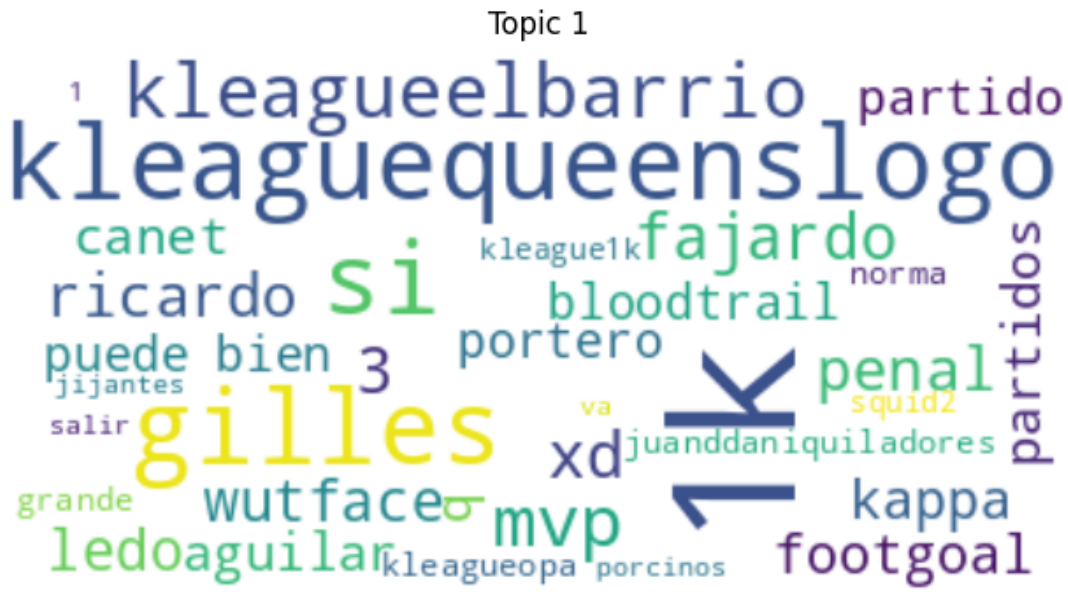

# 9 Topicos Partido #4

## 9.1 El Barrio vs 1K FC (19:00 CET)

[ ]:
```
topic_modeling(partido4["message"])
```

```
Topic ID: 0
Words: 0.016*"1k" + 0.015*"kleaguequeenslogo" + 0.014*"gilles" + 0.014*"si" +
0.013*"kleagueelbarrio" + 0.012*"mvp" + 0.010*"xd" + 0.009*"3" + 0.009*"fajardo"
+ 0.008*"penal"

Topic ID: 1
Words: 0.074*"lul" + 0.028*"resultados" + 0.022*"barrio" + 0.020*"biblethump" +
0.011*"kleaguepeepoelbarrio" + 0.010*"knesleeper" + 0.009*"ibaikekw" +
0.008*"vamos" + 0.007*"f" + 0.007*"seemsgood"
```





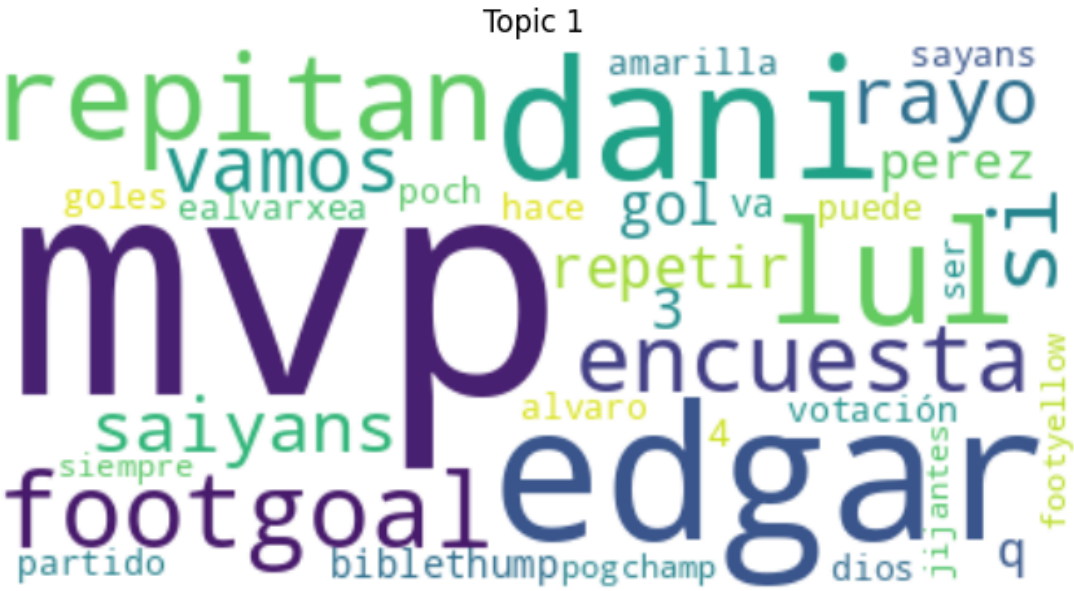
## 10 Topicos Partido #5

### 10.1 Rayo de Barcelona vs Saiyans FC (20:00 CET)

[ ]:
```
topic_modeling(partido5["message"])
```

Topic ID: 0
Words: 0.078*"mvp" + 0.046*"edgar" + 0.034*"dani" + 0.025*"lul" +
0.019*"repitan" + 0.014*"footgoal" + 0.012*"encuesta" + 0.012*"si" +
0.008*"rayo" + 0.008*"vamos"

Topic ID: 1
Words: 0.017*"kleagueopa" + 0.017*"kleaguerayodebarcelona" + 0.012*"resultados"
+ 0.012*"kleaguesaiyans" + 0.012*"golazo" + 0.011*"xd" + 0.007*"kleaguegol" +
0.007*"voteyea" + 0.007*"votacion" + 0.007*"kleaguepeeporayodebarcelona"




# 11 Topicos Partido #6

## 11.1 Kunisports vs Aniquiladores FC (21:00 CET)

```
[ ]: topic_modeling(partido6["message"])
```

```
Topic ID: 0
Words: 0.032*"juanddaniquiladores" + 0.029*"kleagueaniquiladores" + 0.020*"3" +
0.018*"4" + 0.014*"footyellow" + 0.013*"aniquiladores" + 0.010*"juanddescudo" +
0.009*"mvp" + 0.008*"vamos" + 0.008*"kleagueopa"

Topic ID: 1
Words: 0.047*"lul" + 0.031*"kleaguepeepoaniquiladores" + 0.021*"footgoal" +
0.016*"resultados" + 0.016*"seemsgood" + 0.014*"fran" + 0.014*"slakunkunisports"
+ 0.012*"xd" + 0.012*"kleaguejuansguarnizo" + 0.011*"coro"
```



## 12 Topicos Generales de la jornada #4

[ ]:
```
topic_modeling(tabla["message"])
```

```
Topic ID: 0
Words: 0.052*"lul" + 0.037*"kleaguelogo" + 0.029*"juanddaniquiladores" +
```

0.020*"resultados" + 0.017*"kleaguepeepoaniquiladores" + 0.017*"edgar" + 0.016*"kleagueaniquiladores" + 0.015*"4" + 0.013*"footyellow" + 0.012*"vamos"

Topic ID: 1
Words: 0.024*"mvp" + 0.023*"kleagueescudo" + 0.017*"3" + 0.011*"footgoal" + 0.011*"si" + 0.010*"dani" + 0.009*"xd" + 0.008*"aniquiladores" + 0.007*"biblethump" + 0.007*"partido"