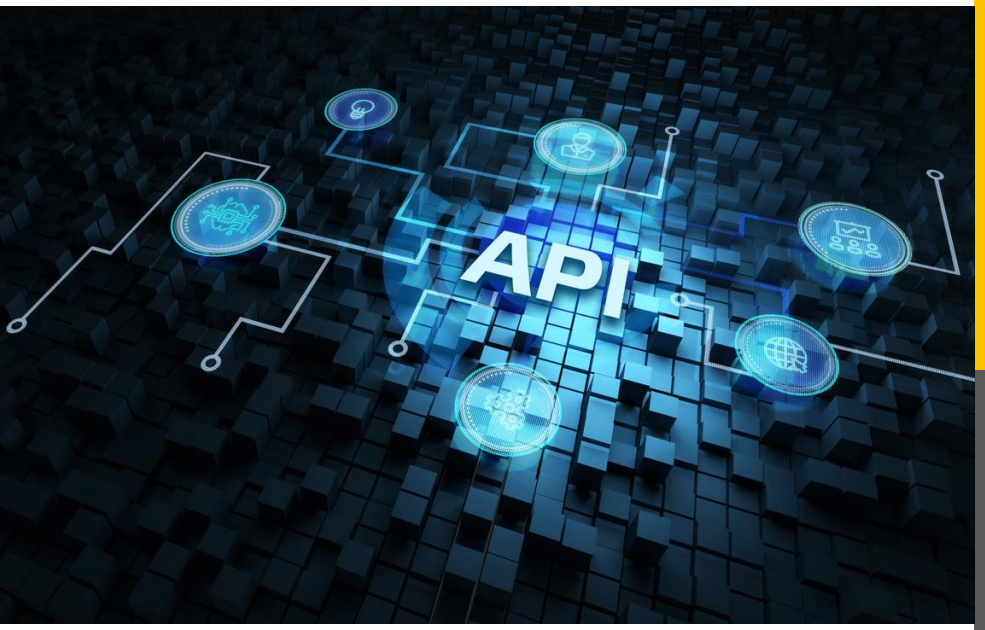
COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI



comillas.edu

Universidad Pontificia Comillas
ICAI

Analítica Social y de la Web
Social and Web Analytics

2.2 Limpieza de datos

Javier Ruiz de Ojeda

Curso 2023-24
Segundo semestre

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

Técnicas básicas de limpieza

Técnicas básicas de limpieza
Introducción

Una vez nos hemos familiarizado con la conexión y respuesta de algunas de las APIs más utilizadas para el análisis social, habremos visto:

- **Endpoints heterogéneos**: Cada API es un monstruo de diferentes características.
- **Diversidad de datos**: La información obtenida es muy diversa y diferente dependiendo de qué plataforma estemos consultando desde perfiles de usuario, estadísticas fundamentales, textos, etc.

Técnicas básicas de limpieza
Introducción

Por supuesto, a la hora de hacer analítica social lo primero que pensamos en datos cuantitativos, pero tál y como hemos podido comprobar con Twitter o Youtube, los datos no estructurados como Twits o comentarios pueden llegar a ser de un valor incalculable.

Sin embargo, a diferencia de los datos cuantitativos que también son recibidos de estas plataformas, su tratamiento para extraer información significativa puede llegar a ser un poco más complejo. El proceso de preparación de tratamiento de datos no estructurados (texto, imágenes, vídeos, etc.) se suele conocer como **limpieza y normalización**. Una vez que los datos hayan sido limpiados y normalizados podremos proceder sin más al análisis social.

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

Pregunta:

¿Por qué limpiar?

Técnicas básicas de limpieza

¿Qué es?

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Técnicas básicas de limpieza

Tipos de datos y codificación

Cuando lidiamos con datos de texto, una de las principales preocupaciones que debemos tener es sobre la codificación de las cadenas de caracteres (**`strings`**) que obtenemos como representación de los tweets, comentarios, etc.

De manera muy resumida, el encoding (codificación) es el proceso mediante el cual una cadena de caracteres se convierte en bytes. Dicha codificación juega un papel muy importante a la hora de poder hacer analítica exacta, sobre todo por el uso de emojis (o emoticonos) como sustitutivos de palabras completas, o el uso de acentos en diferentes lenguajes, etc.

El hecho de que dos strings completamente iguales tengan una codificación diferente puede conllevar una mala categorización de sentimientos, por poner un ejemplo. En Python, el estándar es UTF-8, por lo que como regla general, nos aseguraremos que nuestros datos siempre estén en Unicode UTF-8 para evitar mezclar peras con manzanas.

Técnicas básicas de limpieza
Estructura de los datos

Una de las preguntas claves a la hora de trabajar con datos de cualquier índole es: ¿Cuál es la estructura que mejor representa mis datos?

En el caso del análisis, casi por norma general, tendrá una respuesta sencilla: formato tabular (**`pandas dataframe`**). El porqué es fácil de justificar: la organización en filas y columnas es algo que facilita de forma considerable las operaciones analíticas fundamentales como las búsquedas, agrupaciones, etc.

Recordar que para grandes sets de datos que puedan consumir nuestra memoria RAM, se puede optar por la alternativa **`Sframes`**.

Técnicas básicas de limpieza
Preprocesamiento y normalización

- Esta etapa es una de las más importantes para el posterior análisis, dado que será en esta etapa donde identifiquemos/seleccionemos las partes importantes que posteriormente serán usadas.
- El pre-procesado de columnas numéricas puede incluir:
  - Detección (y corrección) de outliers
  - Detección (y corrección) de **`NaN's`** (*Not a Number*)
  - Dummy-encoding de columnas categóricas
  - Normalización de columnas **`float`**

Técnicas básicas de limpieza
Preprocesamiento y normalización

Por otra parte, el pre-procesado de texto suele involucrar los siguientes pasos:

- Limpieza de espacios en blanco: **`.strip()`**
- Limpieza de símbolos de puntuación: **`re.sub(r"[^\w\s]", "", my_string)`** donde:

```
[      #Character block start.
^      #Not these characters (letters, numbers).
\w     #Word characters.
\s     #Space characters.
]      #Character block end.
```

Técnicas básicas de limpieza
Preprocesamiento y normalización

- Limpieza de elementos HTML: `re.sub(r"<[^<]+?>", "", my_string")`
- Limpieza de URLs: `re.sub(r"^https?:\/\/.*[\r\n]*", "", text, flags=re.MULTILINE)`
- Corrección de palabras con errores
- Limpieza de palabras comunes (vacías de significado), por ejemplo: determinantes, preposiciones o conjucciones.
- Normalización a minúsculas: `my_string.lower()`

Técnicas básicas de limpieza
Preprocesamiento y normalización

- Limpieza de conectores (stop words): Para este caso utilizaremos **`nltk`**:

```
import nltk
nltk.download("stopwords")
from nltk.corpus import stopwords

parrafo = ' '.join([word for word
in parrafo.split() if word not in
(stopwords.words('english'))])
```

```
print(stopwords.words('english'))

['i', 'me', 'my', 'myself', 'we',
'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll",
"you'd", 'your', 'yours',
'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she',
"she's", 'her', 'hers', 'herself',
'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs',
'themselves', 'what', 'which',
'who',…
```

Técnicas básicas de limpieza
Tokenización


```
from nltk import word_tokenize, sent_tokenize, line_tokenize
nltk.download('punkt')
text = ("This is a completely random text in english, and I would like to see the result. "
        "For this is an example that must remain!\nSigned -- Myself")
words = word_tokenize(text)
['This', 'is', 'a', 'completely', 'random', 'text', 'in', 'english', ',', 'and', 'I', 'would', 'like', 'to', 'see', 'the', 'result', '.', 'For', 'this', 'is', 'an', 'example', 'that', 'must', 'remain', '!', 'Signed', '--', 'Myself']
sentences = sent_tokenize(text)
['This is a completely random text in english, and I would like to see the result.', 'For this is an example that must remain!', 'Signed -- Myself']
lines = line_tokenize(text)
['This is a completely random text in english, and I would like to see the result.', 'For this is an example that must remain!', 'Signed -- Myself']
```

Técnicas básicas de limpieza
Traducción de textos

goslate

```
>>> text = 'This site is awesome'
>>> from googletrans import Translator
>>> translator = Translator()
>>> translator.translate(text , dest ='sw').text
'Tovuti hii ni ajabu
```

google translate

```
>>> text = 'This site is awesome'
>>> from googletrans import Translator
>>> translator = Translator()
>>> translator.translate(text , dest ='sw').text
'Tovuti hii ni ajabu
```

textblob

```
>>> from textblob import TextBlob
>>> blob = TextBlob('comment ca va ?')
>>> blob.translate(to='en')
TextBlob("How is it going ?")
```

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

Ejercicios:

Limpieza de datos

Ejercicio
Limpieza de datos

¿Importa el orden?

- Eliminar espacios en blanco
- Eliminar puntuación, tags de HTML y URLs
- Estandarizar palabras (eliminar letras o símbolos repetidos)
- Eliminar palabras concatenadas
- Pasar a minúsculas
- Eliminar palabras comunes
- Tokenizar las palabras resultantes

Ejercicio

Párrafo de ejemplo

The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein, that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age.

Theosophists have guessed at the awesome grandeur of the cosmic cycle wherein our world and human race form transient incidents. They have hinted at strange survivals in terms which would freeze the blood if not masked by a bland optimism. But it is not from them that there came the single glimpse of forbidden aeons which chills me when I think of it and maddens me when I dream of it. That glimpse, like all dread glimpses of truth, flashed out from an accidental piecing together of separated things—in this case an old newspaper item and the notes of a dead professor. I hope that no one else will accomplish this piecing out; certainly, if I live, I shall never knowingly supply a link in so hideous a chain. I think that the professor, too, intended to keep silent regarding the part he knew, and that he would have destroyed his notes had not sudden death seized him.

Ejercicio

Ejercicios a realizar

1. Obtener una lista de palabras en español y otra en inglés con el resultado de limpiar con los pasos anteriores el párrafo de ejemplo
2. Obtener una lista de palabras con el resultado de limpiar con los pasos anteriores el texto de la web www.alotuyo.com
3. Obtener una lista de palabras con el resultado de limpiar con los pasos anteriores las 10 primeras reseñas de la película “Titanic” en www.imdb.com

Entrega de ficheros .`ipynb` por SIFO/email: miércoles 7 de febrero.

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

¡Muchas gracias!