Python

Procesamiento de Lenguaje Natural

Limpieza de Textos

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

Limpieza de Textos

Limpieza de Textos

- **Uno de los primeros pasos para empezar con el análisis del contenido de un archivo es proceder a la limpieza del texto, quitando símbolos y caracteres que por el momento no nos aporten información (aunque para otros tipos de análisis podría ser importante por ejemplo los símbolos de admiración ó interrogación).**

- **Hay varias formas posibles de quitar los caracteres no deseados como son los siguientes:**
  - **Revisar el texto caracter por caracter.**
  - **Revisar el texto a través de Expresiones Regulares.**
  - **Revisar el texto utilizando alguna biblioteca especializada**

Limpieza de Textos

- **Algunos de los posibles caracteres a eliminar podrían ser los siguientes::**
  - **!”#$%&/()=?¡;:_,.-<>´+{}¨*[]**

Limpieza de Textos

- **Código Python:**

```
# Definir una frase
frase_sor_juana = "Hombres necios que acusáis a la mujer sin razón sin ver que sois la ocasión de lo mismo que culpáis: si con ansia sin igual solicitáis su desdén ¿por qué queréis que obren bien si las incitáis al mal?"

# Definir la frase a procesar
texto = frase_sor_juana

# Impresión en pantalla de la frase seleccionada
print("\nTexto Completo:\n", texto)
```

- **Salida**

```
Texto Completo:
 Hombres necios que acusáis a la mujer sin razón sin ver que sois la ocasión de lo mi
smo que culpáis: si con ansia sin igual solicitáis su desdén ¿por qué queréis que obr
en bien si las incitáis al mal?
```

Limpieza de Textos

- **Código Python:**

```
# Eliminar simbolos y caracteres especiales
eliminar = '()[],.""“”¿?:;_-¡!'                    # Agregar los que falten
texto_sin_simbolos = ""
for caracter in texto:
    if(caracter not in eliminar):
        texto_sin_simbolos = texto_sin_simbolos + caracter
```

```
# Impresión del texto final sin caracteres
print('\nTexto final sin simbolos ni caracteres especiales:\n', texto_sin_simbolos, "\n")
```

- **Salida**

```
Texto final sin simbolos ni caracteres especiales:
 Hombres necios que acusáis a la mujer sin razón sin ver que sois la ocasión de lo mi
smo que culpáis si con ansia sin igual solicitáis su desdén por qué queréis que obren
 bien si las incitáis al mal
```

Limpieza de Textos con Expresiones Regulares

Limpieza de Textos

- **Otra técnica es utilizar "Expresiones Regulares", la cual nos permite identificar rápidamente caracteres y patrones de caracteres que se encuentran en un texto y esto lo hace mucho más fácil.**

- **Python cuenta de manera interna con una biblioteca para "Expresiones Regulares" por lo cual no es necesario instalar ningún componente adicional, sólo importarlo en nuestro código como cualquier biblioteca de la siguiente manera:**

```
# Bibliotecas a utilizar
import re                          # Manejo de expresiones regulares
```

Limpieza de Textos con ER

- **Código Python:**

```
# Eliminar simbolos y caracteres especiales usando expresiones regulares
texto_sin_simbolos = re.sub(r'[^\w\s]','', texto)
```

```
# Impresión del texto final sin caracteres
print('\nTexto final sin simbolos ni caracteres especiales:\n', texto_sin_simbolos, "\n")
```

- **Salida:**

```
Texto final sin simbolos ni caracteres especiales:
 Hombres necios que acusáis a la mujer sin razón sin ver que sois la ocasión de lo mismo que culpáis si co
n ansia sin igual solicitáis su desdén por qué queréis que obren bien si las incitáis al mal
```

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

rogelio.fe@morelia.tecnm.mx

rogeplus@gmail.com

xumarhu.net

@rogeplus

https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg

rogelioferreiraescutia