Logotipo

Descripción generada automáticamente

Sistemas de Información no estructurada

Trabajo del Máster Universitario en Ingeniería Informática

Memoria Practica Final

Victor Fagúndez Poyo

E.T.S de Ingeniería Informática

Curso 2021/2022

|  |  |  |
| --- | --- | --- |
| Historial de Revisiones | | |
| **Versión** | **Fecha** | **Descripción** |
| 0.0.0 | 29.03.22 | Creación del documento |

**Contenido**

[1. Enunciado 1](#_Toc102117061)

[2. Consideraciones iniciales 3](#_Toc102117062)

[3. Algoritmo base 3](#_Toc102117063)

[3.1. Descripción de la implementación realizada 3](#_Toc102117064)

[3.1.1. Problemas encontrados 3](#_Toc102117065)

[3.2. Descripción de los resultados obtenidos 4](#_Toc102117066)

[3.2.1. Evaluación usando 10 cross-fold validation 4](#_Toc102117067)

[4. Conclusiones 4](#_Toc102117068)

[4.1. Impacto de las variables en los resultados 4](#_Toc102117069)

[4.2. Combinaciones de variables 4](#_Toc102117070)

[4.3. Mejor conjunto de variables 4](#_Toc102117071)

[5. Bibliografía 5](#_Toc102117072)

Índice de Ilustraciones

[Ilustración 1. Error no encuentra el fichero de entrenamiento 3](#_Toc102117073)

[Ilustración 2. Comprobación de la existencia de los datasets 4](#_Toc102117074)

[Ilustración 3. El script indica que el dataset no existe en esa ruta y finaliza 4](#_Toc102117075)

[Ilustración 4. Se añade un punto a la ruta, para acceder a la carpeta anterior a la actual 4](#_Toc102117076)

Índice de Tablas

**No se encuentran elementos de tabla de ilustraciones.**

# Enunciado

Se deberán realizar un mínimo de 4 modificaciones. En concreto, el objetivo de la práctica es mejorar el conjunto de variables utilizadas en el clasificador SVM, ya sea mediante la modificación de las variables propuestas o mediante la inclusión de nuevas (generalmente mediante la concatenación para cada vector). La idea de la práctica es jugar con diferentes tipos de variables, analizar su funcionamiento y rendimiento mediante la evaluación, así como compararlas, ya sea por separado o combinadas. Algunas sugerencias de modificaciones son:

* Modificar el tamaño del diccionario para el cálculo de tf-idf, ya sea mediante la eliminación de Stop Words, mediante la selección de las XX palabras más usadas, 9 o la elección de ciertos conjuntos gramaticales (verbos y adjetivos, o nombres y adjetivos).
* Generación de la matriz de tf-idf mediante otros elementos que no sean palabras, como pueden ser bi-gramas, tri-gramas o n-gramas, así como el uso de lemas en lugar de palabras, o mediante concatenación de pares como puede ser palabra/categoría-Gramatical.
* Uso de conceptos en lugar de palabras mediante algoritmos de desambiguación semántica.
* Inclusión de elementos como los vectores de Word Embeddings. • Inclusión de categorías gramaticales más frecuentes.
* Uso de léxicos afectivos o de palabras agresivas para identificar palabras polares (SentiSense, Spanish Emotion Lexicon, HurtLex, etc.)

Otra opción de mejora, es la modificación del algoritmo de entrenamiento, pudiendo utilizar otras familias de algoritmos como pueden ser los arboles de decisión, los sistemas basados en regla, o la combinación de varios mediante sistemas de votación.

Como ya se ha expuesto, los puntos aquí planteados son sugerencias de posibles modificaciones, pero no es necesario restringirse a estos. Se pueden incluir cualesquiera originales, o ya vistos en la bibliografía o estado del arte de otros artículos científicos en el área de procesamiento del lenguaje o clasificación de texto.

Por último, se deberá realizar una memoria en la que se detalle el diseño de las modificaciones realizadas sobre el clasificador básico propuesto, su evaluación, análisis de resultados y conclusiones extraídas. En concreto la memoria deberá contener:

* Descripción de la implementación realizada, problemas encontrados, soluciones aportadas, etc.
* Descripción de resultados obtenidos para diferentes variables, y de todas las alternativas/combinaciones realizadas, con un análisis exhaustivo de los mismos en base a las distintas métricas utilizadas.
* Conclusiones que se puedan obtener a partir del análisis de resultados: cómo afectan las variables a los resultados, sus combinaciones, mejor conjunto de variables y por qué, etc

# Consideraciones iniciales

Dado que los datasets de entrenamiento y test proporcionados para la practica superaban con las 6000 y 4000 líneas respectivamente, se ha decidido, con el fin de acelerar las pruebas intermedias durante el desarrollo de las modificaciones del algoritmo, que se reducirán estos datasets a un tamaño de 100 líneas cada uno de las secciones en castellano. Una vez se hayan finalizado las modificaciones, y se tenga un prototipo de algoritmo viable, se realizará una ejecución con los datasets originales.

# Algoritmo base

## Descripción de la implementación realizada

Se realizará en primer lugar una ejecución del algoritmo base sin modificaciones a modo de referencia para el resto de las modificaciones que se implementen.

### Problemas encontrados

#### Problema 1

##### Descripción del problema

Cuando tratamos de ejecutar el script por primera vez, nos devuelve un mensaje de error, el cual se muestra en la ilustración 1.

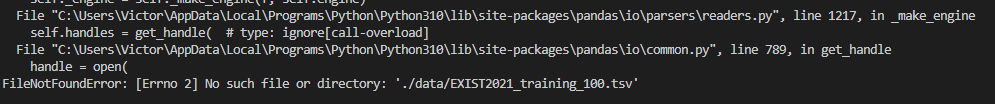


Ilustración . Error no encuentra el fichero de entrenamiento

##### Solución

Con el fin de realizar una comprobación de la existencia de los archivos de entrenamiento antes entrar a la función processEXISTTraining(), se ha optado por añadir una condición if que compruebe la existencia de los ficheros en la ruta proporcionada.

Texto

Descripción generada automáticamente

Ilustración . Comprobación de la existencia de los datasets

Si volvemos a ejecutar el script nos devuelve los siguiente:

Texto

Descripción generada automáticamente

Ilustración . El script indica que el dataset no existe en esa ruta y finaliza

En base a esto se interpreta que el script no esta consiguiendo acceder a la carpeta en la que se encuentran los datasets en la carpeta *data* del nivel anterior. Por ello, en la ruta descrita para pathTraining y pathTest se añade un punto en la parte anterior a la ruta ya descrita para que retroceda al nivel de jerarquía anterior. Por lo tanto, el código quedaría de la forma siguiente.

Texto

Descripción generada automáticamente

Ilustración . Se añade un punto a la ruta, para acceder a la carpeta anterior a la actual

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Tras ejecutar el script obtenemos los resultados mostrados en la ilustración 5.

Calendario

Descripción generada automáticamente

Ilustración . Resultados del algoritmo base

Si observamos los resultados del algoritmo base tenemos que tiene una precisión o accuracy del 0.66.

# Algoritmo base

## Descripción de la implementación realizada

### Problemas encontrados

#### Problema 1

##### Descripción del problema

##### Solución

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

# Conclusiones

## Impacto de las variables en los resultados

## Combinaciones de variables

## Mejor conjunto de variables

# Bibliografía