Legibilidad en Textos Académicos

José M. Tapia Téllez

Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica

Recuperación de Información

Contenido

Sección 1

Obtención de los Documentos

Inaoe Corpus

Se genera un archivo xml con los textos académicos y su respectiva sección de interés.

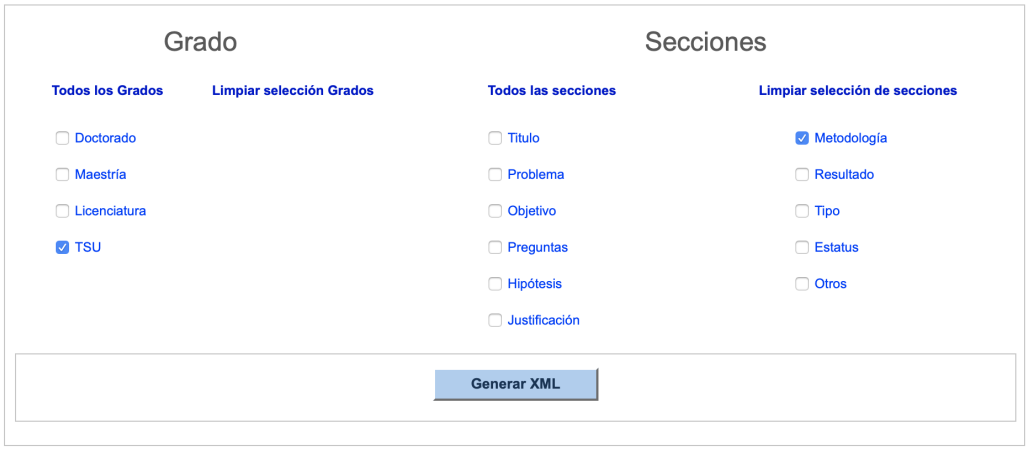


Figura: Caption

Obtención del texto de interés

Metodologías

Del archivo xml se obtiene el texto de nuestro interés y se genera un archivo donde cada metodología se identifica con un Metodologia al principio y un Metodologia al final.

```
<Metodologia>Para el desarrollo de este proyecto de software se aplicará SCRUM,
el cual es conjunto de buenas
prácticas para trabajar colaborativamente, en equipo, y obtener el mejor resulta
do posible
de un proyecto. SCRUM se basa en entregas parciales y regulares del producto fin
al por lo que
SCRUM está indicado para este tipo de proyectos en donde suelen ser entornos com
plejos.
En la metodología SCRUM se establece una lista de tareas las cuales son desarrol
ladas en una
o varias iteraciones, al finalizar cada iteración se obtiene un incremento opera
tivo del producto.
Como resultado de estas iteraciones son el desarrollo ágil del proyecto y SCRUM
gestiona esa
evolución a través de reuniones breves y diarias. SCRUM maneja 2 actividades, la
 planificación,
inspección y adaptación.
</Metodologia>
```

Limpieza y Estructuración de Metodologías

1. Las metodologías se estructuraron en un diccionario de la forma: Metodología-VectorDePalabras
2. Se obtuvo un archivo con palabras vacías del español y se eliminaron de las metodologías

Frecuencia Relativa y Log Frecuencia

1. Por cada palabra en cada metodología se obtuvo el número de veces que ésta, aparecía en la metodología y se dividió entre el tamaño de la metodología.
2. Se obtuvo un archivo de las palabras más comunes del español con su frecuencia y con éste se creó un diccionario Metodologia-Palabra-LogFrecuencia.

Trabajo por hacer I

1. Programar el proceso de la concatenación de los vectores.
2. Utilizar el algoritmo para los otros textos académicos (Licenciatura, Maestría y Doctorado).
3. Introducir los datos a SVM y con ello hacer pruebas y experimentos.

Sección 2

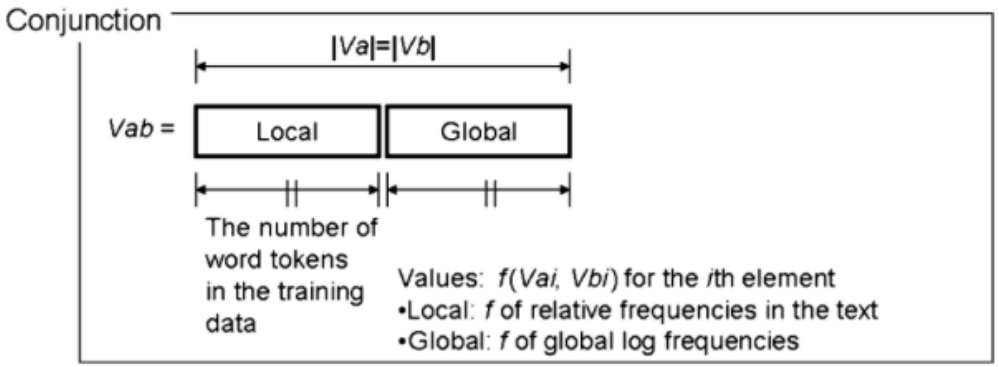
Resta y Concatenación

Resta

Se programó el proceso de resta entre los vectores, esto como resta entre elementos entrada a entrada. Se programó tanto para TSU-Lic como para Lic-TSU.

Concatenación

Se programó y se realizó el proceso de concatenación vector a vector tanto para TSU-Lic como para Lic-TSU



Datos de Entrenamiento y Resultados

Datos de Entrenamiento

Se creó la matriz con los datos de entrenamiento y el vector con los datos de clasificación para la matriz.

Resultados

A través de Scikit Learn y con SVM se entrenó a una máquina SVM con un 80 % de datos. Se dejó un 20 % para pruebas y los resultados son los siguientes:

	precision	recall	f1-score	support
-1.0	1.00	0.77	0.87	13
1.0	0.88	1.00	0.93	21
accuracy			0.91	34
macro avg	0.94	0.88	0.90	34
weighted avg	0.92	0.91	0.91	34

Trabajo por Hacer II

1. Realizar el mismo procedimiento pero ahora incluyendo Maestría y Doctorado.
2. Los datos de entrenamiendo aumentarían, así que se incluyen todos los nuevos y se entrena nuevamente la máquina SVM.
3. Si los resultados son favorables comenzar con la tarea de ordenamiento de los textos (Preguntar.)