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.

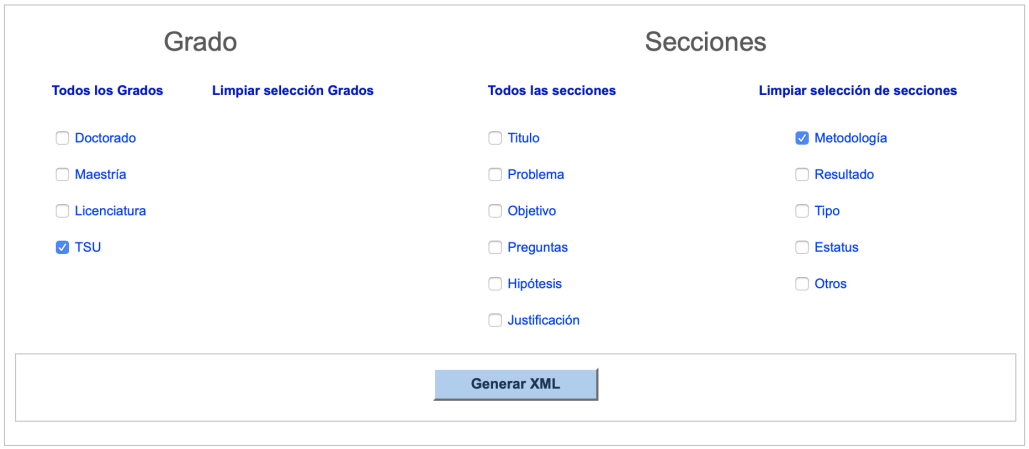


Figura: Caption

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

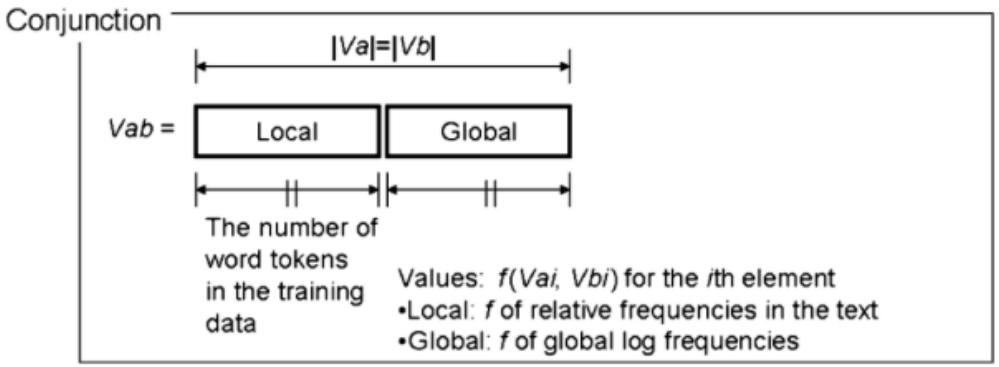
Resta y Concatenación

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.)

Sección 3

Arreglo Clasificatorio en SVM

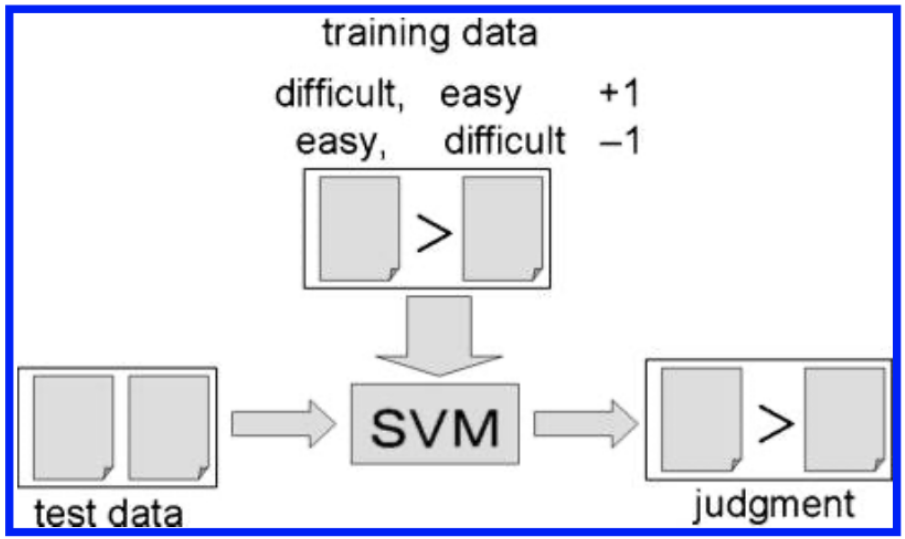


Figura: Arreglo sobre SVM

Maestría y Doctorado

Maestría y Doctorado

Se incluyeron las justificaciones de Maestría y Doctorado y se realizaron diferentes entrenamientos con SVM.

Entrenamiento General

Primero se entrenó con todos los datos juntos, esto para TSU, Licenciatura, Maestría y Doctorado.

	precision	recall	f1-score	support
-1.0	0.92	0.92	0.92	38
1.0	0.93	0.93	0.93	40
accuracy			0.92	78
macro avg	0.92	0.92	0.92	78
weighted avg	0.92	0.92	0.92	78

Figura: TSU, Licenciatura, Maestría y Doctorado Concatenados

Resultados Local y Global Concatenados

TSU y Licenciatura

Resultado para entrenamiento con TSU y Licenciatura con concatenación de vectores locales y globales.

	precision	recall	f1-score	support
-1.0	0.87	0.87	0.87	15
1.0	0.89	0.89	0.89	19
accuracy			0.88	34
macro avg	0.88	0.88	0.88	34
weighted avg	0.88	0.88	0.88	34

Figura: TSU y Licenciatura Concatenados

Resultados Local y Global Concatenados

Licenciatura y Maestría

Resultado para entrenamiento con Licenciatura y Maestría con concatenación de vectores locales y globales.

	precision	recall	f1-score	support
-1.0	0.93	0.87	0.90	15
1.0	0.90	0.95	0.92	19
accuracy			0.91	34
macro avg	0.91	0.91	0.91	34
weighted avg	0.91	0.91	0.91	34

Figura: Licenciatura y Maestría Concatenados

Resultados Local y Global Concatenados

Maestría y Doctorado

Resultado para entrenamiento con Maestría y Dctorado con concatenación de vectores locales y globales.

	precision	recall	f1-score	support
-1.0	0.86	0.86	0.86	7
1.0	0.67	0.67	0.67	3
accuracy			0.80	10
macro avg	0.76	0.76	0.76	10
weighted avg	0.80	0.80	0.80	10

Figura: Maestría y Doctorado Concatenados

ResultadosVector Local

TSU, Licenciatura, Maestría y Doctorado

Resultado para entrenamiento con TSU, Licenciatura, Maestría y Doctorado de vectores locales.

	precision	recall	f1-score	support
-1.0	0.62	0.59	0.61	39
1.0	0.61	0.64	0.62	39
accuracy			0.62	78
macro avg	0.62	0.62	0.62	78
weighted avg	0.62	0.62	0.62	78

Figura: Caption

ResultadosVector Local

TSU, Licenciatura, Maestría y Doctorado

Resultado para entrenamiento con TSU y Licenciatura.

	precision	recall	f1-score	support
-1.0	0.88	0.94	0.91	16
1.0	0.94	0.89	0.91	18
accuracy			0.91	34
macro avg	0.91	0.91	0.91	34
weighted avg	0.91	0.91	0.91	34

Figura: TSU, Licenciatura, Maestría y Doctorado Local

Resultados Vector Local

TSU y Licenciatura

Resultado para entrenamiento con TSU, Licenciatura, Maestría y Doctorado de vectores locales.

	precision	recall	f1-score	support
-1.0	0.62	0.59	0.61	39
1.0	0.61	0.64	0.62	39
accuracy			0.62	78
macro avg	0.62	0.62	0.62	78
weighted avg	0.62	0.62	0.62	78

Figura: TSU y Licenciatura Local

Resultados Vector Local

Licenciatura y Maestría

Resultado para entrenamiento con Licenciatura y Maestría de vectores locales.

	precision	recall	f1-score	support
-1.0	0.86	0.92	0.89	13
1.0	0.95	0.90	0.93	21
accuracy			0.91	34
macro avg	0.90	0.91	0.91	34
weighted avg	0.91	0.91	0.91	34

Figura: Licenciatura y Maestría Local

Resultados Vector Local

Maestría y Doctorado

Resultado para entrenamiento con Maestría y Doctorado de vectores locales.

	precision	recall	f1-score	support
-1.0	1.00	1.00	1.00	6
1.0	1.00	1.00	1.00	4
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

Figura: Maestría y Doctorado Local

Resultados Vector Global

TSU, Licenciatura, Maestría y Doctorado

Resultado para entrenamiento con TSU, Licenciatura, Maestría y Doctorado de vectores globales.

	precision	recall	f1-score	support
-1.0	0.71	0.14	0.24	35
1.0	0.58	0.95	0.72	43
accuracy			0.59	78
macro avg	0.65	0.55	0.48	78
weighted avg	0.64	0.59	0.50	78

Figura: TSU, Licenciatura, Maestría y Doctorado Global

Resultados Vector Global

TSU y Licenciatura

Resultado para entrenamiento con TSU y Licenciatura de vectores globales.

	precision	recall	f1-score	support
-1.0	0.71	0.14	0.24	35
1.0	0.58	0.95	0.72	43
accuracy			0.59	78
macro avg	0.65	0.55	0.48	78
weighted avg	0.64	0.59	0.50	78

Figura: TSU y Licenciatura Global

Resultados Vector Global

Licenciatura y Maestría

Resultado para entrenamiento con Licenciatura y Maestría de vectores globales.

	precision	recall	f1-score	support
-1.0	0.71	0.14	0.24	35
1.0	0.58	0.95	0.72	43
accuracy			0.59	78
macro avg	0.65	0.55	0.48	78
weighted avg	0.64	0.59	0.50	78

Figura: Licenciatura y Maestría Global

Resultados Vector Global

Maestría y Doctorado

Resultado para entrenamiento con Maestría y Doctorado de vectores globales.

	precision	recall	f1-score	support
-1.0	1.00	1.00	1.00	5
1.0	1.00	1.00	1.00	5
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

Figura: Maestría y Doctorado Global

Trabajo por Realizar

1. Realizar pruebas con selección de vector aleatoria representativo.
2. Realizar pruebas con selección de vector representativo a través de centroides.

Sección 4

Comparación aleatoria

Obtener Texto de Prueba

- Se realizó la función obtainTest(A,B), que recibe dos matrices con los los vectores de justificación. Una para local y otra para global.
- La función regresa estos dos vectores por separado.
- Es importante señalar que la función escoge una justificación aleatoria y la remueve para que ésta ya no pueda ser usada en el entrenamiento.

Clasificar el Grado de Legibilidad

random

- Se creó la función random(A,B). Ésta obtiene un vector local y uno global de algún nivel de dificultad.
- El vector que se obtiene es aleatorio, de aquí el nombre de Comparación Aleatoria para este método.
- La función regresa los dos vectores correspondientes a la justificación aleatoria.

Clasificar el Grado de Legibilidad

Obtención de Vectores Aleatorios

- Se creó la función obtainRandomVecs(A,B, C, D, E, F, G, H). Ésta recibe los vectores locales y globales de todos las clases de justificaciones.
- La función crea dos listas, una para vectores globales y otra para locales, éstas se llenan haciendo uso de la función random.
- La función finalmente regresa estas dos listas.

Clasificar el Grado de Legibilidad

Obtener el Grado

- Se creó la function obtainGrade, que recibe las dos listas creadas anteriormente así como el vector local y global de la prueba y la matriz de entrenamiento con su vector de clasificación.
- Para cada una de las posibles clasificaciones de grado se utiliza la máquina de valoración de dificultad, si nos regresa que el texto es más fácil, ahí mismo para y nos regresa ese nivel de grado.

Resultados

- Para TSU son buenos, pero como observación es interesante que nunca regresa del mismo valor.
- Para Licenciatura son buenos, y, de nuevo, siempre nos regresa grado máximo de Maestría
- Para Maestría son buenos y nunca se regresa en clasficación, lo que sí a veces lo dictamina hasta mejor que doctorado.
- Para Doctorado tengo pésimos resultados. Nunca lo cataloga mejor que licenciatura.

Trabajo por Realizar

1. Revisar la parte relacionada con Doctorado.
2. Realizar la clasificación utilizando centroides.

Sección 5

Datos del Trabajo

Cuadro: Justificaciones por Tipo de Datos

Tipos de Datos	Cantidad
TSU	192
Licenciatura	84
Maestría	166
Doctorado	25

Datos de Trabajo

Cuadro: Datos Concatenados

Nivel de Justificación	Número de datos
TSU y Licenciatura	82
Licenciatura y TSU	82
Licenciatura y Maestría	82
Maestría y Licenciatura	82
Maestría y Doctorado	24
Doctorado y Maestría	24

Resultados Selección Aleatorios

Cuadro: Resultados Selección Aleatoria

Pruebas	TSU	Licenciatura	Maestría	Doctorado
0	M	M	M	M
1	M	M	M	M
2	M	M	D	M
3	M	L	M	M
4	L	M	L	M
5	D	D	M	D
6	M	M	M	M
7	L	M	L	L
8	M	L	D	M
9	M	L	M	M
Precisión	0 %	30 %	60 %	10 %

Centroides

Obtención de Centroides

- Se realizó la función de centroides.
- Como entrada recibe tanto la frecuencia relativa como la frecuencia logarítmica de un grado y regresa su centroide.

Resultados Centroides

Cuadro: Resultados Centroides

Pruebas	TSU	Licenciatura	Maestría	Doctorado
0	T	L	D	D
1	T	T	M	D
2	T	T	D	D
3	T	D	D	D
4	T	T	D	D
5	T	D	D	D
6	T	T	D	D
7	L	T	D	D
8	T	D	D	D
9	L	T	D	D
Precisión	80 %	10 %	10 %	100 %

Sección 6

Mínimos, Máximos y Promedio

```
{'Smallest': 13, 'Biggest': 581, 'Average': 157.7958115183246}
{'Smallest': 21, 'Biggest': 441, 'Average': 127.6867469879518}
{'Smallest': 17, 'Biggest': 871, 'Average': 153.75151515151515}
{'Smallest': 32, 'Biggest': 421, 'Average': 184.16666666666666}
```

Figura: Máximos, Mínimos y Promedios por Grado

Ampliación de Objetos

```
TSU-Lic: 15853
Lic-TSU: 15853
Lic-Maestria: 13695
Maestria-Lic: 13695
Maestria-Doctorado: 3960
Doctorado-Maestria: 3960
```

Figura: Cantidad de Objetos por Combinación

Objetos Totales

67,016

Resultados Desviación Estándar y Promedio

5 Corridas con 20 % de Datos de Prueba

Entrenando con TSU-Lic: DE = 0.0, Promedio = 1.0
Entrenando con Lic-Maestría: DE = 0.0, Promedio = 1.0
Entrenando con Maestría-Doctorado: DE = 0.0, Promedio = 1.0
Entrenando con todas juntas:

Sección 7

Resultados 25 Vectores TSU-Lic-Maestria-Doctorado

Cuadro: Accuracy Report

	P1	P2	P3	P4	P5	average	s.d.
svm	0.59	0.65	0.65	0.63	0.56	0.6075	0.0349
random	0.25	0.25	0.25	0.3	0.25	0.266	0.032
centroid	0.25	0.20	0.25	0.35	0.31	0.2719	0.052

Resultados 25 Vectors TSU-Lic-Maestria

Cuadro: Accuracy Report

	P1	P2	P3	P4	P5	average	s.d.
svm	0.63	0.61	0.50	0.53	0.62	0.578	0.052
random	0.1	0.33	0.33	0.33	0.33	0.284	0.092
centroid	0.27	0.33	0.22	0.33	0.06	0.242	0.099

Resultados 50 Vectores TSU-Lic-Maestria

Cuadro: Accuracy Report

	P1	P2	P3	P4	P5	average	s.d.
svm	0.63	0.54	0.50	0.53	0.49	0.5359	0.050
random	0.46	0.36	0.33	0.36	0.36	0.368	0.047
centroid	0.46	0.30	0.24	0.33	0.36	0.337	0.072