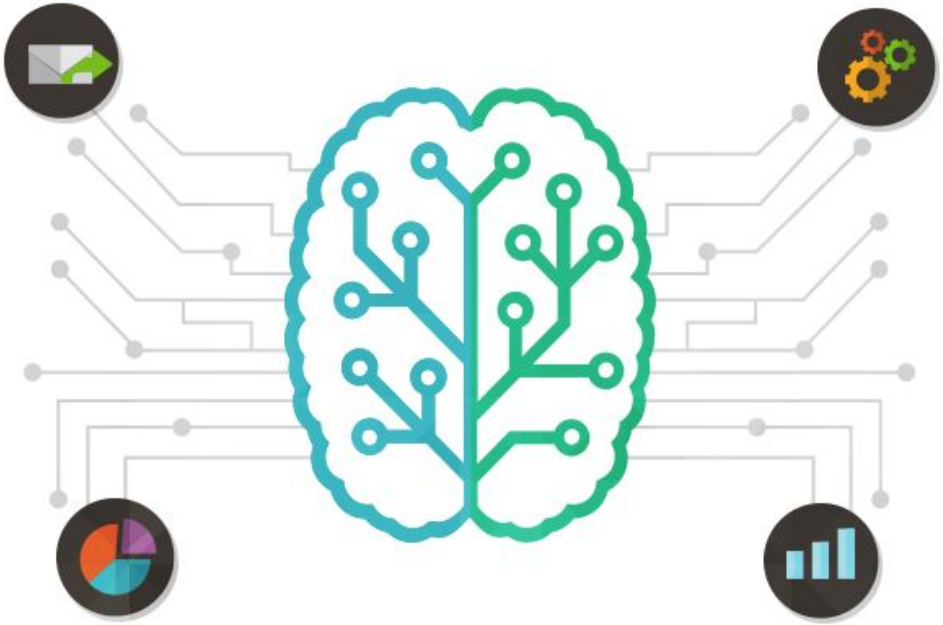
Clasificación de sitios web según su contenido

Daniel Marín
Dayana Rodrigues
Luis Carlos Díaz

Objetivo

El objetivo del proyecto es aplicar algoritmos de aprendizaje de máquina +tipo+ para separar los dominios que contengan información significativa o cuyo tópico principal sea el Software y negocios.



Justificación

La intención de separar los dominios relacionados con Software y negocios es encontrar diferentes herramientas, como por ejemplo lenguajes, frameworks, soluciones SaaS (Software as a Service), librerías, etc. que puedan ayudar a solucionar algún problema específico de la mejor manera posible.

Técnicas y herramientas

Herramientas:

- Alexa.
- Common Crawl.
- Weka.

Técnicas:

- Minería de Texto.
- Aprendizaje supervisado.

Datos empleados

Datos sin procesar:

- Url
- Texto Visible

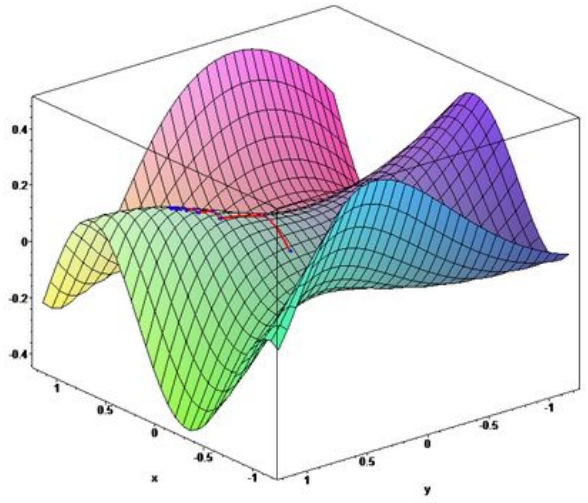
Datos procesados:

- Url
- 985 valores fueron tomados en cuenta luego de realizar el preprocesamiento.

Solución

Para la tarea planteada, se crearon 3 modelos usando clasificadores distintos listados a continuación:

- Naive Bayes
- Support Vector Machine
- Stochastic Gradient Descent



Resultados

- Train/Test Split (70%)

	TP	FP	TN	FN	Recall
Naive Bayes	15	1	19	1	0,944 %
SVM	14	2	20	0	0,944 %
SGD	12	4	16	4	0,778 %

- Cross-validation (10 fold)

	TP	FP	TN	FN	Recall
Naive Bayes	51	8	55	5	0,891 %
SVM	45	14	59	1	0,874 %
SGD	46	13	49	11	0,798 %

Conclusiones

Al ser este un problema de clasificación (Aprendizaje supervisado), existen múltiples alternativas para su solución.
La selección de la herramienta óptima es una decisión que se reflejará tanto en la efectividad de la solución, como en su tiempo y costos.