Clasifiación de Preguntas de StackOverflow de Acuerdo a su Estado de Solución

Juan Leonardo Padilla Gómez
Maestría Ingeniería Sistemas y Computación - 2021-01

UNIVERSIDAD NACIONAL DE COLOMBIA

Introducción

- StackOverflow es un sitio web creado en 2008 con el propósito de facilitar la interacción entre programadores y facilitar el intercambio de preguntas y respuestas enfocado al mundo de la programación.
- El presente proyecto, tiene como objetivo desarrollar un modelo de clasificación aplicado a un cuerpo de preguntas realizadas en el sitio de StackOverflow con el fin de determinar, a partir de un conjunto de características de la pregunta, si la pregunta fue o no resuelta.
- En este proyecto, se pretende explorar el uso de diferentes algoritmos de procesamiento de lenguaje natural como apoyo en la tarea de clasificación de preguntas.

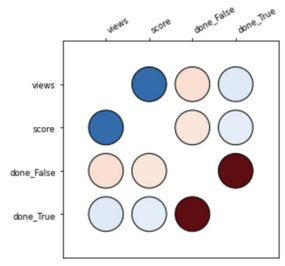
Descripción Conjunto de Datos y Exploración

- Se realiza el análisis univariado de algunas de las variables encontradas en el conjunto de datos. En la siguiente tabla se presenta las frecuencias absolutas para la variable “Lenguaje de programación”, de esta tabla se concluye como los lenguajes de programación más populares en la comunidad de StackOverflow son Python, seguido por Javascript, mientras que el menos popular es R.

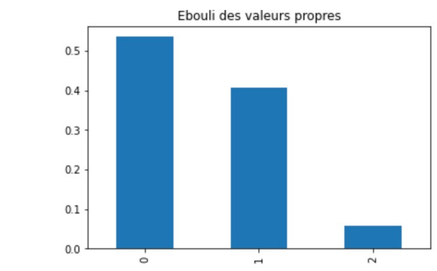
Lenguaje	Conteo
python	19530
javascript	17897
java	16418
php	12304
R	9185

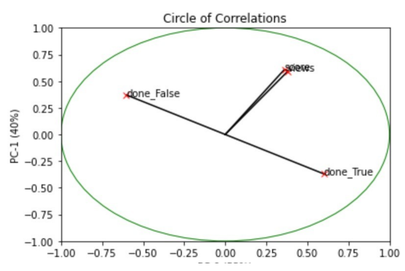
En la siguiente tabla se presentan las frecuencias absolutas para la variable “Estado de la pregunta”. Es importante notar como más del 70% de las preguntas no son marcadas como correctamente cerradas, indicando que la mayoría de las preguntas quedan sin una respuesta satisfactoria..

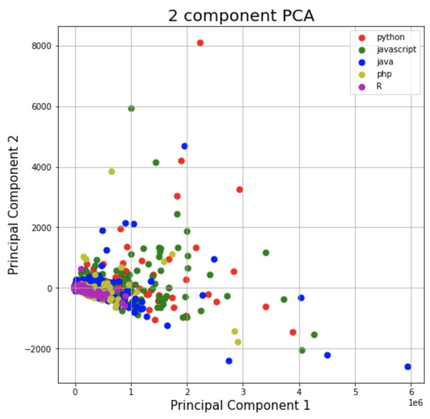
Estado	Conteo
FALSO	48373
VERDADERO	26961

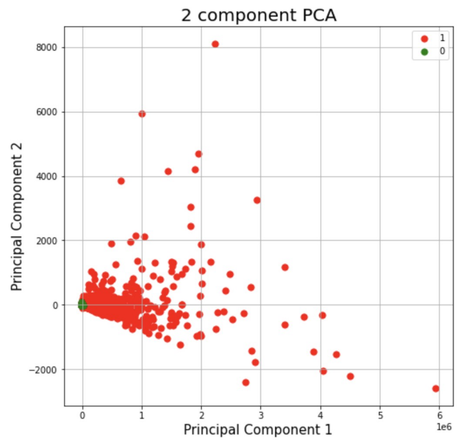


Con el fin de complementar el análisis univariado, se precede con un análisis de correlaciones entre las variables. Más precisamente, se procede con un análisis de componentes principales, esto una vez que la mayoría de variables en el conjunto de datos o son dicotómicas, o dicotomizables, o numéricas.







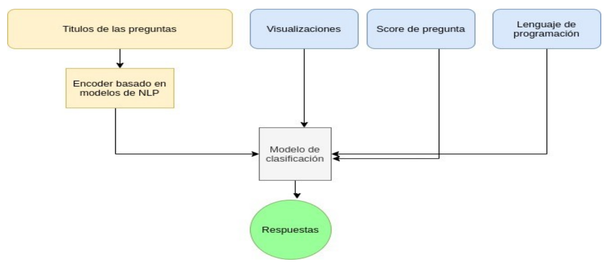


Metodología

En el presente proyecto el modelo propuesto para el análisis textual usará modelos de NLP para procesar y alimentar el clasificador que se va usar para el seguimiento de preguntas subidos a StackOverflow.

Con el objetivo de validar el mejor modelo, se propone utilizar las métricas de precision, recall y f1-score, una vez son métricas clásicas para validación de este tipo de modelos de Machine Learning.

Un bosquejo, inicial,de la arquitectura completa que se implementa se puede encontrar en la siguiente imagen.

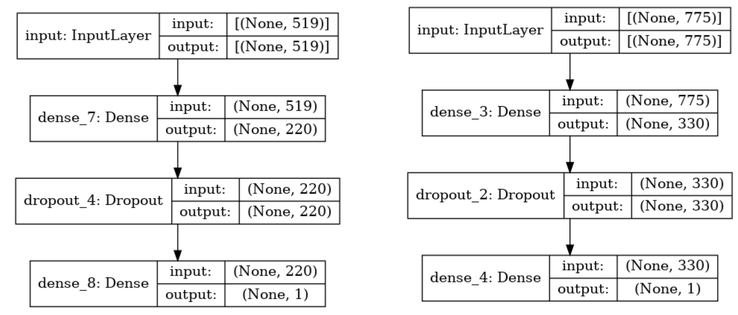


Resultados

En esta sección se muestran los resultados asociados a los modelos ajustados:

- (Modelo 1) Modelo de clasificación basado en el score y número de visualizaciones de la pregunta, sin incluir features textuales;
- (Modelo 2) Modelo de clasificación basado en el score y número de visualizaciones de la pregunta, se usa el modelo LaBSE- BERT como encoder para incluir los títulos de las preguntas;
- (Modelo 3) Modelo de clasificación basado en el score y número de visualizaciones de la pregunta, se usa el modelo Universal Sentence Encoder como encoder para incluir los títulos de las preguntas.

Un ejemplo de la arquitectura del clasificador es presentada en la siguiente figura.



Los resultados de los tres modelos son preentados a continuación. De izquierda a derecha se presntan los modelos 1, 2 y 3, respectivamente.

	precision	recall	f1-score	support
0	0.78	0.90	0.83	9645
1	0.75	0.54	0.63	5422
accuracy			0.77	15067
macro avg	0.76	0.72	0.73	15067
weighted avg	0.77	0.77	0.76	15067

	precision	recall	f1-score	support
0	0.80	0.89	0.84	9645
1	0.76	0.60	0.67	5422
accuracy			0.79	15067
macro avg	0.78	0.75	0.76	15067
weighted avg	0.79	0.79	0.78	15067

	precision	recall	f1-score	support
0	0.81	0.87	0.84	9645
1	0.74	0.64	0.69	5422
accuracy			0.79	15067
macro avg	0.77	0.76	0.76	15067
weighted avg	0.79	0.79	0.79	15067

Conclusiones

Una vez ajustados los tres modelos, se puede concluir, que:

- en términos del f1-score, precision y recall el mejor modelo es el Modelo de clasificación basado en el score y número de visualizaciones de la pregunta, se usa el modelo Universal Sentence Encoder como encoder para incluir los títulos de las preguntas.
- Es importante pensar en el principio de parsimonia, en cuyo caso el modelo más sencillo que no incluye features textuales debe ser considerado como un candidato a ser el modelo seleccionado, esto una vez que las métricas no difieren significativamente con respecto a aquellas de los otros dos modelos más complejos.
- Parece existir una correlación entre las preguntas asociadas al lenguaje R y las preguntas con menor probabilidad de ser marcadas como cerradas.

Referencias

[1] Sitio web: https://es.wikipedia.org/wiki/Stack_Overflow
[2] Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python, June 2009, Publisher(s): O'Reilly Media, Inc.
[3] Sitio web: https://www.kaggle.com/nazeboan/starter-svc-rf-rnn-and-lstm
[4]Sitioweb: https://www.kaggle.com/nazeboan/stackoverflow-questions-classification-challenge
[5] Sitio web: https://programmerclick.com/article/4622171728/
[6] Aaron Courville, Ian Goodfellow y Yoshua Bengio, Deep Learning, 2015.
[7] https://tfhub.dev/google/universal-sentence-encoder/1

Link del video

https://youtu.be/wZ8WxL8aFXA