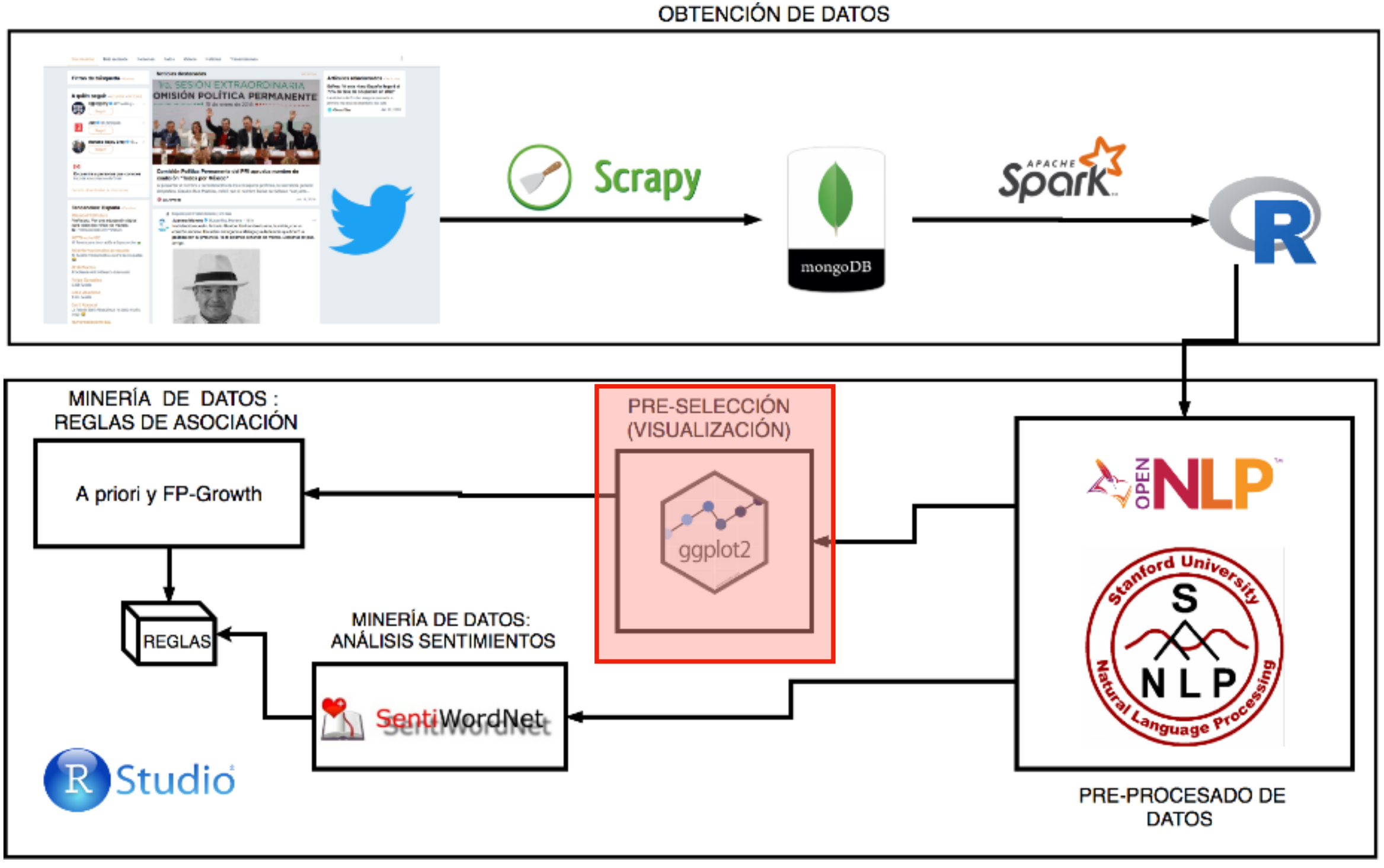
Metodología



Preprocesado: Selección de datos

Es necesario obtener la **matriz de términos. Problema:** Matriz muy grande imposible de mantener en memoria La solución: Eliminar datos. Obtenemos el histograma de palabras por tamaño y vemos algunas demasiado grandes. Eliminamos estas y las que aparecen menos de **30 veces**.