Topic modeling en datos de Twitter: Una aplicación en el contexto político peruano

Jesús Eduardo Gamboa Unsihuay[1]

[1] Universidad Nacional Agraria La Molina, Departamento Académico de Estadística e Informática, Lima 12, Perú
jgamboa@lamolina.edu.pe

Palabras Claves: minería de texto, modelamiento de temas, redes sociales, corpus, política

La minería de textos está compuesta por técnicas que permiten descubrir patrones en un conjunto grande de documentos. El primer paso de este análisis consiste en el procesamiento de los textos originales a fin de convertirlos en una matriz de términos de documentos, previa creación y limpieza del corpus; este paso puede ser desarrollado empleando el paquete tm. La siguiente etapa consiste en aplicar una técnica en particular: en esta investigación se presenta y aplica el modelamiento de temas, más conocido como Topic Modeling, el cual permite identificar los asuntos de los cuales trata un documento y se construye en base a dos principios: cada documento está estructurado según una mixtura de temas y cada tema es explicado mediante una mixtura de palabras. Para ello, emplea el modelo bayesiano de Alocación Latente de Dirichlet cuyos parámetros no pueden ser estimados analíticamente sino mediante el algoritmo EM con inferencia variacional o algoritmos MCMC, siendo posible fijar el número de temas de antemano u optimizar dicha cantidad; para esta tarea se hace uso de los paquetes topicmodels y LDAvis.

En las redes sociales, los usuarios y en especial los políticos exteriorizan sus pensamientos a través de sus publicaciones y de esa manera exponen sus prioridades a sus seguidores; en Twitter dichas publicaciones son conocidas como tweets. La aplicación de Topic Modeling se realiza en datos extraídos de cuentas de Twitter de políticos peruanos en el periodo de diciembre del 2017 a junio del 2018 con el fin de analizar los temas que tuvieron mayor relevancia en dicho periodo. La labor congresal en la comisión lavajato, los pedidos de vacancia presidencial, el indulto a Alberto Fujimori y las reacciones ante ello son algunos de los temas que pudieron ser identificados mediante el análisis de tweets, así también las palabras empleadas y los grupos políticos que dieron mayor relevancia a cada tema. Es posible constatar que tales temas presentan concordancia con los sucesos que se dieron en la realidad.