CURSO: CMP 0575 - TÓPICOS 2 (DATA MINING)
COLEGIO: POLITÉCNICO
Semestre: Primer Semestre 2019/2020

Proyecto 3: Ejercicio usando la técnica *TextMining*

Problema:

1. El 24 de marzo de 2019 El Ecuador celebró sus elecciones seccionales, en las cuáles se eligieron alcaldes, prefectos, miembros del consejo de participación ciudadana y control social, entre otros cargos públicos. Para estas elecciones se eligieron alrededor de 11000 autoridades de cerca de 81300 candidatos.

   En este ejercicio, su tarea consiste en analizar de una lista de candidatos (candidatos a alcaldes de Quito y Guayaquil, y los candidatos a miembros del consejo de participación ciudadana y control social) y de los tweets generados durante la campaña electoral que los mencionaban o que fueron generados por ellos, lo siguiente:

   a) Candidatos que más utilizaron la red social Twitter durante la campaña electoral.
   b) De los candidatos que más utilizaron la red social, determinar las palabras más utilizadas en campaña por ellos.
   c) Generar una línea de tiempo por candidato y sus ofertas de campaña principales en la red social (top 10 de términos). Ejemplo: si el candidato "PEPE" tiene asociado a su campaña los términos "violencia, robo, salario, etc" como los más importantes en su candidatura, entonces, el gráfico que se pide debe contener una serie (línea en el tiempo por día, semana o més) de uso por cada palabra importante (previamente determinada en el inciso b). ver Fig. 1.
   d) Establecer un ranking por candidato, atendiendo al *query* introducido por el usuario.

   Para dar solución a este ejercicio, se considerará lo siguiente:

   I. La colección de los candidatos (*account_info.csv*) y *tweets* relacionados a ellos (*livetweets_data.csv*) se puede descargar del siguiente link: https://estudusfqedu-my.sharepoint.com/:f:/g/personal/driofrioa_usfq_edu_ec/EqzuTVe0CsdDjRj5M6tbORkBZU5HSn9_1dTbFaoQxRLycg?e=MtduDo
   II. Es obligatorio mostrar la trazabilidad (pora el inciso b) de la técnica *TextMining* durante la ejecución del programa (mostrar una tabla con todos los índices de cálculo por términos: *tf, df, idf, tf-idf*).
   III. El criterio de similaridad debe ser basado en la métrica del coseno y sobre el espacio vectorial.

IV. Atendiendo a que se tienen los datos de los tweets de aproximadamente 4 meses. Sería factible visualizar la solución del inciso c) por semanas.

V. Cargar al D2L los códigos implementados (fichero compactado que incluye el ejecutable ej: el .JAR de java) dentro del plazo de entrega.

VI. +1

- Implementar una idea de *query* semántico ☺ (Ejemplo: el o los candidatos que prometieron aumentar el salario básico)

Nota: En cada fase de evaluación el profesor aplicará puntos de chequeo sobre el código implementado y basado en la trazabilidad. Además, los ficheros de análisis son pesados y deben ser descargados con la mayor brevedad posible.

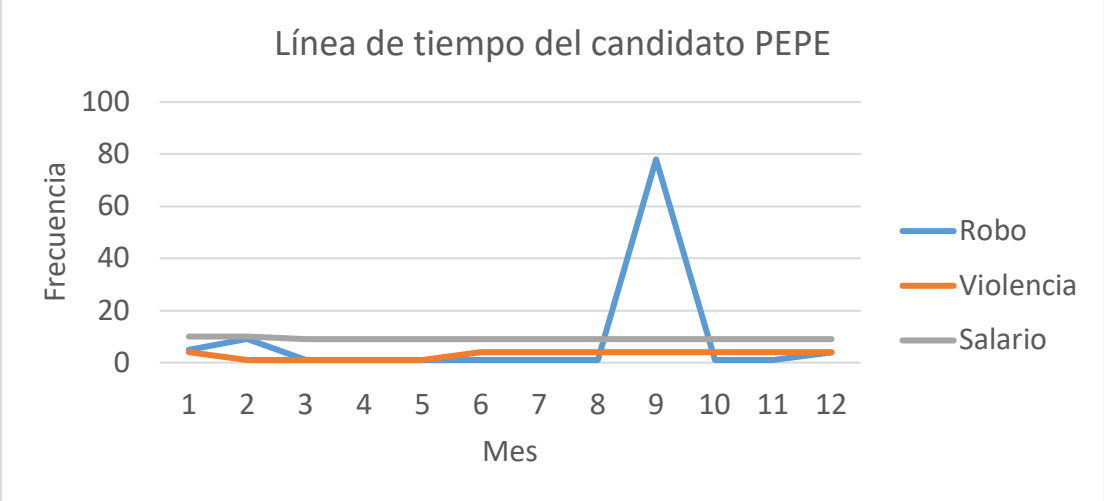

Fig. 1 Ejemplo de línea de tiempo por mes para el candidato hipotético "PEPE"