Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Tarea 2

Curso: Web Intelligence IN5526.
Semestre: Primavera 2017.
Entrega: 03/11/2017.

Profesor: Juan D. Velásquez.
Andrés Córdova.
Auxiliar: Valeria Lobos-Ossandón.

Contexto

Twitter es una red social de microblogging, en la que usuarios pueden publicar mensajes de hasta 140 caracteres, llamados tweets. En esta red, un usuario puede responder a un tweet de otro usuario, mencionarlo, o retwittearlo, publicando un mensaje idéntico a uno de otro usuario, dándole el crédito. También un usuario puede seguir a otro, de forma que los tweets que el seguido postea aparecen primero en el seguidor. De estas interacciones se desprende el concepto de influencia de un usuario en la red. A los usuarios les interesa ser seguidos por una mayor cantidad de usuarios y por usuarios más influyentes, aumentando la probabilidad de generar impacto en la red con sus opiniones.

Problema a resolver

El SENDA (Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol) lo contrata a usted para encontrar usuarios chilenos influyentes en el tema del consumo de marihuana en Twitter. El SENDA desea conocer la prevalencia del consumo de marihuana en Chile, por la cual le entrega a usted una base de datos con tweets concernientes a la marihuana, los usuarios que los emiten, y los seguidores de esos usuarios. Usted debe identificar a los usuarios más influyentes y los tópicos más relevantes relacionados al tema.
En específico, usted debe:

1. Construir un grafo de usuarios con los datos entregados usando el seguimiento de un usuario como arista.
2. Calcular el Pagerank de cada usuario. Muestre los primeros 20 más influyentes y los últimos 20 menos influyentes.
3. Comparar los pageranks obtenidos con la popularidad, es decir, el número de seguidores de los usuarios.
4. Construir un indicador con métricas basadas en el contenido del usuario, esto es, con:
    a. Número de tweets en el juego de datos.
    b. Número de retweets promedio.
    c. Respuestas y menciones.

    Comparar los pageranks obtenidos con los resultados de este nuevo indicador.
5. Por último, se requiere conocer los tópicos más relevantes en el consumo de marihuana. Para ello, realice un análisis LDA (Latent Dirichlet Allocation). Usted determine el número de tópicos y palabras por tópicos relevantes.

Entregables

1. Informe autocontenido que detalle claramente el trabajo realizado. Piense que su objetivo es entregar este informe a la alta dirección del SENDA. Este informe debe contener como mínimo: Resumen Ejecutivo, Introducción, Desarrollo, Resultados, Análisis y Conclusiones
2. Scripts, modelos, etc, que permitan reproducir los resultados del informe.