Text Mining en R para Ciencias Sociales
Estación Lastarria – 2023

Contexto general

La explosión de datos digitales ha llevado a un creciente interés en el análisis de texto, el cual puede ser de gran utilidad para las ciencias sociales debido a que gran parte de su investigación se basa en el estudio de documentos, encuestas y otros datos textuales. El text mining (minería de texto) permite a los investigadores explorar grandes cantidades de información textuales y extraer conocimientos valiosos y significativos utilizando métodos estadísticos y de aprendizaje automático, como la identificación de patrones y tendencias en los datos.

Con la disponibilidad de herramientas de software como R y RStudio, el análisis de texto se ha vuelto más accesible para los investigadores y profesionales. Estas herramientas ofrecen una amplia gama de paquetes de análisis de texto que permiten a los usuarios procesar y analizar grandes cantidades de datos textuales de manera eficiente y efectiva.

Este curso está diseñado para brindar las habilidades necesarias para utilizar técnicas avanzadas de text mining en R. Los participantes aprenderán a utilizar diversas técnicas de análisis de texto, desde la representación de palabras hasta el modelado de tópicos, y obtendrán una comprensión más profunda de cómo el procesamiento del lenguaje natural (NLP) puede ser aplicado en el análisis de datos en las ciencias sociales

Descripción

En este curso de 5 sesiones se enseñará a los estudiantes los fundamentos del procesamiento del lenguaje natural y técnicas avanzadas de minería de texto. Se cubrirán técnicas como TF-IDF, análisis de sentimientos, modelos de tópicos como Latent Dirichlet Allocation (LDA) y Structural Topic Model (STM), y técnicas de word embedding como Word2vec y glove. Los participantes obtendrán habilidades en la manipulación, visualización y análisis de datos de texto utilizando R y RStudio.

Responsable

Ignacio Toledo

Ingeniero Civil Electrónico, Doctorado (Ph.D.) en Ciencias de la Complejidad Social de la Universidad del Desarrollo y una Maestría (M.Sc.) en Ciencias de la Ingeniería con mención en Ingeniería Eléctrica de la Universidad de Concepción.

Ha desempeñado roles de analista de datos en distintos contextos, tanto en la industria como en la academia. Anteriormente, se desempeñó como científico de datos y coordinador de analítica en los Programas TIDEM y RedBios, liderados por la Facultad de Diseño de la Universidad del Desarrollo (UDD) y financiados por el Gobierno Regional del Biobío.

Actualmente, ejerce como investigador en la Facultad de Diseño de la Universidad del Desarrollo. Su investigación se distingue por el uso de métodos provenientes de las ciencias sociales computacionales para el estudio de ecosistemas regionales de innovación, la gestión del diseño, el diseño e innovación sostenible, y el diseño sistémico.

Público objetivo

Estudiantes de pregrado y posgrado, y profesionales en ciencias sociales interesados en aprender técnicas avanzadas de análisis de texto

Requisitos mínimos

- Conocimientos básicos en estadística y programación en R y RStudio.
- Se requiere un conocimiento general de metodología y técnicas de investigación.
- Cada estudiante deberá contar con un computador portátil operativo.
- Es deseable un manejo intermedio del idioma inglés.

Objetivos

Objetivo general:

Enseñar a los participantes técnicas avanzadas de análisis de texto utilizando R y RStudio.

Objetivos específicos:

1. Comprender los fundamentos del procesamiento del lenguaje natural (NLP).
2. Familiarizarse con técnicas de minería de texto avanzadas como TF-IDF, análisis de sentimientos, modelos de tópicos y word embedding.
3. Desarrollar competencias en el uso de R y RStudio para manipular, analizar y visualizar datos de texto.
4. Proporcionar a los participantes la habilidad para implementar técnicas de minería de texto en situaciones reales.

Metodología

El curso se basará en la metodología de enseñanza activa, donde los estudiantes aprenderán a través de la práctica y la resolución de problemas reales. Se realizarán talleres prácticos y se asignarán ejercicios para completar fuera del aula.

Desarrollo de talleres

Se llevarán a cabo 5 talleres prácticos, cada uno de 2,5 horas de duración. Los talleres se centrarán en la enseñanza de técnicas de text mining utilizando R y RStudio, y se aplicarán en procesamiento, análisis y modelamiento de datos textuales.

Contenidos por sesión

Sesión[1]	Contenidos	Fecha
1. Introducción a la minería de texto y R.	● Introducción a técnicas de análisis de texto en R. ● Preprocesamiento de texto: tokenización, limpieza, stemming y lematización. ● Modelos de representación de texto con bag of words y TF-IDF. ● Análisis de sentimientos	22-11-2023
2. Fundamentos de modelos de tópicos	● Modelos de tópicos probabilísticos ● Latent Dirichlet Allocation (LDA) ● Structural Topic Model (STM)	29-11-2023
3. Modelamiento de texto con modelos de tópicos	● Evaluación de modelos y selección de número de tópicos ● Especificación de modelos e interpretación de resultados ● Visualización de tópicos con LDAvis y stmBrowser	06-12-2023
4. Fundamentos de word embedding	● Modelos de Word Embedding ● Word to vector (Word2vec) ● Gloval Vectors (GloVe)	13-12-2023
5. Modelamiento de texto con word embedding	● Análisis semántico: Similaridad y analogía entre palabras. ● Visualización de vectores de palabras: PCA (Análisis de componentes principales) y t-SNE (t-Stochastic Neighbor Embedding). ● Herramientas y recursos adicionales para la minería de texto en R. ● Discusión de proyectos de los participantes y preguntas frecuentes	20-12-2023

[1] La duración de cada sesión son 2,5 horas cronológicas.

Bibliografía sugerida

- Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".
- Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, December). The structural topic model and applied social science. In Advances in neural information processing systems workshop on topic models: computation, application, and evaluation (Vol. 4, pp. 1-20).
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).