Programa del seminario:
Análisis cuantitativo de textos

Valor académico: 1.5 UMA,s (22.50 horas presenciales)

Profesor: Javier Brolo (javierbrolo@gmail.com)

Descripción del curso:

La abundancia y disponibilidad de textos tiene el potencial de revelar invaluable información sobre las posiciones políticas de sus autores; por ejemplo mediante el análisis de discursos públicos, declaraciones a medios, comentarios en redes sociales, columnas de opinión, estatutos partidarios, entre otros. Sin embargo, las limitaciones de tiempo y recursos reducen la capacidad para considerar sistemáticamente grandes cantidades de contenido en textos a la vez.

Una forma de enfrentar la dificultad de evaluar grandes cantidades de contenido es automatizando los procesos con la ayuda de computadoras. Por ello, el seminario "Análisis cuantitativo de textos", ofrece una introducción a la extracción de información y análisis sistemático de textos utilizando el software estadístico R y el paquete quanteda (Benoit y Nulty, 2016).

El seminario tiene un enfoque práctico. Al finalizar, el estudiante contará con capacidad de identificar textos relevantes, extraer sus características y utilizar técnicas estadísticas para analizarles. Entre las técnicas a utilizar destacan: contar palabras, calcular índices, construir diccionarios, clasificar textos y ajustar escalas ideológicas. La metodología del seminario consistirá en presentaciones del profesor para exponer los conceptos y técnicas a utilizar seguidas de ejercicios aplicados. También, se discutirán diseños de investigaciones emblemáticas como: "Análisis computarizado de transcripciones de textos de Al-Qaeda" (Pennebaker y Chung, 2007); "Medición de populismo, comparando dos técnicas de análisis de contenido" (Rooduijn y Pauwels, 2011); "Extracción de posiciones ideológicas de textos políticos utilizando palabras como datos" (Laver, Benoit y Garry, 2003); y "Aplicación de análisis automático de contenido para mejorar la investigación legal empírica" (Evans, McIntosh, Lin y Cates, 2007).

El seminario constará de 17 sesiones a impartirse en la fecha y horarios acordados por medio de la plataforma en línea WebEx. Previa solicitud del estudiante, los días acordados de 11:00AM a 12:00PM estarán disponibles para atender dudas por medio de video-llamada; alternativamente, se atenderán dudas por correo electrónico. Aquellos estudiantes con requerimientos especiales deberán hacerlos ver con anticipación.

Objetivos del curso:

1. **General:**
   - Comprender la relación entre textos y las posiciones políticas de sus autores
   - Identificar y obtener textos relevantes para el análisis político
   - Extraer características de textos y transformarlos en matrices
   - Utilizar técnicas estadísitcas para analizar textos

2. **Específicos:**
   - Utilizar el paquete estadístico R y quanteda para analizar textos
   - Manejar distintos formatos de texto y los caracteres especiales del español
   - Calcular estadísticas descriptivas de los textos
   - Analizar textos mediante diccionarios
   - Clasificar textos utilizando "Bayes ingenuo"
   - Ajustar escalas ideológicas de textos utilizando "Wordscores"

Prerrequisitos:

Los estudiantes deben haber completado el curso de estadística y se recomienda que hayan recibido el curso de métodos y técnicas cuantitativas de investigación. Adicionalmente, deben contar con acceso a una computadora y tener instalado el software para análisis estadístico R, su interfase R-Studio, y Sublime Text (Windows) o TextMate (Mac). Tomar en cuenta que varias de las lecturas serán en inglés.

Perfil del estudiante al finalizar el curso:

Al finalizar el curso, el estudiante:

- Comprenderá el modelo de "bolsa de palabras" para analizar textos
- Contará con capacidad de extraer información de textos con la ayuda de la computadora
- Sabrá utilizar técnicas estadísticas para analizar textos
- Conocerá diseños emblemáticos de investigaciones académicas que analizan texto

Metodología del curso:

El seminario constará de 17 sesiones a impartirse en la fecha y horarios acordados por medio de la plataforma en línea WebEx. Cada sesión consistirá en presentaciones del profesor para exponer los conceptos y técnicas a utilizar seguido de ejercicios prácticos. La evaluación del seminario contempla participación, ejercicios prácticos, trabajo escrito, presentación y examen final escrito.

Programa en detalle:

METAS DE APRENDIZAJE	CONTENIDO	NÚMERO DE SESIÓNES	ACTIVIDADES	FUENTES/LECTURAS
Presentar la logística y objetivos del curso	R; R-Studio; Quanteda; Sublime Text; TextMate	1	Instalar software a utilizar durante el curso	Programa del curso R Studio Tutorial
Seleccionar textos, definir sus características y mediciones	Unidades de análisis, conceptos básicos	1	Recolección de textos	Getting Started with quanteda Alonso, S., Volkens, A., & Gómez, B. (2012). Capítulo 1 (p. 11-44)
Contar características de textos	Frecuencia de palabras, Ley de Zipf,	1	Verificar Ley de Zipf	Manning, C. D., Raghvan, P., & Schütze, H. (2009). Capítulo 5.1.2 (p. 89-90)
Modelar la generación de texto	Selección de muestra, ponderación (tf-idf), truncar, bigramas	1	Calcular tf-idf	Manning, C. D., Raghvan, P., & Schütze, H. (2009). Capítulo 6.2 (p. 117-120)
Describir textos	Palabras en contexto, resúmenes descriptivos	1	Interpretar palabras en contexto; crear nubes de palabras	Neuendorf, K. A. (2002). Chapter 3 (p. 47-54)
Resumir textos	Escalas, índices de lectura y diversidad de léxico	1	Calcular diversos índices	Roberto, J. A., Martí, M. A., & Salamó, M. (2012). Capítulo 3.1 (p. 99-100)

METAS DE APRENDIZAJE	CONTENIDO	NÚMERO DE SESIÓNES	ACTIVIDADES	FUENTES/LECTURAS
Analizar documentos como vectores	Mediciones de similitud entre textos	1	Calcular coeficientes de similitud	Manning, C. D., Raghvan, P., & Schütze, H. (2009). Capítulo 6.3 (p. 120-128)
Identificar "colocaciones"	Colocaciones	1	Utilizar la prueba Chi-cuadrado	Sánchez R, M. Á. (2005). (p. 73-74)
Medir nivel de incertidumbre	"Bootstrapping" (remuestreo)	1	Realizar un remuestreo	Ledesma, R. (2008). (p. 52-54)
Utilizar diccionarios	Diccionarios	1	Analizar sentimiento de textos	Rooduijn, M., & Pauwels, T. (2011).
Construir diccionario propio	Diccionarios	1	Construir un diccionario propio	LIWC Pennebaker, J. W., & Chung, C. K. (2007).
Clasificación de textos	Bayes ingenuo; k-NN; SVM	1	Decidir si un texto es de determinado tipo	Manning, C. D., Raghvan, P., & Schütze, H. (2009). Capítulo 13 (p. 253-287)
Evaluar rendimiento de un método de clasificación	Precisión, discriminación, exactitud, F1	1	Determinar rendimiento de una clasificación	Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007) Capítulo II (p. 1010-1013)
Determinar estabilidad, reproducibilidad y precisión de una codificación	Confiabilidad	1	Determinar la confiabilidad de una clasificación	Krippendorff, K. (2004). Capítulo 11 (p. 211-235)
Emplear técnicas supervisadas para clasificar textos	Wordscores	1	Determinar "ideología" a partir de textos	Laver, M., Benoit, K., & Garry, J. (2003).
Mostrar aprendizaje	Presentaciones de estudiantes	1	Presentaciones de estudiantes	Provisto por estudiantes
Mostrar aprendizaje	Material del curso	1	Evaluación escrita	
		17		

Calificaciones:

Los puntajes quedan definidos de la siguiente forma:

ESTRUCTURA DE LA ZONA	
Participación y asistencia	15 puntos
Laboratorios (2 puntos cada uno)	24 puntos
Trabajo escrito	20 puntos
Presentación	16 puntos
PRUEBA FINAL	
Examen final	25 puntos
TOTAL	100 puntos

Honestidad:

Se requiere absoluta honestidad académica por parte de cada alumno, tanto en términos de exámenes como de trabajos de investigación. **Cualquier sospecha de copia o plagio será tratada severamente de acuerdo al reglamento de la UFM**.

Bibliografía:

**** Básica:***

Alonso, S., Volkens, A., & Gómez, B. (2012). *Análisis de contenido de textos políticos. Un enfoque cuantitativo*. Madrid: Centro de Investigaciones Sociológicas.

Benoit, K., & Nulty, P. (n.d.). Getting Started with quanteda. Retrieved October 1, 2016, from https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

Kenneth Benoit and Paul Nulty (2016). quanteda: Quantitative Analysis of Textual Data. R package version 0.9.8. https://CRAN.R-project.org/package=quanteda

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039.

Krippendorff, K. (2004). *Content Analysis. An Introduction to Its Methodology* (Second). London: Sage Publications.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *Source: The American Political Science Review*, *97*(2), 311–331.

Ledesma, R. (2008). Introduccción al Bootstrap. *Tutorials in Quantitative Methods for Psychology*, *4*(2).

Manning, C. D., Raghvan, P., & Schütze, H. (2009). *Introduction to Information Retrieval* (Online). Cambridge: Cambridge University Press.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

Pennebaker, J. W., & Chung, C. K. (2007). Computerized text analysis of Al-Qaeda transcripts. In K. Krippendorff (Ed.), *A content analysis reader*. Thousand Oaks, CA: Sage Publications.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roberto, J. A., Martí, M. A., & Salamó, M. (2012). Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, *48*, 97–104.

Rooduijn, M., & Pauwels, T. (2011). Measuring Populism: Comparing Two Methods of Content Analysis. *West European Politics*, *34*(6), 1272–1283.

Sánchez R, M. Á. (2005). Uso metodológico de las tablas de contingencia en la Ciencia Política. *Espacios Públicos*, *8*(16), 60–84.

** Complementaria:*

Bunea, A., & Ibenskas, R. (2015). Quantitative Text Analysis and the Study of EU Lobbying and Interest Groups. *European Union Politics*, *16*(3), 429–455.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Serial Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722.

Statsoft. (2016). Naive Bayes Classifier. Retrieved October 1, 2016, from http://www.statsoft.com/Textbook/Naive-Bayes-Classifier

Laver, M., & Garry, J. (2000). Estimating Policy Positions from Political Texts. *American Journal of Political Science*, *44*(3), 619–634.

ACTUALIZADO: OCTUBRE 2016