Ludovic Lebart, André Salem
Analyse Statistique de Données Textuelles
Dunod, 1988
Prólogo de Christian Baudelot.

En este libro los autores han querido resumir una experiencia de casi diez años de tratamiento estadístico de datos textuales y presentar una síntesis de los métodos de análisis elaborados.

Ludovic Lebart es Doctor en Estadística y Director de Investigaciones del CNRS. Tiene una larga experiencia en la concepción, el diseño y la elaboración de instrumentos estadísticos e informáticos para el análisis de grandes encuestas. Citemos, en particular, la concepción del sistema «Étude des Aspirations des Français» y la práctica del mismo durante 10 años. Es uno de los conceptores de los sistemas informáticos de Análisis de Datos SPAD, SPAD.N y del sistema informático de análisis de datos textuales SPAD.T, compatible con los anteriores.

André Salem es Doctor en Estadística e Investigador del Centre de Lexicométrie Politique de l'Université de Paris III. En el marco de este laboratorio, ha participado en múltiples estudios sobre textos políticos e históricos y ha elaborado nuevas técnicas de tratamiento estadístico de datos textuales. Es uno de los conceptores del sistema informático LEXICLOUD.

En el libro reseñado aquí, se dirigen a todo tipo de investigadores que necesiten afrontar el tratamiento y el análisis de una información de tipo textual, ofreciendo una presentación de las herramientas matemáticas, orientada a lectores no especialistas.

Entre los métodos de análisis presentados, los de análisis estadístico multidimensional ocupan un lugar privilegiado. Los autores muestran como estos métodos, construidos y probados para el tratamiento exploratorio de grandes tablas de datos, constituyen una herramienta poderosa para el análisis de una información textual y para el análisis de una información mixta, textual y cualitativa, sobre numerosos individuos. Se consideran también métodos más propios del campo textual para la comparación de diferentes textos. Todos estos métodos son considerados elementos complementarios de una herramienta global.

El primer capítulo está consagrado a los datos de encuesta. Las respuestas a preguntas cerradas y abiertas de numerosos individuos constituyen, en efecto, el ejemplo-tipo de datos a los cuales se pueden aplicar estos métodos. Se recuerda la aportación específica de las respuestas abiertas y se muestra la im-

posibilidad de sustituirlas por respuestas cerradas. A continuación, se explicitan las unidades estadísticas consideradas: la forma gráfica o grafema (cada forma llena derivada de una misma palabra se considera como una forma gráfica distinta) y los segmentos repetidos.

El segundo capítulo presenta los métodos lexicométricos y la terminología empleada en el estudio cuantitativo de los textos. Se comentan los documentos lexicométricos más habituales: índices, concordancias, comparación de vocabulario.

Las técnicas de análisis de datos, como técnicas de tratamiento de datos de encuestas, se exponen en el capítulo 3. Después de una breve exposición de los principios básicos, se comentan las técnicas de análisis de correspondencias simples, las de correspondencias múltiples y, finalmente, los métodos de clasificación automática. Los métodos se presentan con la ayuda de pequeños ejemplos, lo que permite entender la elaboración de los resultados estadísticos obtenidos y las reglas de interpretación.

Los capítulos 4 y 5 tratan del análisis estadístico de datos textuales propiamente dicho. El capítulo 4 estudia el análisis de las tablas léxicas. La tabla léxica o tabla Individuos*Formas contiene la frecuencia con la cual cada individuo emplea cada forma. La tabla léxica agregada contiene la frecuencia con la cual cada grupo de individuos —grupo formado según criterios apropiados al problema estudiado— emplea cada una de las formas. Son, por lo tanto, tablas de contingencia particulares a las cuales se pueden aplicar los métodos de análisis de correspondencias y de clasificación. La aplicación de los métodos a datos textuales se explica mediante ejemplos de aplicación a datos reales, lo que permite entender el interés de los resultados obtenidos.

El capítulo 5 está consagrado al tratamiento de la unidad estadística «Segmento repetido». Esta nueva unidad de segmentación del corpus, introducida por A. Salem, permite contextualizar las formas y completa los estudios anteriores. En efecto, se pueden construir tablas segmentales o tabla de frecuencia de uso de los segmentos por los individuos o por los grupos de individuos, tablas a las cuales se aplican los mismos métodos.

Finalmente, los autores inician una reflexión sobre el alcance de estos métodos, la significación de los resultados obtenidos y la relación que puede haber entre las diferencias de formas y las diferencias de contenido. En anexo, se presentan los sistemas informáticos LEXICLOUD y SPAD.T.

Este libro será muy útil y apreciado por los investigadores en ciencias sociales. Presenta, en efecto, de manera muy clara una herramienta novedosa para el análisis y el tratamiento comparativo de textos. Dicha herramienta, desarrollada por los autores, abre nuevas vías de investigación sobre este material tan rico que constituyen los textos.

MÓNICA BÉCUE