Text mining se originó por la necesidad de catalogar textos, por ejemplo libros en una biblioteca. Y luego se fue moviendo hacia extracción de datos utilizando técnicas de NLP.

Incluso en 1999 se notaba que no tenía casi ningún practicante, principalmente porque la definición común de text mining excluía Information Retrieval.

A partir de 1999 investigación y aplicaciones comenzaron a desarrollarse rápidamente.

Sus orígenes se remontan a la necesidad de acceder a información enterrada entre vastas cantidades de texto que está almacenada en distintos formatos. En base a esto existían dos procesos: information retrieval, que encuentra un set de documentos como respuesta a un query con set de key words; e information extraction, que busca extraer información específica de documentos y luego analizarla por tendencias y otros datos.

Para este problema se necesitaba algún tipo de proceso de resumen para reducir el tamaño del texto a algo utilizable, y luego clasificarlos.

Existían tres métodos para abordar esto:

• Library science for text summarization and classification

Uno de los primeros ejemplos de resumen y clasificación fue el catálogo de bibliotecas, atribuido por primera vez a Thomas Hyde en 1674 para la Bodelian Library en la Universidad de Oxford. Luego las tarjetas de índices fueron introducidas por Melvil Dewey en 1876 para formar el catálogo de tarjetas de biblioteca.

Otro paso fue el resumen de texto para generar abstracts, que empezó en 1898 con una colaboración entre la Institution of Electrical Engineers y la Physical Society of London.

En 1958 Luhn usó una computadora para generar abstracts, a partir de Word Frequency Analysis. Desde su inicio text mining para information retrieval y manipulación estuvo centrado en library science.

• Information science

• Natural language processing

Estoy siguiendo este documento:

<https://learning.oreilly.com/library/view/practical-text-mining/9780123869791/xhtml/CHP001.html>