Analytics of text data

How text data are generated?

Text data are generated whenever language—written, spoken, or digital—is created or captured in a machine-readable form. They originate from multiple sources, including human communication such as social media posts, news articles, and emails; machine-generated content like chatbot responses, system logs, and automated reports; and digitized materials produced through technologies like optical character recognition (OCR) and speech-to-text conversion. Additionally, text data can be extracted from the web or APIs for analysis. Altogether, these sources provide vast and diverse collections of text that form the foundation for modern text mining and natural language processing applications.

Topics

Text data encompass all forms of language captured or produced in digital form, making them a crucial foundation for data science and natural language processing. They originate from various sources, including human-generated content such as social media posts, news articles, and online reviews; machine-generated text like chatbots, reports, and system logs; and digitized materials obtained through OCR or transcription of printed or spoken language. Researchers and analysts often collect these data using web scraping, APIs, or other automated methods. Once gathered, text data are preprocessed and analyzed through techniques such as tokenization, sentiment analysis, and topic modeling to uncover patterns, trends, and meanings within language use.

Sentiments

Sentiments refer to the emotions, opinions, or attitudes expressed in text data. In text analysis, sentiment analysis (or opinion mining) is the process of identifying and categorizing these emotional tones—typically as positive, negative, or neutral—based on the words and context used. Sentiments are often derived from written sources such as product reviews, tweets, comments, or survey responses, providing insight into public opinion, customer satisfaction, or social trends. By applying natural language processing (NLP) and machine learning techniques, researchers can detect not only overall polarity but also more nuanced emotions like joy, anger, fear, or sadness. Sentiment analysis thus helps organizations and researchers understand how people feel about products, events, policies, or experiences on a scale.

Positions and polarity

In sentiment analysis, positions and polarity describe how opinions are expressed and distributed within text data. Polarity indicates the emotional orientation of a statement—whether it conveys a positive, negative, or neutral attitude—and may also reflect the intensity of sentiment on a scale. Position refers to where these sentiments occur in the text, such as at the beginning or end of a sentence or paragraph, revealing how opinions may shift or contrast throughout the message. Together, analyzing polarity and position helps researchers capture not only the overall tone but also the structure and progression of sentiments in written communication.

Prepare a list of text data useful for your team project (e.g. web, government documents, news).

For our project, several text-based data sources will support the analysis of crime, climate, and socioeconomic patterns across the United States. FBI Uniform Crime Reporting (UCR) documentation provides narrative explanations of crime definitions, reporting practices, and year-to-year changes, helping us interpret crime trends more accurately. NOAA climate summaries and technical reports offer written descriptions of temperature anomalies, weather events, and methodological notes, which are useful for understanding environmental influences on crime. U.S. Census Bureau ACS documentation includes definitions, sampling details, and economic explanations that help contextualize income and demographic indicators used in the analysis. Finally, government and academic research reports on crime, climate, and social behavior provide background theories and previously established findings that support our project's interpretive framework.