TDM Glossary

Application Programming Interface (API)
A set of protocols and tools that allows different software applications to communicate with each other, facilitating access to data and services in a machine-readable format.

Corpus
A (text) corpus is a collection of written or spoken material, such as journal articles, or any other documents, used for linguistic analysis or research.

Crawling
An automated process used to discover and follow links within a website to extract information from web pages.

Entity
Refers to a distinct, identifiable real-world object or concept, such as a name, organization, or event (e.g., "cat" as an animal).

Extensible Markup Language (XML)
A web standard for document mark up, designed to simplify and provide flexibility to Web or other digital media authorship and design. It is not a fixed-format language, in contrast to HTML, for example. Hypertext Mark-up Language (HTML)

Hypertext Markup Language (HTML)
This is a text based coding language, interpreted by web browsers and used to construct web pages.

Information Extraction
The process of automatically identifying and extracting specific pieces of information or data from unstructured text.

Machine Learning
A subset of artificial intelligence that employs algorithms and statistical models to analyze and learn from data, enabling systems to identify patterns and make decisions with minimal human intervention.
In essence, mathematical and statistical methods (algorithms) that automatically identify patterns in data. The “learning” is the finding of those patterns.

Natural Language Processing (NLP) Tools
Software systems or services facilitating the automatic analysis of text, e.g. named entity extraction

Ontology
The organization of a specific domain with the entities that belong in it and their relationships. So, for example a domain could be “genes” or “chemistry”, the entity could be a specific gene e.g. 1245 and all the forms that it might show up in a text. This is how the human genome has been mapped.

Linguistic Parsing

Linguistic parsing refers to the process of (syntactic) analysis of text, i.e. identifying how a sentence follows the grammatical rules of a language. It breaks down a unit/sentence into its component parts. You can also parse files into their component parts

Relationship Extraction

Process of automatically finding relationships between two (or more) entities within a text (semantic relationship), e.g. A cat sits on a mat.

Semantic Relationship

A linguistic relationship between two or more entities so that machines can understand that relationship, e.g. “is_a” as in a cat is an animal

Sentiment Analysis

The computational study of opinions, sentiments, and emotions expressed in text, identifying positive, negative, or neutral sentiments based on specific words or phrases.

Scraping

The automated technique of extracting data from websites by visiting the site and copying the information for use elsewhere.

Taxonomy

A structured classification system that organizes concepts or terms in a hierarchical manner, often enriched with synonyms and relationships (e.g., "cat" as a "feline" and "mammal").

Text and Data Mining (TDM)

Text mining is the data analysis of natural language works (articles, books, etc.), using text as a form of data. It is often joined with data mining, the numeric analysis of data works (like filings and reports), and referred to as "text and data mining" or, simply, "TDM." TDM depends on the assembly of a working set of data/content against which an analytic process is run. This process breaks down digital information into raw data and text, analyses it, and comes up with new connections, from unexpected patterns. This can eventually lead to the development of a new drug, to subtle shifts in weather patterns that might predict a downturn in the price of wheat.

Treebank

A linguistic resource consisting of a corpus of texts that have been syntactically parsed, used primarily for training and evaluating natural language processing models.