Glossary

Text And Data Mining

Name	Definition
Application Programming Interface (API)	The technical window/programming language interface through which users can access and obtain vast quantities of information (text/data/objects) in a machine-readable format.
Corpus	A (text) corpus is a collection of documents, e.g. web pages, journal articles.
Crawling	It is an automatic method to find and follow links within a website, in order to scrape the information.
Entity	Refers to a real world thing (e.g. a name such as cat).
Extensible Mark-up 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)	This is a text based coding language, interpreted by web browsers and used to construct web pages.
Information Extraction	Automatically isolating specific words or information from unstructured text.
Machine Learning	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.
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 extraction of words or phrases which convey emotional meaning. For example, the sentence "The chicken curry was salty and overpriced" indicates a negative sentiment as "overpriced" has an emotional meaning.
Scraping	Scraping information is an automatic way to visit a website, copy and paste the information somewhere else.
Taxonomy	A controlled vocabulary organised in a hierarchical manner, or enriched with synonyms and non-hierarchical relationships e.g. cat is a feline, is a mammal, etc.
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	This is as corpus of syntactically parsed documents used to train TDM models.