Text Mining In A Nutshell

This is a gentle introduction to the why's and how's of text mining—text mining in a nutshell.

Van Gogh's *Starry Night*

At its core, text mining is about discovering patterns and anomalies in sets of written documents. Invariably, the process begins with the creation of a digitized corpus of materials. The content of the corpus may then be cleaned up, marked up, and organized, thus making it easier for computers to read and parse. From there "tokens"—usually words—are identified, counted, and tabulated. These techniques—usually employed under the rubric of "natural language processing"—form the basis for more sophisticated applications.

Simple examples

Some simple examples of text mining include:

- **frequency counts of words beginning with specific letters** - By creating an alphabetical list of all the words in a document along with their counts, it is relatively easy to look up a specific word and see how its variations are used.
- **frequency counts of common words and phrases** - Ordering the list of words in a document (or n-length phrases) by their frequency allows the reader to get a sense of what words are mentioned most often in a corpus, or just as importantly, what is not mentioned. These sorts of frequency counts are the basis of visualizations known as word- or tag-clouds.
- **concordances** - These are the world's oldest keyword-in-context indexes. For any given word or phrase, display n number of words surrounding the given word to see how it is used in a particular circumstance.

More sophisticated applications

Textual documents are expressions of written language, and written language follows sets of loosely formulated rules. (If there were no "rules", then none of us would be able to understand each other.) Taking English as an example, all sentences must begin with a capital letter and end with a punctuation mark. Likewise, proper nouns are expected to be capitalized. In general, adverbs end in "ly", and many gerunds end in "ing". Based on these and other sorts of rules, more sophisticated applications can be created, such as:

- **parts-of-speech (POS) analysis** - English is made up of nouns, verbs, adjectives, etc. POS tools allow the reader to count and tabulate these occurrences and then make evaluations. Is this text more masculine or feminine? Is it more singular or plural? Are there a preponderance of action verbs? What about colors? Are the sentences long or short?
- **named entity extraction** - Nouns can be subdivided into different types: names, places, dates, times, etc. Identifying and tabulating these particular types help to characterize a text. It can also help to "read a text from a 'distance'" as well as answer some basic questions. When does action take place? Where does action take place? Who are the major (or minor) characters in the text. The answers to these questions can then be linked with things like bibliographic indexes, image

databases, atlases, gazetteers, or other textual documents to enhance the reading process.

- **sentiment analysis** - Similar to named entity extraction, the “feeling” or subject content of a text can be determined through text mining and analysis. The result of such a process may classify a document as expressing happiness, sadness, anger, etc. Think how this could be applied to movie or book reviews. Think how it could be applied to children’s literature or Gothic novels.

- **summarization** - This process is used to generate brief statements outlining the substance of a text. This is usually done through stylistic analysis such a looking for sentences starting with phrases such as “In conclusion”, “In summary”, or headings starting with “Abstract”.

- **classification and clustering** - These processes are used to divide a large set of documents (or maybe even sets of paragraphs) into a number of smaller sets. Classification—a kin to the library cataloging process — assumes the existence of broad categories where documents are expected to be placed. Clustering creates sets of broader categories based the content of the documents.

Tools

On the Web there exist at least a few of directories of text mining (digital humanities) tools. These tools come in a myriad of formats and flavors for both a number of different computers as well as on the Web:

- **Bamboo Dirt** (http://dirt.projectbamboo.org) - “Bamboo DiRT is a tool, service, and collection registry of digital research tools for scholarly use... and makes it easy for digital humanists and others conducting digital research to find and compare resources ranging from content management systems to music OCR, statistical analysis packages to mindmapping software.”

- **Hermeneuti.ca** (http://hermeneuti.ca/voyeur/tools) - A collection of text mining tools accompanied by sets of text (an online book) describing the how’s and why’s of text mining in modern scholarship.

- **TAPoRware** (http://taporware.ualberta.ca) - “TAPoRware is a set of text analysis tools that enables users to perform text analysis on HTML, XML and plain text files, using documents from the users’ machine or on the web.”

If there were only one suggested text mining tool, the I would suggest the following, which is really a collections of tools built into a single Web interface:

- **Voyant Tools** (http://voyant-tools.org) - “Voyant Tools is a web-based text analysis environment. It is designed to be user-friendly, flexible and powerful. Voyant Tools is part of the Hermeneuti.ca, a collaborative project to develop and theorize text analysis tools and text analysis rhetoric.”

Text mining and the humanities

Text mining is simply the application computing techniques applied against the content of human expression. Their use is similar to use of the magnifying glass by Galileo. Instead of turning it down to count the number of fibers in a cloth (or to write an email message), it is being turned up to gaze at the stars (or to analyze the human condition). What he finds there is not so much truth as much as new ways to observe.

Eric Lease Morgan
September 6, 2013