Common Terminology in Text Analysis

- **Corpus (and corpora)** – A dataset or collection of text documents. The root of "Corpus" comes from Latin and means "body".

- **Document** – A document is a unit of text that belongs together like row in a table; e.g. email, tweet, news article, open answers to a question, etc. Analogy: An observation is to dataset, as a document is to corpus.

- **Tokenizing** or **parsing**– the process of breaking apart a set of text or document into terms. E.g., By parsing, "I love statistics and computer science", we get the following list: "and", "computer", "I", "love", "science", "statistics".

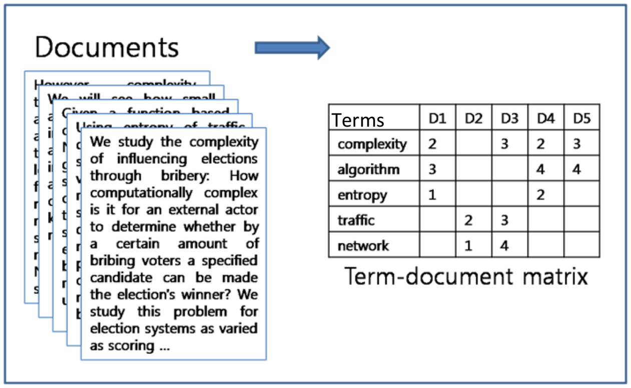
- **Term or word or entity** – a word or a string of characters separated from other words by a space or punctuation; e.g. hello, pound, Greg, David, Alaska, hasn't, etc.

- **Dictionary -** the set of all unique terms in a corpus. These terms may be raw or cleaned as desired, so long as it is communicated to others.

- **Bag of Words** – A commonly used method of text analysis for which a document is represented as a bag words, disregarding grammar and even word order. The frequency or occurrence of each word is used as a feature for learning from text.

- **Synonyms** (and polysemes) – a list of words with similar meaning.

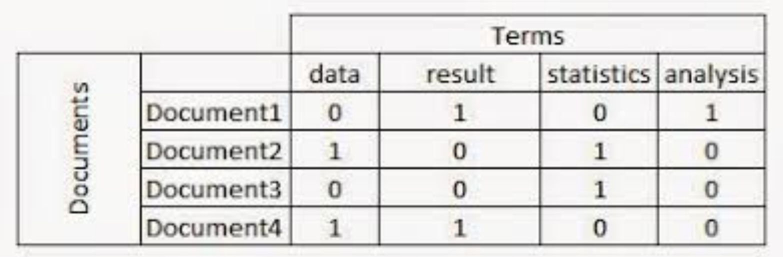
- **Term–by–Document Matrix (TDM)** – Given a dictionary with p terms from n documents, a TDM has the following dimension: p x n. The transpose of a TDM is n x p. Each cell of a TDM includes a count or numeric value that reflects the presence (e.g., frequency) of a term in a document.



We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

Terms	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

A term -by- document matrix. Notice the column headers and the row names.



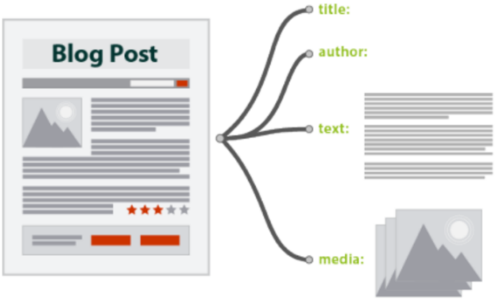
		Terms			
		data	result	statistics	analysis
Documents	Document1	0	1	0	1
	Document2	1	0	1	0
	Document3	0	0	1	0
	Document4	1	1	0	0

The transpose of a term -by- document matrix. Notice the column headers and the row names.

- **Natural Language Processing (NLP)** – A Computer Science field connected to Artificial Intelligence and Computational Linguistics which focuses on interactions between computers and human language and a machine's ability to understand, or mimic the understanding of human language. Examples of NLP applications include Siri and Google Now.

- **Information Extraction** – The process of automatically extracting structured information from unstructured and/or semi-structured sources, such as text documents or web pages for example.

Extraction:



- **Sentiment Analysis** – The use of Natural Language Processing techniques to extract subjective information from a piece of text. i.e. whether an author is being subjective or objective or even positive or negative. (can also be referred to as Opinion Mining)

Sentiment Analysis: