Affordances and limitations of algorithmic criticism

Verhaar, P.A.F.

Appendices

Appendix A: Glossary of Technical and Statistical Terms

Euclidean distance	The Euclidean distance between two points represents the general similarity of the vectors that are associated with these two points. To calculate the Euclidean distance, all the values for these two points need to be subtracted. These differences are then squared and summed. Euclidean distances can effectively be visualised using a dendrogram.
Chi-Squared test	Chi-squared test is a statistical method which can be used to determine whether or not the variation within the values that are collected for a particular variable follows standardised probability rules. In the null hypothesis, there is no difference between the expected values and the observed values. If the result of the formula exceeds a given critical value, however, this null hypothesis must be rejected. A rejection implies that the variable which is investigated indeed has an effect on the values.
Cosine similarity	Cosine similarity is a formula which can give an impression of the similarity between two vectors. This formula normally results in a value in between one and minus one, but if this formula is used to evaluate the similarity of two texts, based on their word frequencies, these frequencies can obviously not be negative. In this specific situation, the cosine similarity lies in between zero and one. If the two vectors are completely identical, the value is one. If the value is zero, all the values in these two vectors are dissimilar.
KWIC	The acronym KWIC stand for “Keyword in Context List”. It is an operation in which a computer retrieves all the passages which contain a specific search term. These search results are then shown within their original context, consisting of a number of words before and after

	the occurrence of this search term. In many cases, users of KWIC software can specify the number of words or characters that are shown to clarify the context. In this thesis, the term is used synonymously with "concordance".
OAC	The *Open Annotation Collaboration* is a data model which represents the way in which scholars can annotate particular sources, or fragments within these sources. This model stipulates that annotations consist of a target, which is the source that is annotated, and a body, which refers to the comment which is made about this source. This model can be implemented using semantic web technologies. The main advantage of using such technologies is that the annotations are not tied exclusively to a single software environment, and that they can be shared more easily across different platforms and across different applications.
PCA	Principal Component Analysis is a form of multi-variate analysis in which a large number of variables can be replaced by a much smaller number of variables. The method is based on the calculation of the eigen vectors of the covariance matrix of all the data values. The method aims to create new variables which can account for most of the variability in the full data set. These new variables are referred to as the principal components. If the first two principal components account for most of the variability, the nature of the full data set can be clarified by plotting these two principal components.
Perl	The *Practical Extraction and Report Language* is a programming language, originally developed by the linguist Larry Wall. It offers extensive possibilities for regular expressions and for object-oriented programming.
Processing	Processing is a programming language, based on the Java language, which consists mainly of methods and classes for the creation of data visualisations. Processing is also a software environment in which Processing code can be executed.

Naive Bayes Naive Bayes is a supervised machine learning algorithm. It often starts with a process in which human beings train software applications by manually supplying categories for the data in a training set. The application subsequently develops a model on the basis of these labelled data. This model can be used to make predictions about new unlabeled data. Naïve Bayes is based on Bayes' theorem, which describes the probability of an event, based on occurrences of events that are related.

N-gram An n-gram can be described generally as a sequence of textual units. These "grams", or textual units, are usually words or characters. More rarely, they can also be syllables or phonemes. The "n" in this term refers to the length of the sequence. A bigram (or a two-gram), for instance, can consist of a sequence of two words or of two characters. Analyses of bigrams or trigrams can be useful in applications which focus on occurrences of specific phrases.

R R is both a software environment and a programming language. The software environment can be used to perform statistical analyses and to produce graphics for the clarification of such analyses. R offers support for a wide range of standard statistical operations, and its functionalities can be extended considerably through the installation of additional packages. R was developed at Bell Laboratories, and it is available as free software under a GNU Public License.

RDF The *Resource Description Framework* provides a generic model for the description of resources. It envisages descriptions or annotations broadly as assertions consisting of three parts: a subject, a predicate and an object. The assumption is that any statement can be expressed using this tripartite structure. RDF has emerged as a standard model for the exchange of data on the web. RDF triples are often visualised as graphs.

Standard deviation The standard deviation is a statistical measure, which can give an indication of the degree of variation within the values of a dataset. Standard deviations are calculated by

taking the square root of the variance of all the values. The variance, in turn, is produced by calculating the average of the squared differences between all values and their mean. If the standard deviation is low, this means that all values are close to this mean value. A high value means that the values are dispersed more widely. A standard deviation of one means that the values are distributed according to a standard normal distribution, and that a plot of these values results in a bell curve.

Td-idf

The abbreviation td-idf stands for frequency-inverse document frequency. It is a statistical operation which was designed to indicate the importance of a specific term within the context of a corpus. The td-idf formula assigns a high weight to a term if it occurs in a small number of documents. Terms which occur in all documents will receive the value zero. This is generally the case for very frequent terms, such as articles, prepositions or pronouns. The formula can thus be use to retrieve the rarer, more distinctive words from a corpus.

Token

Tokens result from the process of tokenisation. In this latter process, a full text is divided into its constituent units. The aim of tokenisation is often to separate a text into individual words. In this situation, the term “tokens” refers to the total number of words in a text.

Topic Modelling

The type of searches that are enabled by search engines or by library catalogues are generally based on the assumption that people know beforehand what they are searching for, and that they can supply specific search terms. Topic Modelling is a fundamentally different approach, in which algorithms organise the textual data by themselves, and in which they try to extract some of the topics that are discussed in a corpus. Topic Modelling is based on the Latent Dirichlet Allocation model, which was first developed by David Blei. The model considers the frequencies of all the words that occur in the corpus, and combines these with data about the documents in which these words are used. On the basis of these data, Topic Modelling algorithms can produce lists of words which are often used in combination, and which can be

assumed to refer to the same topic. These clusters of words are initially unlabelled, and researchers who use Topic Modelling must interpret and label these groups of words themselves. These word clusters can give a rough indication of the topics that are discussed within a collection of text documents.

Type Types are the distinct words that occur in a text. Types are often associated with frequencies, which reflect the number of times the type is used. Data about the number of types and the number of tokens can be used to calculate the type-token ratio, which gives an impression of the diversity of the vocabulary.

XML The *eXtensible Markup Language* can be used to provide explicit descriptions of specific aspects of text fragments through the addition of inline encoding. As is also manifest in the final two letters of the name of the standard, XML is used to 'mark up' specific aspects of a textual document. Marking up a text entails two things: (1) selecting or situating a certain logical or structural component within the text and (2) giving information about the fragment which is selected. The descriptive terms which are added to the document are referred to as *elements*. The elements which are allowed in a particular XML-based encoding language are listed in a *Document Type Definition* or in a *Schema*.

z-score The z-score of a value in a data set expresses its distance to the mean of this data set. This distance is expressed in standard deviations. They provide an intuitive indication of the position of an individual value within the context of the full collections of values. A negative value means that the value is below the mean, and a positive value indicates that this value is higher than the mean. When all the values in a dataset are converted to z-scores, the mean of these values will be zero, and the standard deviation of this collection of z-scores will have value one. An important advantage of z-scores is that they do not have a unit of measurement. Data sets with different distributions and with different units of measurement can be compared effectively by firstly converting all their values to z-scores.

Appendix B: Ontology of Literary Terms

Devices based on repetitions of sounds	Alliteration		
	Assonance		
	Consonance		
	Internal Consonance Rhyme		
Devices based on changes in meaning	Metaphor		
	Simile		
	Metonomy		
	Personification		
	Synechdoche		
Devices based on repetition or word Order	Paronymy		
	Chiasmus		
	Anaphora		
Prosodic Techniques	Rhyme	Perfect Rhyme	
		Slant Rhyme	Assonance Rhyme
			Consonance Rhyme
		Semi Rhyme	
		Deibhide rhyme	
		Aicill Rhyme	
		Internal Rhyme	
	Rhythm		
	Metre	Iambic	
		Trochaic	
		Spondaic	
		Dactylic	
		Trimeter	
		Tetrameter	
		Pentameter	
		Hexameter	
Form	Couplet		
	Three line form	Tercet	
		Triplet	
	Quatrain		
	Quintain		

	Sestet	
	Fourteen line form	Sonnet
Texture		
Diction		
Syntax		
Volume		
Tone		
Mood		
Structural terms	Poem	
	Stanza	
	Line	

This ontology is not intended as an exhaustive list of all existing literary techniques. It concentrates principally on the terms which were discussed in Chapter 2 of this dissertation, and on those terms which were relevant for the case study that was conducted for this dissertation.