Plagiarism and Spam Filtering

Michael P. Oakes
University of Wolverhampton

Plagiarism, Authorship and Spam

- The annual PAN conference on uncovering plagiarism, authorship and social software misuse (especially spam filtering)
- All three are forms of text classification, and often concerned with uncovering fraudulent behaviour.
- Potthast et al. (2009) defined plagiarism as "unacknowledged use of another author's original work".
- A major problem in publishing and academia, especially since electronic texts are so widely available on the internet.

Types of plagiarism and related phenomena

- Cryptomnesia (Taylor, 1965)
- The “dodgy dossier” on Saddam Hussein’s “weapons of mass destruction”
- Carroll (2004) estimates 10% of student work in USA, Australia and the UK may be plagiarised.
- Essay “mills” (Times, 2010).
- Use of computer code without permission in industry
- Press agencies wishing to know the extent of (legitimate) reuse of newsfeeds (METER)
- Removal of duplicate content in databases and search engine hit lists (Metzler, 2005) “diversity”
- Repeated sequences of amino acids in biology
- Is a file identical “undamaged and untouched” after transmission (Manber, 1994).

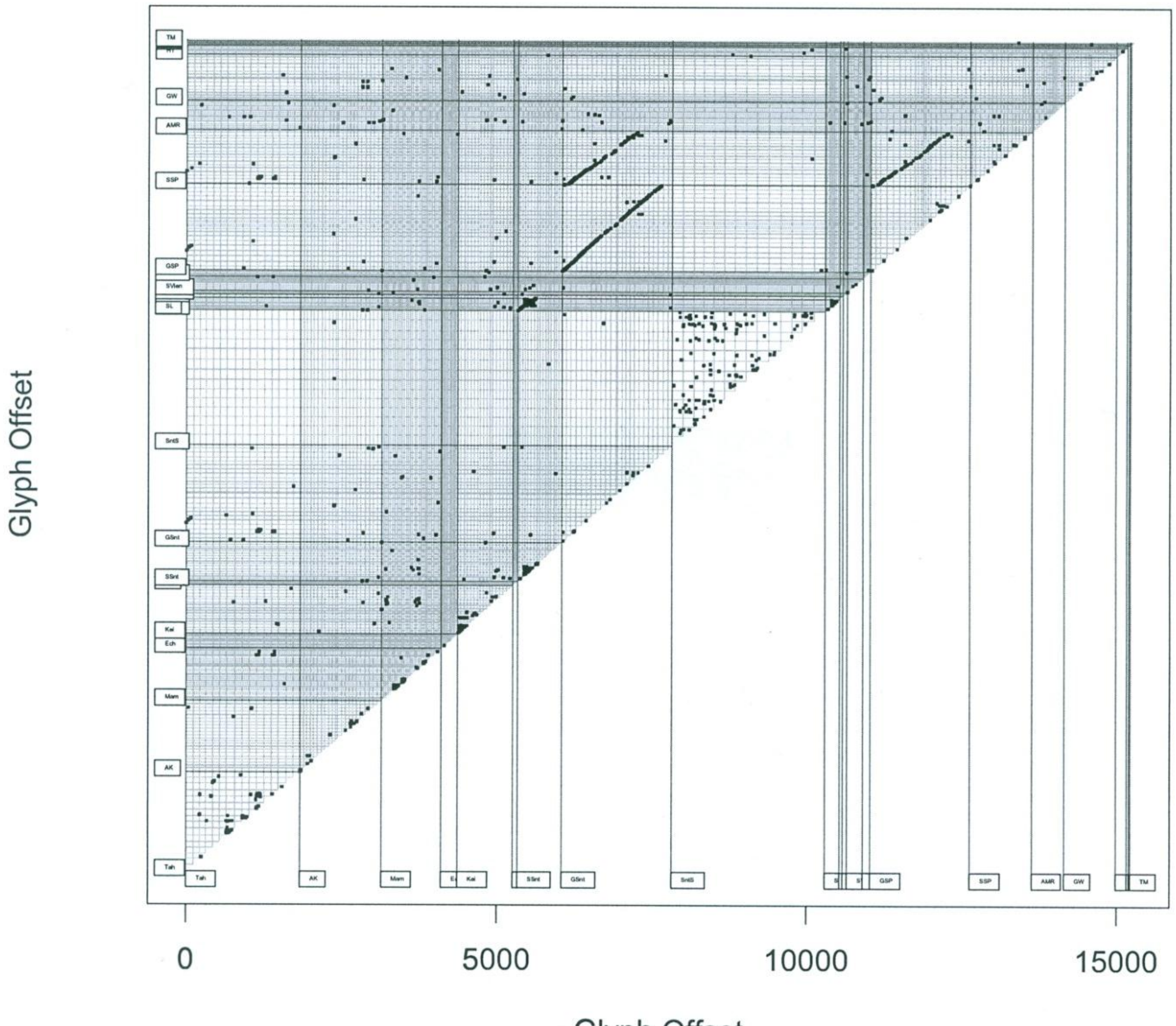
Grades of plagiarism

- Word for word copying is the easiest to detect
- “obfuscation” such as paraphrasing, deletion, substitution of synonyms, changes in word order and tense, active to passive voice (Clough 2000, 2003).
- Patchwriting – fragments taken from several sources and joined in a possibly incoherent order, or resulting in abrupt changes in style.
- Students may insert passages of their own writing, which may be in a poorer style than the rest of the text (Weber-Wulff, 2010).
- Translate from another language, claim as your own work.
- Structural plagiarism, the ideas rather than the exact form of the words are used without credit, e.g. main argument, background sources, experimental design, overall research findings.
- Lukashenko et al. (2007) describe idea plagiarism as “using similar ideas which are not common knowledge”.
- Missing references, failure to enclose direct quotations in quote marks.
- Music and artistic plagiarism (here we consider text only)
- Easy to detect manually, but time consuming and stressful to prove – hence the need for automated software.

Plagiarism detection software

- “take a document as input, and output the pieces of text which have been taken from another source” (Tsatsaronis et al. 2010).
- The “dotplot” is a display showing where matches might be found (Clough, 2000).
- Weber-Wulff (2013) tested 26 commercial plagiarism detection systems – only 3 were even “partially useful”
- Most widely used in the UK is JISC’s “Turnitin” which shows percentage similarity, matching sections and their original sources.
- Submit 5 to 6 words from a suspicious phrase into “Google” – the source should be high on the hitlist. This laborious task is partially automated by SNITCH (Niezgoda and Way, 2006).
- Online searching: search the web, often using search engine metrics
- Offline searching: search a closed set e.g. the lab reports for one class: pairwise similarities can be studied in more detail (“collusion”)
- Intrinsic plagiarism detection: no recourse to original document, but can still look for internal inconsistencies in the essay.

Dot plot (Sproat): all-pairs string matches in the rongorongo corpus



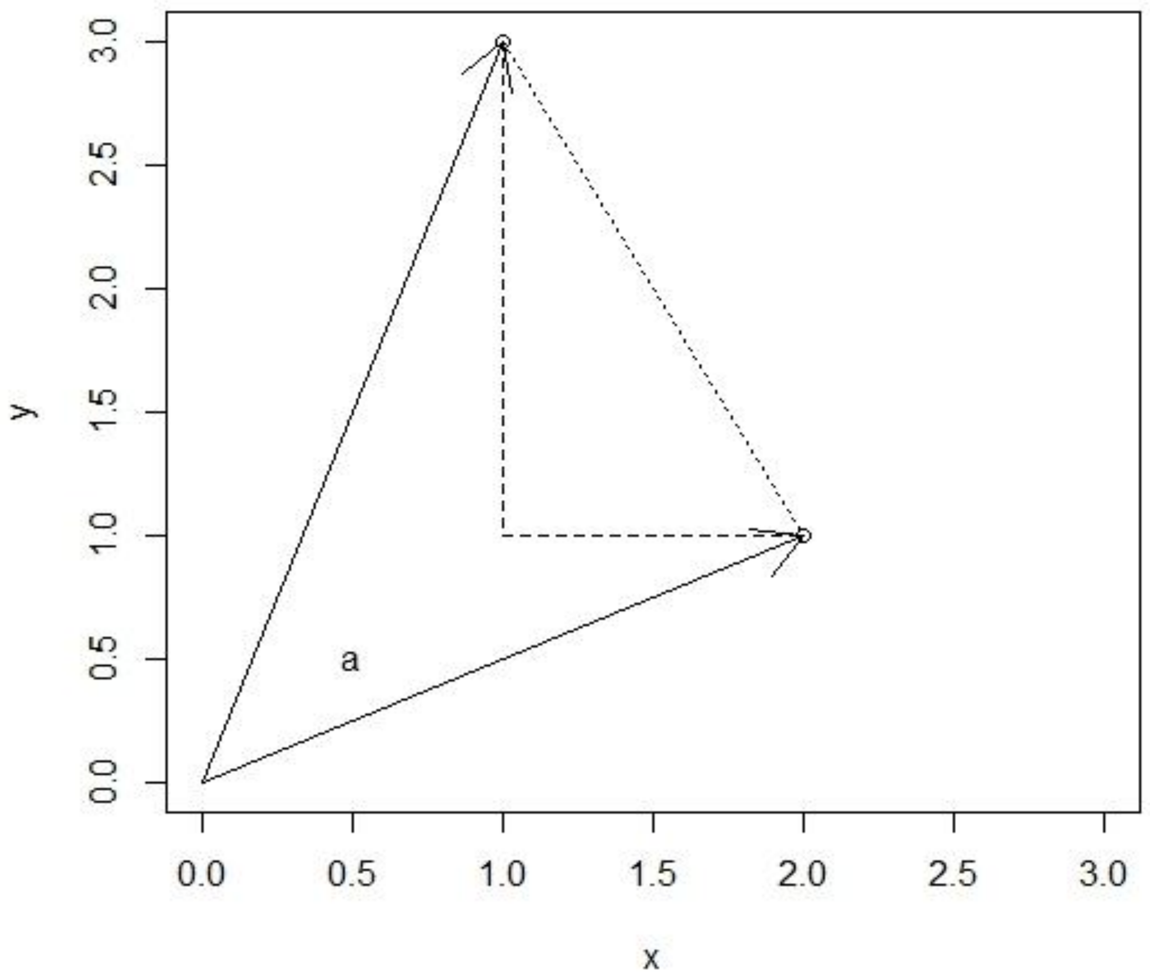
Preprocessing of Corpora and Feature Extraction

- Before developing plagiarism detection software, obtain or compile a corpus of plagiarised and original texts, e.g. Clough and Stevenson, METER, PAN.
- Which statistical and linguistic features should be used to characterise the texts, e.g. single words, sequences of characters or multiple-word phrases?
- The non-trivial task of splitting up a text into features for comparison is called “tokenisation”.
- Word forms can be standardised by removal of punctuation, stop-listing, reducing all words to lower case and replacement of all numerics by a single code, replacement of all words by lemmas, stemming rules(Ceska and Fox, 2009).
- “Matching” techniques: exact matching, ranking using measures from information retrieval, hashing and fingerprinting, language models.

Sequence comparison and exact match

- Possibly the first program to find matching segments of texts and computer programs was the Unix “diff” command.
- A technique which can be used even when the order has changed is the Levenshtein (1966) metric: the least number of “edits” (deletions, insertions and substitutions) required to convert one text into another.
- E.g. “quantitative methods in linguistics” → “statistics in linguistics” requires “statistics → quantitative” (substitution) and “methods (to be deleted), two steps in all.
- Divide by the length of the longer string = 2 / 4 = 0.5.

Search-engine approaches: a = cosine measure, --- is Manhattan distance, ….. is Euclidean distance



Semantic relatedness measures.

- While cosine measure just uses the surface form of the words, Chong and Specia (2011) use WordNet for estimating semantic similarity between document pairs.
- Instead of counting the number of matching words, the count the number of matching synsets.
- The number of synsets common to the suspicious and candidate original document is normalised by the number of synsets in either document (the Jaccard coefficient).
- A review of semantic relatedness measures is given by Budanitsky and Hirst.

Fingerprinting

- Each section of the document is represented by a “fingerprint”, and by comparing fingerprints, one can see which parts of the two documents are identical.
- Using a mathematical function similar to a hash function, each section of text is given a numeric value.
- In Karp-Rabin matching, a code for the first 50-character substring is generated by first finding the ASCII codes for each of these characters numbered from t_1 to $t_{50:}$
- $F_1 = (t_1p^{49}{}_{+} t_2p^{48} + t_3p^{47} + \ldots + t_{50})$ mod M
- where p and M are constants.
- Alignment using anchor points e.g. “acte” and Greedy String Tiling.

Language Models (Barrón-Cedeño and Rosso, 2008)

- Manning and Schütze describe a statistical language model as one which “tries to predict a word given the previous words”
- For example, following “a mile” we might predict “long”, or next most likely, “high”.
- For every word triple in the suspicious document, we calculate p(c|a,b), the probability of word c following the sequence (a,b) = count (a,b,c) / count(a,b).
- Perplexity $PP = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(c|a,b)}}$
- A high PP value for a sentence means it is more likely to have originated from a source external to the rest of the document, since it contains word sequences rarely encountered in the other parts of the text.

Natural Language Processing techniques

- Leung and Chan (2007) consider that a word should match itself with a weight of 1, match a semantically unrelated word with a weight of 0, a WordNet synonym with weight 0.8, hypernyms and hyponyms with weight 0.6, any other related words with a weight of 0.4.
- For example, the similarity between “I go to school every_day” and “I walk to class daily” = (1 + 0.4 + 1 + 0.4 + 0.8) = 3.6 / 5 = 0.72.
- Chong, Specia and Mitkov (2010) combined LM 2-gram and 3-gram PP and a measure of similarity in dependency relations, as combined features in a Naïve Bayes classifier.
- Bär et al (2012) got even better results using stylistic features (such as frequency of function words, type/token ratio), structural features (such as stop words, n-grams and lemma pair distance) and alignment by Greedy String Tiling.
- Uzuner et al (2005) use vectors of syntactic elements, weighted by tf-idf, to characterise documents, then find the similarity between them using the cosine measure.

Intrinsic plagiarism detection

- Most studies of plagiarism consider external plagiarism detection with a reference corpus.
- Intrinsic plagiarism detection is more difficult since no reference corpus is given – we can only detect plagiarised sections of a suspicious document by detecting inconsistencies in the writing style of that document.
- However, an important task, as not all texts are available electronically.
- A similar task to disputed authorship (where we might look for one extraneous book in the author's entire works), but here the text samples are much smaller (a suspicious passage in an essay).
- Method is to characterise portions of text in some numerical way, and then to work out this number for a "sliding window" over the text.
- Discontinuities in (for example) word frequency profile or readability index show points where text might have come from an external source (Stein and Meyer zu Eissen(2007)
- Seaward and Matwin (2009) use Kolmogorov complexity as the numeric measure.

Plagiarism of program code

- Plagiarism detection software designed for text often works well for finding duplicates of computer code.
- However, students disguise plagiarised code by changing comments, renaming variables and procedures, reordering of statements and functions, and adding and removing white space and comments.
- Thus we must look for similarities in programs even when they are superficially dissimilar (Janowitz, 1998):
- Convert program into a tree showing which procedures call which other procedures; look for identical branches;
- Other characteristics which tend to remain constant are overall number of variables used, numbers of reserved words, number of assignment statements, and number of conditions (such as repeat/while, for, case, with).
- For each criterion, there are empirically found acceptable degrees of similarity between procedures.

Distance between translated and original text (1)

- Another form of plagiarism is to translate a text from another language, then claim it as original work.
- Ilisei et al (2010) showed that a computer can be trained to recognise whether a text is a translation or an original, based on the "translationese" hypothesis, which is felt to occur even in very high quality translations.
- Translationese has common characteristics, regardless of the source and target languages, e.g. simplification (lower vocabulary richness, more readable, shorter) (Corpas et al., 2008).
- Ilisei et al. used various automatic classifiers, and an overall "meta-classifier" (majority vote) worked best.
- From a linguistic point of view, Jrip and the decision tree were the most interesting, since they provide output which can readily be understood by humans.
- For example, Jrip gives the rule (lexical richness <=0.16) and (ratio finite verbs <= 0.08) => translation(86/15).

Distance between translated and original text

- Potthast et al. (2010) used a technique drawn from Cross-Language Information Retrieval. The suspicious document is represented by a bag of keywords, which are then translated and used as a query for a conventional web search.
- Barrón-Cedeño et al. 2010) found the degree of character overlap can be effective, e.g Basque “sozialdemokrata” and Spanish “socialdemócratas”.
- Pinto et al. 2009 use a technique related to IBM model 1 for machine translation. This uses an empirically derived bilingual dictionary where p(x,y) is the probability that x can be translated by y.
- For each word, p(x,y) is found for every possible translation of x which appears in the target document, and these are added togther.
- These sums (one for each x) are all multiplied together to calculate an overall probability that the source text is a translation of the target text.

Direction of plagiarism

- A human reader can often infer the direction of plagiarism since "the plagiarist has to avoid the very words which become most naturally, and which, probably, are already in the text being copied" (Olsson, 2009).
- This suggests the use of readability indexes.
- Once a good measure of similarity has been decided upon, the original can be determined as follows (Grozea et al., 2009):
- If two texts A and B contain a plagiarised fragment C, and A is found to be more "stylistically consistent" with C than B is to C, then A is more likely to be the original.
- Ryu et al. use "phylogenetic" measures, originally for finding out which species was the ancestor and which was the decendant.
- These techniques need an asymmetric distance metric such as PEP similarity values or Kruskal's MST algorithm, which shows in which direction the transformation requires most effort (compare with lenition e.g. f → h).
- Computational stemmatology: when comparing several versions of a text, when did certain errors creep in?
- Use of translation universals – is the source simpler than the target (e.g. lexical richness, number of frequent words – Volansky et al., 2012).

Chong and Specia (2012) used RIPPER and an SVM to find 12 most important features, as sorted by Information Gain, to discriminate between original and rewritten text.

Feature no.	Feature Type	Feature
F2	Simplification	Average sentence length
F3	Simplification	Information load: proportion of lexical words to tokens
F6	Simplification	Proportion of sentences without finite verbs
F13	Morphological	Proportion of finite verbs over tokens
F14	Morphological	Grammatical cohesion rate: grammatical words over lexical words
F19	Statistical	Number of characters in the segment
F20	Statistical	Language model 3-gram log probability
F21	Statistical	Language model 3-gram PP (all tokens)
F22	Statistical	Language model 3-gram PP (no </s>)
F23	Statistical	Language model 5-gram log probability
F24	Statistical	Language model 5-gram PP (all tokens)
F25	statistical	Language model 5-gram PP (no </s>)

Case Study 1: Hidden influences from hidden sources in the Gaelic tales of Duncan and Neil MacDonald

- William Lamb (2012) describes the case of Duncan MacDonald (1883-1954), a renowned storyteller of Gaelic tales from the West Highlands of Scotland.
- On five occasions when he orally recounted the tale “Fear na h-Eabaid”, it was transcribed.
- In addition, a text from Duncan’s brother Neil, who was also a gifted storyteller, has been found in the collection of Duncan’s son, Donald John MacDonald.
- The six versions are astonishingly close in their wording.
- Lamb notes that the version from Neil is more similar to that of Craig than any other version.
- This is incongruous, since the MacD text is a) most distant from Craig’s time than any of the others, and b) dictated by a different person.
- Thus it appears that Donald John did not transcribe his version from Neil, but more likely took it from Craig, simply changing some words and phrases and employing his own spelling conventions.

Table of versions of recordings of "Fear na h-Eabaid" recited by Duncan MacDonald

Doc	Code	Recorded by	method	Year	Words
1	Storn	John Lorne Campbell	Recorded on wire, later transcribed	1950	7492
2	MacL47	Calum Maclean	Ediphone recording, later transcribed	1947	6571
3	MacL53	Calum Maclean	Tape recording, later transcribed	1953	7381
4	MacD	Donald John MacDonald	Dictated by Neil MacDonald	1955	7381
5	Clements	Peggy McClements	Dictation	1936	5171
6	Craig	K C Craig	dictation	1944	6571

Dice's similarity coefficient

- Since the six texts were written using different orthographic standards, Lamb standardised them all according to a common set of the Gaelic Orthographic Conventions of 2009.
- All six texts were compared with each other by Dice's similarity coefficient:
- $Dice(A,B) = \frac{2 \times number\ of\ words\ in\ common}{total\ words\ in\ A + total\ words\ in\ B}$
- As a baseline, a different tale recorded by Craig "An Tuairisgeul Mor" had a Dice's coefficient of 0.42 to 0.45 with each of the 6 texts of this study.
- These results confirm that the most similar pair of texts are Craig and MacD.

Comparison of six Gaelic texts by Dice's similarity coefficient

	Storn	MacL47	MacL53	MacD	Clements	Craig
Storn	1	0.79	0.82	0.77	0.75	0.81
MacL47	0.79	1	0.80	0.77	0.76	0.81
MacL53	0.82	0.80	1	0.78	0.75	0.82
MacD	0.77	0.77	0.78	1	0.76	0.87
Clements	0.75	0.76	0.75	0.76	1	0.77
Craig	0.81	0.81	0.82	0.87	0.77	1

The cosine similarity measure

The similarity between the vectors is found by the following formula:

$$Cosine(doc_1, doc_2) = \frac{\sum_{k=1}^{t}(term_{ik}, term_{jk})}{\sqrt{\sum_{k=1}^{t}(term_{ik})^2 \cdot \sum_{k=1}^{t}(term_{jk})^2}}$$

Where $term_{ik}$ and $term_{jk}$ are the frequencies of word k in documents i and j respectively.

Comparison of six Gaelic texts by Cosine similarity coefficient

	Storn	MacL47	MacL53	MacD	Clements	Craig
Storn	1	0.872	0.896	0.873	0.847	0.891
MacL47	0.872	1	0.867	0.846	0.861	0.855
MacL53	0.896	0.867	1	0.865	0.844	0.875
MacD	0.873	0.846	0.865	1	0.866	0.912
Clements	0.847	0.861	0.844	0.866	1	0.863
Craig	0.891	0.855	0.875	0.912	0.863	1

Lamb: Conclusion

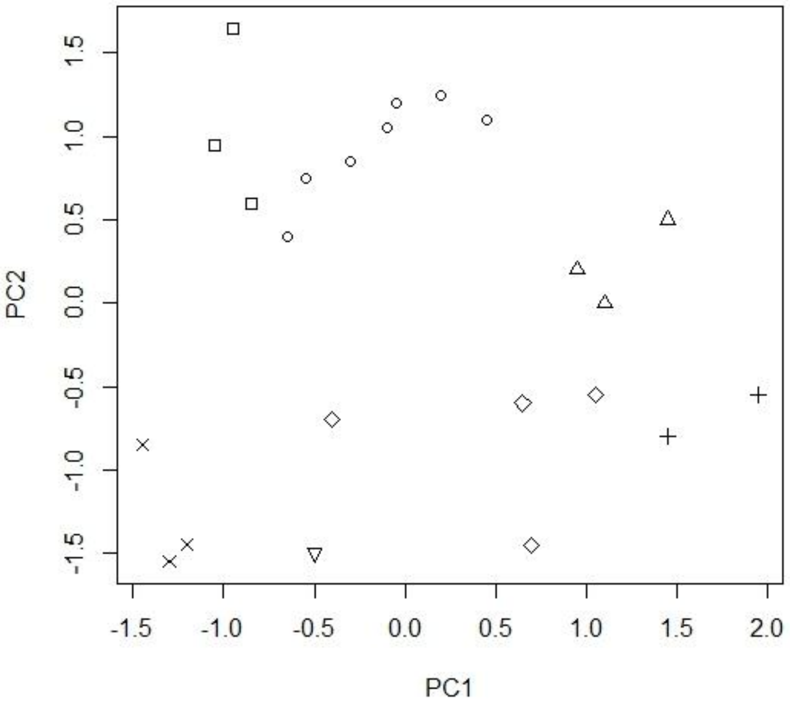
- Lamb also obtained the pairwise cosine similarity measures for the 6 texts.
- The 50 most common words were ignored, otherwise the texts would have been very similar to each other.
- Once again, the most similar pair are Craig-MacD, again suggesting that the MacD manuscript is plagiarised from Craig.
- If these findings are correct, it means that corrupt evidence has given “erroneous impressions to other scholars on the constancy of oral traditions”.

Case study 2: General George Pickett and related writings

- 50 years after the battle of Gettysburg in the American Civil War, LaSalle Corbell Pickett, the widow of General George Pickett who fought in that battle, published a book of letters supposedly written by her husband from the battle front.
- This book, “The Heart of the Soldier” sold very well, and became “part of the canon of American Civil War literature”.
- However, some historians, notably Gary Gallagher, suspected that LaSalle had written them.
- To investigate this claim, Holmes, Gordon and Wilson (2001) ran a PCA on the letters in “Heart of a Soldier” and some related texts as controls.
- Following the “Burrows approach”, the input to the PCA was a matrix of text samples against the frequencies of the 60 most common words in the whole data set.
- All of the 7 sources in the PCA plot are internally consistent, forming coherent clusters.
- Samples of LaSalle’s autobiography fall very close to the disputed letters in “Heart of a Soldier”, suggesting that she was a more probable author than George.
- Gallagher felt that the letters in “Heart of a Soldier” may also have been partly plagiarised from Harrison’s book, but the PCA analysis does not show this.

Texts submitted to a Principal Components Analysis

Text	Symbol
Heart of a Soldier	◊ (small)
LaSalle Pickett's autobiography	□
Inman papers: handwritten papers by George Pickett	+
Letters by LaSalle Pickett	Δ
Book by Walter Harrison "Pickett's Men"	×
George Pickett's war papers	∇
George Pickett's personal papers	◊ (large)



Evaluation of plagiarism detection systems used at PAN

- Recall = number of characters of text labelled as plagiarism by both machine and human judges, divided by the total number of characters labelled as plagiarism by the human judges.
- Precision = number of characters of text labelled as plagiarism by both machine and human judges, divided by the total number of characters labelled as plagiarism by the machine.
- "macro" R and P.
- $F = 2RP / (R + P)$
- Granularity. The number of different detections which overlap with a plagiarised passage is found for each plagiarised passage and the average A is found. Granularity = $\log_2(1 + A)$.
- Overall score = F / granularity is in the range 0 to 1
- 5th PAN also measured cost effectiveness: the average workload per suspicious document = average number of queries and downloads until the first true positive detection has been made.

Plagiarism: Conclusion

- Most databases hold only a small proportion of the sources that might be used, and have no access to “essay banks” or ghost writing.
- However, if the originals can be found, commercial systems such as Turnitin save much time for tutors once they have seen signs that what they are marking is not the student’s own work (Rowell et al., 2009).
- Intrinsic plagiarism detection systems can detect inconsistencies with in a document – but it must be at least 50% original – hence cannot detect ghost writing.
- Consider a comparison between the student’s undisputed own work and the suspicious document.
- Plagiarism software is not a “silver bullet” – we need to teach students what constitutes plagiarism, Universities should train teachers how to recognise plagiarism without the use of software, assignments should be designed to encourage making rather than finding answers, Universities should have clear and consistent policies on plagiarism.
- There is a danger of over-reliance on technology.
- “Detecting plagiarism is academic judgement – and only people can do that” (Rowell et al., 2009).

Spam Filtering.

- Naïve Bayesian classification is a popular method of spam filtering, though it has been superseded performance wise by machine learning methods.
- The first implementation was by Jason Rennie (http://people.csail.mit.edu/jrennie/ifile/old/readme-0.1A)
- Is P(spam|message), the probability that an email is spam (the hypothesis) given the features of the message (the evidence) greater than its complement P(not spam | message) ?
- P(spam) is the so-called prior probability of the message being spam, estimated for example as the proportion of emails arriving in mail boxes which are spam, and P(not spam) is the prior probability of the message not being spam.
- P(message|spam) = P(first word | spam) × P(second word |spam) … P(last word |spam)
- For example, if we have 1000 spam emails in our training corpus, and 300 of them contain the word “cash”, then P(cash | spam) = 0.3.
- We find P(message | not spam) in the same way.
- We keep only the best discriminators, where P(word | spam) / P (word | not spam) is greatest.

Spam: conclusion

- Evaluation is by junk precision (the proportion of messages classified as spam which actually are spam); junk recall (the number of messages classified as spam divided by the total number of spam messages in the corpus; legitimate precision; legitimate recall
- The feature which distinguishes spam filtering from other text classification tasks is that we cannot afford to risk deleting any legitimate message.
- This can only be done by showing the recipient the list of messages deemed to be spam, so he or she gets a chance to retrieve them.
- Since this requires the recipient to at least read the subject header, the gist of the spammer's message will still get across.