Analysis of turnitin.com

An application for detecting plagiarism and self-plagiarism in written work

By Jens Trotzky, jtro012@ec.auckland.ac.nz, 3832929

1 Abstract

This paper offers an experimental analysis of a tool for plagiarism detection. It will cover the definition of plagiarism and self-plagiarism, difficulties in document similarity measurement and its limitations. Three varieties of plagiarised work will be used to perform tests on turnitin.com in an experiment.

Results are being taken to make a statement about the limits of a plagiarism detection tool and indicate possible weaknesses in document evaluation and similarity. A final step estimates the probability of false positives generated by plagiarism detection tools and evaluates the responsibility of supervisors and users of such tools.

2 Introduction

Plagiarism is a crime and depending on the form of plagiarism it might be hard to identify. In a world of internet and computers, making copies of work is easy, cheap and doesn't need particular knowledge. Because copying has become so easy, it is more and more important to verify credibility of work and their assumed authors and to make sure, that dishonest use can be detected.

This also applies to reuse described as self-plagiarism. Even though not unlawful, sometimes disreputable, reuse of someone's own work might be motivation enough to discover.

We will start off with some characterization of reused work, present classifications and offer information about similarity measures and algorithms. Taking into account difficulties and boundaries, we will conduct a couple of experiments, to get a better understanding on what common applications are able to achieve and whether there are parts that need to be observed seriously.

2.1 Reused work and Plagiarism

Plagiarism is a term used quite frequently in the academic environment.

The Guidelines of the University of Auckland state that: "Plagiarism means using the work of others […] and presenting it as your own [work]
without explicitly acknowledging – or referencing -- where it came from."
(AUCKLAND 2003)

This means, that plagiarism is mainly about reusing work in an illegal, dishonest way. This concludes that plagiarised work needs to be identified as such by first recognizing a particular piece of work as being reused. Plagiarism is linked closely to the statement of reused work being presented without acknowledgment.
This is important, as citation and quotation of sources is also reuse of work, but with implicit acknowledgment of the source, which enriches the academic society by spreading the word.

2.2 Plagiarism and Self-plagiarism

While plagiarism is considered to be a crime, a misconduct called self-plagiarism is at least highly unethical, depending on the type, and being prosecuted by some universities. “Self-plagiarism occurs when an author reuses portions of their previous writings in subsequent research papers.” (Christian Collberg 2005)
Collberg illustrates self-plagiarism by giving a couple of examples and classifies some types of self-plagiarism.
This papers purpose is to analyse an application to detect reused work, plagiarised, self-plagiarised or just reused, if the sources has been used in a copy and paste style, as this can be solved using document comparison. Reuse of other authors work by incorporating ideas into someone’s own work is still a possibility but almost impossible to track without having a system that can compare documents according to their semantic similarity.

3 Detection of reused work

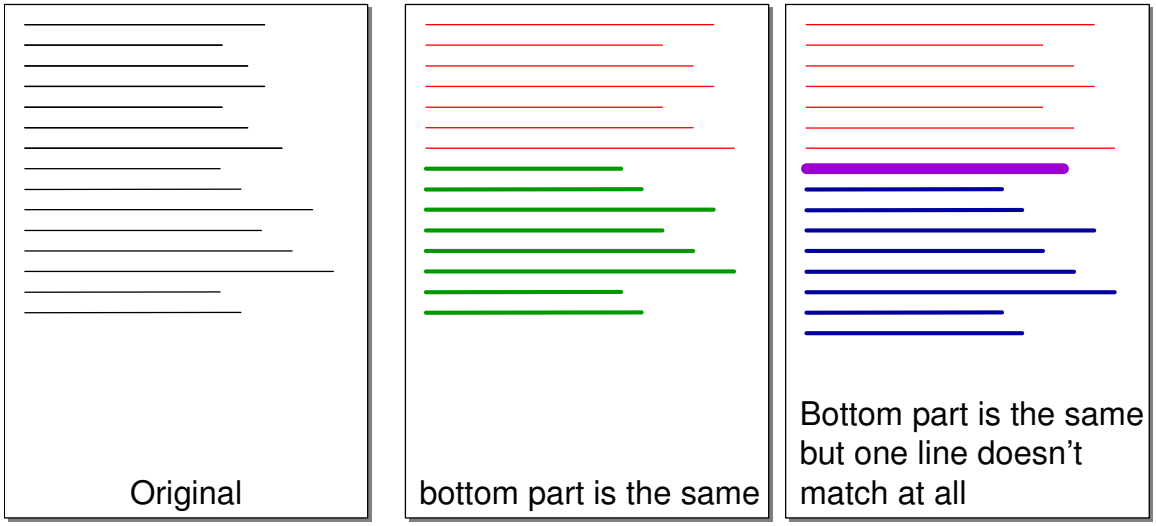
To detect reused work we need to compare documents and calculate their similarities. This comparison can be applied to parts of papers, paragraphs, sentences or even words. At some stage, during document comparison, we might need to use each level of detail to retrieve information about document similarity.
The calculated similarity of two sources can be used to approximate the probability of work being reused and is a first step to make a judgement in an ethical or legal discussion, whether a document is in fact plagiarised.

3.1 Document comparison and similarity

Comparing documents is an important, even though sometimes not an easy, task. This part focuses on two general purpose comparison methods and points out the pros and

cons of those mechanisms to propose a more useful comparison technique for documents, which meet the requirements for detecting reused work.

When comparing two documents we want to make sure that we can compare documents partially as we try to avoid the following:



These documents are being compared line by line without any other restrictions. This works fine, as long as the matching lines have the same ordered number such as line 5 in document one matches line 5 in document 5 (indicated with green colour in document 2). However the same document with one additional inserted line (purple line in document 3) can't be compared even though the lines are existent (blue lines), because there are no matching lines with ordered numbers such that line 5 matches line 5 (there are no operations that would check whether line 5 matches with line 7 in the other document).

This is in fact a simple example but should illustrate, that those things make comparisons a lot harder if we cannot make assumptions about the order of sentences etc. Therefore comparisons need to be based on search, rather than a simple read and test for equality procedures.

3.1.1 Hamming distance

One of the easiest ways to compare two strings (words, sentences, paragraphs or papers) is to use the Hamming distance, so "the number of bit positions in which a pair of words differ"(Watkinson 2000) . This basic concept refers to "two binary strings"(Chapman 2005) and as every electronically stored document can be represented as a binary string comparison of these is easy.

The following example should be a first motivation in comparison of strings and is being discussed in the next section as to how to improve this idea.

Example

Two words *dog* and *cat* can be encoded in ASCII. Both words equal a set of 3 numbers ASCII(dog)=(100,111,103) and ASCII(cat)=(99,97,116) (Watson 1996) in a decimal based number system.

As ASCII limits the number of characters to 2^8=256, we can write each word:

$$dog = 100 \cdot 256^2 + 111 \cdot 256^1 + 103 \cdot 256^0 = 6553600 + 28416 + 103 = 6582119$$
$$= (110010001101111011001111)_2$$

$$cat = 99 \cdot 256^2 + 97 \cdot 256^1 + 116 \cdot 256^0 = 6488064 + 24832 + 116 = 6513012$$
$$= (110001101100001011101000)_2$$

Those two binary strings

```
110010001101111011001111
1100011011000010111010100
```

Have a Hamming distance of 9 out of 24 bits, as 9 out of the 24 bits are different from one another. Taking this as a similarity measure is understandable but also shows that even two words that have no characters in common are pretty similar in their binary representation. In this case one could argue that *cat* and *dog* are

$\frac{24-9}{24} = 62.5\%$ similar.

Application on sentences

The idea of the Hamming distance can be used to compare sentences in a similar way. Instead of using binary strings, one "bit" could be one word. We assume sentences to be a "bit"-string of words. We consider the following two sentences:

The dog ran home because it was raining

and

The cat ran home because it was sunny

We now can compare those sentences word wise:

We get:

The	dog	ran	Home	because	it	was	raining
The	**cat**	ran	home	because	it	was	**sunny**

And calculate the Hamming distance to be 2 out of 8 words (atoms). We say, the two sentences are $\frac{8-2}{8} = 75\%$ similar.

3.1.2 Levenshtein distance

The "Levenshtein distance is a measure of the similarity between two strings". "The distance is the number of deletions, insertions, or substitutions required to transform" one string into the other string. (Gilleland 1998)

Therefore the similarity can be calculated in a negative way. Two strings, words or sentences are more similar the fewer changes need to be made. An algorithm based on the Levenshtein distance gives a penalty for each word that needs to be changed to match the actual sentence in a database. The penalty could be modified depending on the word to be added and could be linked to the word frequency. We will analyse word frequencies in more detail in the next section, as they are of some importance for paper comparisons.

3.1.3 Method of hits and misses

A different proposed method to determine similarity of documents is to calculate the occurrence of words (word types) in a document. (Lia Combrink-Kuiters 1999)

The idea is that we can create a list of all occurring words in each of the documents and for each word that is present in both documents we have a hit, increasing the similarity. For each absent word we decrease the similarity. In the first approach this is being used without taking word frequencies into account. However, the frequency with which the words appear in a document, are useful to increase the accuracy of this similarity method.

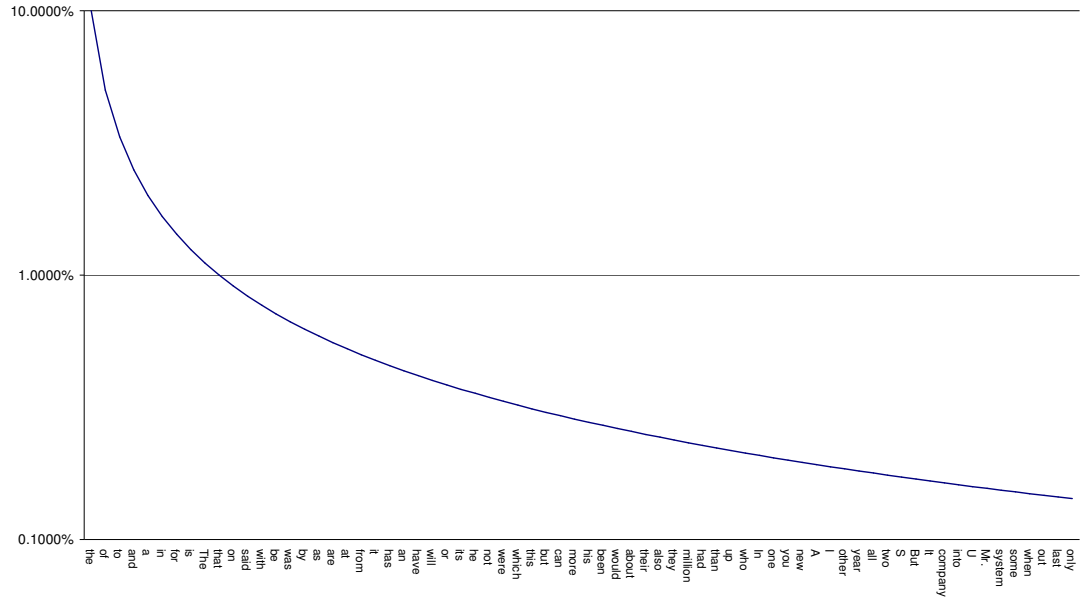
3.2 Word frequency

Observing written English language shows that words appear with a different probability. Words like "the" are a lot more frequent than words like "elephant". This observation has been analysed in various ways by (Welsh 1988) and has been visualized in projects like WordCount™ (Harris 2003).

There are many resources providing word lists ordered according to their frequency in written language. The distribution of the words is being shown the graph below based on the word list from (Leipzig 2001).

The distribution matches a law called Zipf's law, which states, that "if the words of any natural language are ordered in decreasing order of their probabilities of occurrence, […] then a good approximation to these probabilities is given by the

formula $p_n = A/n$"(Welsh 1988). The graph has been drawn assuming A to be `0.1` a first approximation for English language.



Even though this is just a first approximation it is a good first model for English language and many other languages. We leave the discussion about the variance of real languages to language scientist; however, a more accurate similarity can be calculated using more accurate models. More information and analysis is provided in (Hideaki Aoyama 1998)

3.3 Difficulties

3.3.1 Performance

One of the major issues arising from document comparison with one of the proposed ideas is a performance issue. Especially Levenshtein is a very powerful tool to measure similarities between papers taking penalties as a measurement, nevertheless this technique requires a lot of memory as every single possibility of combinations needs to be worked out and in a final step the path of minimal change has to be calculated. Things get even more complicated, if each word is being valued differently. All these tasks can be done by using matrix operations and result in the difficulty to calculate the matrix for a whole document fast.

3.3.2 Word similarity

Based on results from creating spellcheckers it is not a good idea to increase the vocabulary of a spellchecker or in this case the vocabulary of a similarity checker

over a certain number as spelling mistakes wouldn't be identified as such or words might not be recognized as similar anymore.

A good example is given by the word "baht" (currency of Thailand). The question is whether it should be added to the vocabulary or not as it might be more likely to be a misspelled word "bath" instead.(community 2005) Using Zipf's law and analysing typed language one might find out that even the misspelled word "bath" as "baht" is more frequent than the actual word "baht". If that's the case it would be better, not to add "baht" to the vocabulary, but to mark "baht" as misspelled and provide the suggestion to correct it to "bath" instead. "in practice […] an optimal size [of vocabulary]for English appears to be around 90,000 entries" (community 2005)

3.4 Document comparison and similarity measure

Taking the presented ideas into account the best way to compare documents is a combined technique using word occurrence or absence as well as a word comparison approach. Counting hits and misses in a document is easy and fast and avoids matrix calculations. Levenshtein distance is a good way to calculate similarity of sentences but shouldn't be used for whole paragraphs or documents due to performance issues.

A proposed way of analysing a paper is

1. Create a list of word occurrences for each document to get a first similarity measure.
    a. Provide different values for different words depending on the overall frequency in English language as given through Zipf's law.
    b. Work with thresholds to avoid misses because of doubled or misspelled words.
2. Documents that tend to be more similar should be analysed at sentence level using methods like Levenshtein's algorithm taking word frequencies into account.
3. A sentence, that matches another sentence in another document, is a first, good indication for reused work. If more than this one sentence in a row are the same it becomes a strong indication for copied work and should be marked as such.
4. The overall similarity of a document should be the sum of those sentences, which have been marked as possibly copied.

4 Experiments

4.1 Provided papers

We prepare three papers to compare with turnitin.com to retrieve a similarity report for each of them. Each of these three test documents is prepared in a different way:

1. The first paper, to compare with turnitin.com, is a newly written work that hasn't been published before and uses sources that are referenced at the end of the document with acknowledgments of their original authors.
2. The second provided paper is a full copy of a document found on the internet. It has been copied from only one source without acknowledging any authors.
3. The third paper is a fully copied paper that is paste together from about twenty different sources. The range of copied work ranges from single sentences up to whole paragraphs.

The intention of the prepared documents is to determine, how well turnitin.com discovers reused work and what sources it suggests, that work has been copied from. The prepared third document should give more insight into whether turnitin.com compares sentences or paragraphs or whether certain sentences are not being marked as reused at all.

4.2 Turnitin.com reports

Turnitin.com returns an Originality Report for each submitted document. The document is reduced to text only and doesn't show any major forms of formatting and design. Pictures, photos and graphs are being removed and text only is being compared by the turnitin.com application.

5 Results

The reports provided by turnitin.com show different similarity indices for each of the provided papers. While the non copied paper scores a similarity index of 3%, the fully plagiarised a similarity index of 90% and the paper that has been copied together from different sources still scores 82%.

These are the first, most obvious results that are analysed in more depths in the next subsections.

5.1 Fully copied paper

This paper was completely copied from one internet source and only minor changes have been made, such as correcting spelling mistakes or deleting information about

the actual authors. The processing with turnitin.com, however, indicates that the paper has been copied from different sources. Even though turnitin.com states that most of the paper has been copied from the actual copied source it gives another 5 sources that cover about 60% of the copied material.

An interesting result is that Turnitin.com seems to compare submitted paper against each other first, rather than comparing a paper against the document database.

This theory is being supported by the fact that the multi source copied paper contained a part that has as well been used in the fully copied paper. As all the three papers have been submitted at the same time turnitin.com compared those documents against each other first. It marks the part contained in both documents as being copied from one another, rather than copied from the internet source.

This is interesting in terms of plagiarism where people might copy work from one another or work on a project together and use the prepared work as their individual work. These cases are being investigated first.

5.2 Multi source copied paper

The results from the multi source copied paper are diverse. Parts indicate that turnitin.com was unable to identify text as being copied; other results indicate copied material but from different sources and some results show the actual source that the text was copied from. We are going to analyse the results for each of the three outcomes briefly to get an overview on what most likely happened during document comparison.

5.2.1 Text marked as copied but indicates a different source

A couple of sources are marked as copied and a source is given that it is similar to. The comparison with the prepared paper shows that some of these sources are different from the once actually used. Even though not necessarily very important for determining whether a piece of work has been copied, at least an interesting fact to keep in mind as it already gives some information about the use of material in the internet and throughout books. It shows that a couple of authors providing information on the web have used other peoples work as well. Some of the authors acted according to academic ethics and a few people without acknowledging the work of others. This statement is in fact hard to proof as the internet is changing all the time and there is simply no way of deciding which information was there first but one might expect that especially documents at places like www.wikipedia.org are

referenced accordingly even though those cases would be classified as self-plagiarism, rather than plagiarism as most of the authors for Wikipedia simply provide their own work and make it available to the public.

5.2.2 Text marked as copied having a link to the right source

About 30% of the copied work is being linked with the actual source and correctly marked as copied. Interestingly some highlighted parts don't include frequent words such as "the", "and". Doubled words that have been mistyped are being deleted (a sentence: "*My name is is …"* will be treated like *"My name is …"*). Even though the actual comparison algorithm is not available it incorporates a couple of proposed features from the sections before. Especially frequently used words shouldn't have a big impact on the similarity of a sentence.
Working with a threshold in a hit and miss model could support the argument, that a sentence is still the same even if there has been a miss if the miss is related to one of the very frequent words such as "the".

5.2.3 Text not being marked as copied at all

Interestingly turnitin.com didn't mark about 75% of the copied sentences where only one single sentence was copied from a source. It makes sense to avoid marking single sentences simply because the variety of sentences that are syntactically and semantically correct is limited. A sentence that appears in two documents isn't necessarily a strong indication for copied work as certain sentences might appear in a couple of documents. Another explanation for this is the update rate with which the system updates its data base from the internet. Some of the resources copied were new and have been published just recently. Turnitin.com states that its system "checks papers against […]copies of both current and archived internet content and […] [a] proprietary database of millions of previously submitted student papers"(iParadigms 2005). Information from Google's internet search engine show that "Google's index update occurred on average once per month."(Sobek 2003)
This, however, doesn't affect the results that much. Thinking of plagiarism or self-plagiarism this would mean that for an average paper one must copy a couple of hundred sentences all from different sources to not being detected. This doesn't seem to be logical, as the sentences need to form a proper paper in the end.

5.3 Non-copied paper

Turnitin.com regards the Non-copied paper as not copied indicating this with a similarity index of about 3%. Most of the text written is not being marked as copied from other sources. The Interesting fact from this experiment is about quotes and cited sources. Citation is a major part in scientific writing and therefore part of many scientific papers. Turnitin.com only compares the documents and states that certain parts might be copied from other sources but it can't decide whether a text is actually copied unlawfully. It doesn't provide a system to identify correctly cited or quoted sources. In the non-copied paper quotes are being marked as copied even though they are being referenced in an appropriate manner. Not indicating the probability of correctly cited sources is something that needs to be kept in mind for rating the final similarity index.

6 Conclusion

Analysing turnitin.com with such simple experiment as used here already gave some useful information about how the system works and what users need to be aware of, if using a system for document comparison with the goal to detect self-plagiarised and plagiarised work. The results and experiments show that turnitin.com is in fact a system that can support the process of detecting copied work. Especially the results of similarity indices show that (compare 90% similarity for the copied document and 3% for the self-written one). However certain things need to be kept in mind:

1. The system provides information about document similarity, not a statement whether work is in fact plagiarised. The system is able to provide information about possible copied work to some extent if the source hasn't been modified too much and if the sources have been around for a while.
2. The system doesn't recognize everything. There seems to be some clear threshold that needs to be passed before turnitin.com actually starts marking text as being copied.
3. Information about sources are not accurate and because of non-up-to-date information faulty.
4. Referenced and correctly cited materials will most likely being discovered as copied and there is not indication provided that sources might have been used correctly.

Even though some results didn't turn out to be as expected there have been quite a few interesting facts from the experiments. Systems for plagiarism detection are in fact a useful tool that can support detecting different kinds of textual reused papers.

7 References

AUCKLAND, T. U. O. (2003). GUIDELINES: CONDUCT OF COURSEWORK. Auckland, THE UNIVERSITY OF AUCKLAND. **2005**.

Chapman, S. (2005). Similarity Metrics. **2005**.

Christian Collberg, S. K. (2005). "Self-Plagiarism in Computer Science." Communications of the ACM **48**(4).

community, I. (2005). Spell checker, From Wikipedia, the free encyclopedia. **2005**.

Gilleland, M. (1998). Levenshtein Distance, in Three Flavors. **2005**.

Harris, J. (2003). WordCount™.

Hideaki Aoyama, J. C. (1998). "Word Length Frequency and Distribution in English: Observations, Theory, and Implications for the Construction of Verse Lines."

iParadigms, L. (2005). Plagiarism Prevention, iParadigms, LLC. **2005**.

Leipzig, U. (2001). Rangliste der englischen Wörter (Ranked list of English words). **2005**.

Lia Combrink-Kuiters, R. V. D. M., Henk Elffers, Kees van Noortwijk (1999). Comparing Student Assignments by Computer. 14th BILETA Conference: "CYBERSPACE 1999: Crime, Criminal Justice and the Internet".

Sobek, M. (2003). Google Dance - The Index Update of the Google Search Engine, eFactory GmbH & Co. KG. **2005**.

Watkinson, J. (2000). Hamming distance. The Art of Digital Video, Focal Press**:** 487.

Watson, G. (1996). ASCII Table.

Welsh, D. (1988). Zipf's law and word entropy. Codes and Cryptography, Oxford University Press**:** 97-98.