24.

DIDACTIC EVOLUTION OF SIMILARITY DETECTION SOFTWARE: THE EXAMPLE OF COMPILATIO

Frédéric Agnès

With the collaboration of Élodie Bresse and Lucile Guillermin, in charge of training and support for academic integrity.

Abstract

Since 2005, Compilatio has been offering tools to help detect and prevent plagiarism. Users of similarity detection software were initially attracted by the ability to track down cheaters. They are now more aware of the tools and services offered to create an environment that encourages the adoption of integrity and citizenship values, especially digital ones. They are aware that plagiarism is not a passing evil to be eradicated, but a deep-seated temptation that each individual must learn to overcome. The technology used to help teachers spot cheating has also evolved. The approach was initially syntactic, comparing texts formally to detect similarities. It then became semantic, using so-called artificial intelligence techniques to find similarities between different words with the same meaning. The issues related to plagiarism prevention illustrate how technology and pedagogy can be used together to train individuals for their future professional and civic life.[*]

[*] Corresponding authors: F. Agnès. To quote this chapter: Agnès, F., "Didactic Evolution of Similarity Detection Software: the Example of Compilatio" in: Bergadaà, M., Peixoto, P. (Eds.), *Academic Integrity: A Call to Research and*

1. Introduction

When we created the Compilatio service in 2005, our ambition was to curb student plagiarism in academia. But over the years, needs have changed because of the evolution of cheating practices and the related needs of instructors. Indeed, as checks have been conducted in educational institutions, instructors' needs have become more refined. Compilatio users were initially attracted by the possibility of quickly detecting cheaters who abused the copy-paste function. They are now more interested in creating an environment that encourages the adoption of integrity and citizenship values, especially digital ones.

There is not now, and will never be, a vaccine against cheating (unlike coronavirus), and institutions rarely express a desire to unmask all cheating attempts, after the fact. Plagiarism is an easy temptation that everyone must learn to overcome. Pedagogical services, libraries, and teaching staff must support this learning process. Thus, the service we are offering today is designed to meet this need for support, in terms of early prevention.

The usefulness of similarity detection assistance software is, first, shown by its dissuasive effect. It can prevent the massive and uninhibited use of plagiarism. The implementation and use of such a tool in an institution is also an opportunity to remind students of the rules of integrity and their rights and duties toward their institution. Changing practices in the area of plagiarism prevention, are related to both educational and technological developments. These tools are factual measuring instruments; today they can reveal similarities in form, but tomorrow they will reveal similarities in meaning, with the help of artificial intelligence.

Action, Geneva: Globethics Publications, 2023, pp.581-599, DOI: 10.58863/20.500.12424/4273128 © Globethics Publications. CC BY-NC-ND 4.0. Visit: https://www.globethics.net/publications

2. Changing attitudes in the support of plagiarism prevention

The idea that one can reduce plagiarism in the same way as one would eradicate a disease is illusory, because each new generation of students bears within it the seeds of creativity in circumventing instructions they consider to be tedious. Cheating has always existed. It evolves and progresses with the use of new technologies. Every year, we must start again and teach good writing and citation practices to new students. And every year, some will try to slip through the net.

Our goal is clear: to make plagiarism more complicated, more time-consuming, and more effortful for students than simply respecting copyright. The aim, of course, is to enhance the value of degrees, which depends on the strength of the skills acquired by learners.

2.1 Reveal cheating and hunt down cheats?

With the generalization of the use of word processors and the Internet in educational institutions in the late 1990s, the scope for cheating expanded considerably. The excessive use of copy-paste gives many instructors the impression that their pupils and students spend less time writing their assignments than they themselves do correcting them.

These ‘cheating opportunities’ were brought back to the forefront with the global pandemic of 2020 and the rise of distance learning and hybrid learning. At the same time, instructors are making greater use of Compilatio due to the increased use of distance learning, linked to the COVID-19 epidemic. It was used nearly three times more often between April and June 2020 than in the previous year, and the increase in use throughout 2020 was about 70% compared to 2019; the increase was as high as 400% during the months of lockdown.

When Compilatio was created, at the end of 2004, the service was first presented to users as antiplagiarism software. The first instructors who used it were partly seduced by the change in the balance of power with cheaters. It became possible for them to identify most cases of cheating on a mass basis. Some even nursed hopes of seeing plagiarism disappear.

It soon became obvious to Compilatio users that each discovery of a case of potential plagiarism entailed the implementation of long and tedious procedures to have the cheating recognized. Instructors must deal with pupils or students who have much to lose and little to gain by making amends... They also face obstacles within their own institutions, which have a long-term interest in ensuring the quality of their teachings and the value of their degrees, but which also face the management of a difficult dispute and the risk of bad publicity in the short term. When burying one's head in the sand is the strategy chosen to resolve this dilemma, instructors can find themselves without support. They may then feel discouraged about carrying out an investigation alone, in which they would ultimately have more to lose than to gain. That is why we thought that the best way to serve our customers would be to go beyond supporting instructors in identifying cases of cheating and establishing evidence.

Certainly, the announcement that an educational institution uses Compilatio discourages massive fraud among students. But our role today is to support educational stakeholders who want to create conditions for integrity. We want to help them encourage their students to choose personal work and academic honesty over the apparent ease of cheating.

2.2 Toward the creation of conditions conducive to educate students to value integrity, creativity, and originality over easiness

Encouraging students to make ethical decisions and to unite behind the concepts of digital citizenship has become our challenge over the past five years. To do this, we are committed to a pedagogical approach that raises awareness, rather than a repressive one based on sanctions.[632]

This approach, which is part of institutional philosophy, is based on several pillars that must be driven by an internal project leader and constantly communicated in positive terms to all stakeholders within an institution: management, teaching staff, students, etc.

Based on our ongoing discussions with the people in charge of implementing Compilatio in their schools, we have established that monitoring is necessary, but that it is not intended to punish bad behavior. On the contrary, monitoring creates an opportunity to reward both:

- students, who are rewarded for their ethical work decisions;
- the institution, which validates the mechanisms of its pedagogical approach.

Monitoring will only be perceived as a positive, rewarding process if upstream conditions are conducive to the learning and development of the digital skills associated with integrity: knowledge of indicators for assessing the reliability of sources, understanding of copyright, assimilation of citation standards, etc.

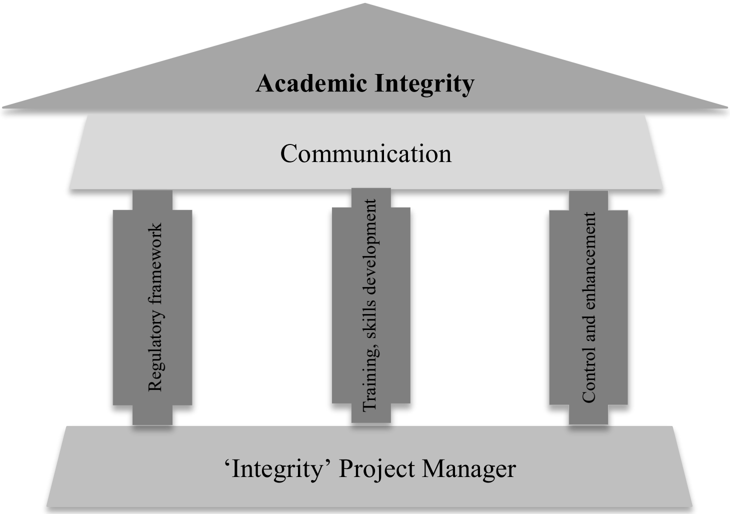
In addition, control becomes legitimate when it is supervised and when the expected behavior is made explicit from the outset, that is, from the moment of arrival at the institution. Establishing a regulatory framework then becomes very important. Defining good and bad behaviors with the associated rewards and sanctions indicates the direction—the path to follow.

Figure 1 presents the three 'pillars of accountability' that we propose putting in place in the institutions we support. The 'integrity' project

[632] M. Peters and S. Gervais, 'Littératies et créacollage numérique', *Language and Literacy,* 18(2) (2016), 62-78.

manager is the main contact person for this approach within the institution.

Figure 1: The pillars of academic integrity.



It is easy to understand the importance of this awareness-raising approach and the related issues by an analogy with road safety.[633] To inform about the framework to be respected on the road, one can observe speed limit signs, advertising posters, advertising campaigns.

For plagiarism prevention at Compilatio, awareness-raising materials are used to communicate the rules to be followed: training sessions, workshops, plagiarism prevention posters, turnkey website dedicated to plagiarism, etc.

To assist in practice, road safety indicates how to apply the regulations: traffic laws, accompanied driving, or a driver awareness course. Compilatio guides its users with a Magister and Studium toolkit.

[633] Compilatio, 'Rules of the Road, Plagiarism Prevention | 1 Goal: Responsible Behaviour', *Compilatio,* 9 April 2020.

Training and surveys are offered to academic institutions so they ca learn how to recognize and prevent plagiarism.

To evaluate compliance, the speedometer is a reliable indicator used on the road. The Studium and Magister software can be used as a 'speed camera' to indicate whether the limit has been crossed. Indeed, the similarity rate is an impartial indication of the percentage of 'copy and paste' in an assignment.

To reprimand fraudsters, radar monitoring warns of danger and non-compliance with the law. It is only by highlighting a complete preventive approach that one can hope to induce everyone to clarify their values when making decisions under conflicting constraints. For students, this means a choice between:

- going faster to get a (good) mark by cheating; or
- taking the time to earn a grade and graduate without cheating their fellow students and instructors.

2.3 Putting the meaning of teaching back at the heart of the anti-plagiarism approach

In a previous article, we argued that, while the development of the Internet has facilitated the sharing of knowledge, it has also led to the idea that it is possible to access and falsify images, books, and even music.[634] Many people believe that knowledge belongs to everyone and that it is therefore unnecessary to mention original works and authors. The question students ask themselves is a legitimate one: 'What is my incentive to respect copyright?'

Let us be clear: people who put all their energy into deliberate cheating will not be 'saved' by the instructor's use of foolproof software, because those cheaters already know that what they are doing is wrong. On the other hand, a clear signal must be sent to guide the well-intentioned in the right direction. Our challenge is to make them

[634] Compilatio, 'Why Is Plagiarism Prohibited? What Are My Incentives to Respect Copyright?', *Compilatio,* 27 July 2021.

admit that plagiarism is forbidden, because a breach of copyright harms not only the plagiarized author but also all the creative people and authors whose ideas are looted. It is therefore essential to value authors and their words.[635] As Paul Desalmand said, ‘a quotation without references is about as useful as a clock without hands’.[636]

So it is not just for the present and in relation to a particular work that we need to act but for a more sustainable education in terms of integrity. For while it is important to understand the issues at stake in not plagiarizing, it is also important to understand issues that go far beyond that, including

- educational issues: to examine the capacity to learn, acquire knowledge, integrate new skills, etc.
- professional issues: to instill the right behaviors for future professional life;
- societal issues: to be honest in all aspects of one’s life, to be an informed citizen.

So anti-plagiarism software does not have to be foolproof—which would make it fun to circumvent—but it must be credible. The zone of freedom and the zone of prohibition must be delineated and made objective so that students learn and integrate compliant behavior

If most people learn best by making mistakes, what better place for them to do so than in school? For example, instructors can turn a potential plagiarism situation into an educational opportunity. They can determine the reasons for an identified case of plagiarism and, by discussing the objective facts revealed by the software, they can ask themselves why students choose this bad behavior: ‘This understanding leads to awareness and enables adjustment of teaching methods in order

[635] See also Compilatio, ‘Comprendre la nouvelle réforme européenne du droit d’auteur’, *Compilatio,* 29 May 2019.

[636] P. Desalmand, *S.O.S. Citations* (Paris: Leduc.s Éditions, 2008), p. 237.

to avoid recurrences'.[637] Revealing mistakes as early as possible in order to transform them into opportunities for progress is a real pedagogical act, which succeeds when behavioral change is observed.

3. Technological evolution of similarity detection software

Like the pedagogical approach, the technology used to help instructors detect cheating has evolved greatly over the past fifteen years. When they first appeared in the early 2000s, similarity detection tools such as Compilatio were designed to identify identical areas of text by comparing the texts formally to detect similarities. This is known as a syntactic approach, as the form (syntax) of texts is compared.

But not all cases of plagiarism are characterized by strictly identical borrowing. Today, in order to enlarge the situations of use and improve the software's performance, new technical approaches are being experimented with. It may also be possible to detect cases of paraphrasing, reformulation, or translation. Advances in artificial intelligence have allowed some promising initial experiments: in 2017, a system designed by Compilatio won a translation detection contest, by identifying more than 80% of the translations between two texts.[638] These experimental advances may soon enrich similarity detection software.

[637] Compilatio, 'Student Plagiarism: Create an Educational Learning Opportunity', *Compilatio,* 9 December 2019.

[638] J. Ferrero and others, 'CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity', in *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)* (2017).

3.1. Detecting similarities with a syntactic approach

The syntactic approach is the most efficient method in terms of computer processing time to reveal the most obvious cases of plagiarism, which is why it is used today in most similarity detection software.

It consists of comparing character sequences from two texts, by grouping ‘n-grams’, as illustrated in Figure 2.

Figure 2: Text comparison using clusters of n-grams.[639]

Clustering two sentences using n-grams

	‘This is a sentence.’	*‘What a sentence is this!’*
N = 1 –> unigrams	[this], [is], [a], [sentence]	[what], [a], [sentence], [is], [this]
N = 2 –> bigrams	[this is], [is a], [a sentence]	[what a], [a sentence], [sentence is], [is this]
N = 3 –> trigrams	[this is a], [is a sentence]	[what a sentence], [a sentence is], [sentence is this]

[639] DeepAI, ‘N-Grams’, *DeepAI,* 17 May 2019.

Comparing two sentences using n-grams	
This is a sentence.	What a sentence is this!
this] [is] [a] [sentence]	[this] [is] [a] [sentence] *-no match-*
this is] [is a] [a sentence]	*-no match-* *-no match-* [a sentence]
this is a] [is a sentence]	*-no match-* *-no match-*

This approach has the advantage of being effective in comparing texts in any language, if they are written in the same language and they contain formally identical passages. To characterize an obvious borrowing, the quality of similarity detection depends on the correct parameterization of the analysis algorithm, according to criteria such as:

- the length of the n-grams retained, and
- the number of successive points of similarity common to both sources.

Of course, this approach has many limitations. For example, if the passages compared are too short, then ‘common’ sentences containing a few identical words could be wrongly considered as similar and generate many false positives (i.e. with the unigram comparison in Figure 2). On the other hand, if the passages compared are too long, then variations in form between two texts, such as changes in tense or the use of synonyms, will cause the recognition of similarities to fail.

The effectiveness of a software will therefore depend on the fineness of the adjustment selected by its designer, to find a fair compromise between not missing any similarities, even if this means having false positives, and presenting only significant similarities at the risk of not detecting the shortest borrowings.

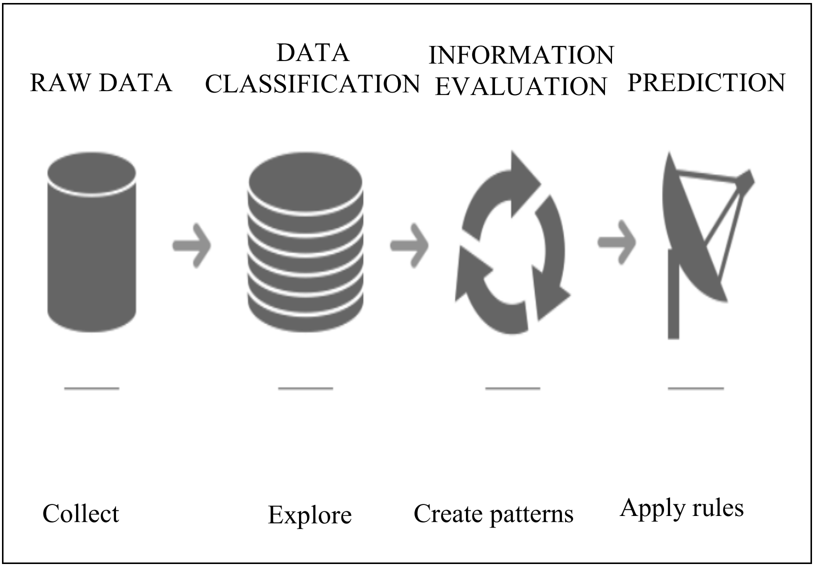
3.2. Principles and contributions of artificial intelligence

The term artificial intelligence (AI) is often used to refer to machine learning and neural networks (deep learning). The progress made in this field has applications in many areas, including text analysis. In the years to come, there will be significant improvements in similarity detection software, expanding the possible fields of application and leading to the development of software to help detect many cases of cheating.

To build an AI, you need to provide a computer system with 'labeled' learning data. They will enable the machine to empirically recognize (learn) rules that accurately describe the learning data set or discriminate among them. The considerable computing capacity available today makes it possible to explore all possible combinations of rules and design an effective classification or recognition method. It is therefore a question of empirically and *a posteriori* observing which rules best suit the desired task, after exploring all the possible paths envisaged by the system.

Once these rules have been 'learned', they are then validated on a test data set, to make sure that the rules apply to both new test data and training data. Learning is considered effective if the rules obtained apply to both data sets. It is therefore assumed that the application of these rules will be satisfactory and they can be used for the analysis of new data, for which the result of the expected processing is not known in advance (Figure 3).

Figure 3: Machine learning: deciphering a revolution in progress.[640]

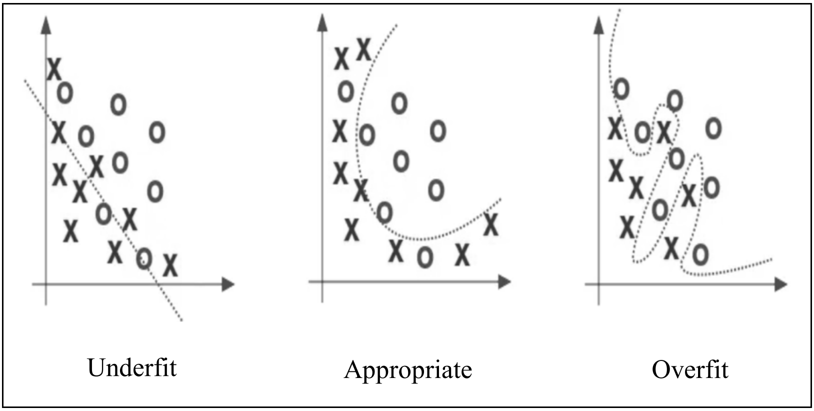


Even if they are applicable by ad hoc computer systems, the rules learned by the AI can be totally incomprehensible, as they are too abstract to be understood by a human mind. In addition to a suitable computer system, the first condition for setting up an AI is large volumes of suitably labeled learning data.

It may seem surprising but building learning on too much data can also reduce the AI's performance, due to the phenomenon of 'overfitting'. Indeed, overfitting the intelligence to its training data can lead to the production of a model that is too accurate or too demanding, which may deviate from the general model adapted to the task (Figure 4).

[640] A. Fall, 'Machine Learning: Décryptage d'une révolution en marche', *Content Shaker,* 4 September 2018.

Figure 4: Diagram showing underfitting, ideal case, and overfitting.[641]



The success of a neural network–based machine learning system therefore depends as much on the volume of data used for learning (training) as on the volume of data used to test and validate the model's adaptation to the desired task.

The use of AI can open many applications in the field of text analysis, for the purpose of detecting various forms of cheating. To mention a few examples of the experiments currently being conducted by Compilatio, or envisaged for the future, the stylistic analysis of a text will make it possible to detect whether passages present anomalies in relation to the rest of the document, and thus to identify passages that may have been written by different authors.

On the basis of different documents identified as written by the same author, it will be possible to determine whether a new document was written by that same author or by a third party. It will also be possible to reconstruct a plan according to the areas of text covered by main or secondary themes and the logical articulation of a text. Similarly, it will be possible to highlight the most characteristic passages of a text to

[641] B. Maurice, 'Comprendre overfitting et underfitting', *Deeply Learning,* 15 September 2018.

facilitate quick reading and good understanding, or to identify documents with similar structures, in terms of the topics discussed and the sequence of ideas presented, which could reveal a theft of ideas. These are all new indicators for characterizing documents and automating searching and comparison. Returning to the detection and measurement of similarities, which are often indicative of plagiarism, it may be possible to detect similarities between reformulated or translated texts.[642]

3.3. Application of AI through a semantic approach

Compilatio is currently working on the design of a AI system specializing in the detection of reformulations and translations, which are more elaborate forms of similarities than the copying and pasting detected with the syntactic approach mentioned above. To overcome the limits of the syntactic approach, it is now possible to favor a semantic approach, built using AI. The goal is to identify similarity points between two texts based on the proximity of meaning of the words used.

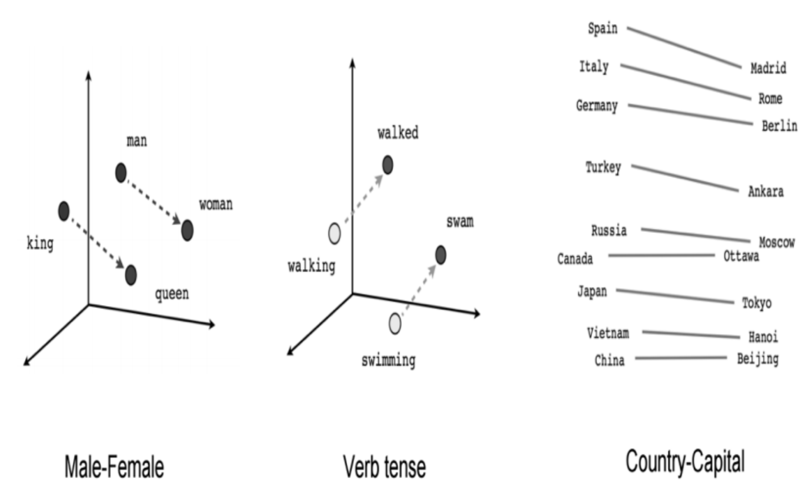
Our research focuses on the detection of similarities using the 'word embedding' technique.[643] This approach makes it possible to compare two texts even if they contain only different words, or even are written in different languages. The principle is to place the words of each language in a 'space' constructed in such a way that words with the same meaning have close coordinates in this space. Similarly, the

[642] Many text analysis technologies already exist (NLP, Natural Language Processing services), such as www.synapse-developpement.fr, www.meaningcloud. com, www.cloudfactory.com, Microsoft Azure Cognitive Services / Language API, IBM Watson, Google Cloud Natural Language, OrphAnalytics, etc.

[643] J. Ferrero, 'Similarités textuelles sémantiques translingues: vers la détection automatique du plagiat par traduction' (unpublished thesis, Université Grenoble Alpes, 2017).

distance between two words with a similar relationship is equal in each space, as shown in Figure 5.

Figure 5: Creating word embeddings: coding the Word2Vec algorithm in Python using deep learning.[644]



Thus, the gap between ‘man’ and ‘woman’ in a space representing terms in French is the same as the gap between ‘king’ and ‘queen’ or ‘male’ and ‘female’. The positions of the words ‘man’ or ‘woman’ in the space representing the French language will also be the same as the positions of the words ‘man’ and ‘woman’ in the space representing the English language. In this way, it is possible to measure the semantic proximity of two words, even if the words are different or belong to different languages, provided that a spatial representation of the words has been constructed in each of the languages studied.

In the long term, one can dream of an automatic system that would alert us to several kinds of similarities or cheating. We could even see the emergence of new systems capable of assisting instructors in all the

[644] E. Bujokas, ‘Creating Word Embeddings: Coding the Word2Vec Algorithm in Python using Deep Learning’ *Towards Data Science,* 5 March 2020.

academic work they supervise in their review, evaluation, and correction tasks. However, pushing the capacity of machines further to detect borrowings, similarities, and all forms of resemblance, both in form and in content, will also raise new questions.

4. Conclusion

What kinds of creative capacity and reflective skills should be assessed? Where do we place the thresholds that separate an original, honest, authentic creation from a legitimate and appropriate borrowing or a reprehensible plagiarism? The emergence of new indicators will inevitably raise questions about the definition of the standards to which it is appropriate to conform, as was the case with the emergence of systems for measuring similarities between two texts, known as *anti-plagiarism software* for convenience.

The years a person spends pursuing an education represent a time for learning, experimenting, and acquiring skills and values.

Compilatio's business model is to design and propose technological, educational, and methodological tools, at the service of teaching. They are not a substitute for instructors' correction and judgment, which are part of their pedagogical mission. We reveal the similarities, but we do not judge whether or not those similarities are reprehensible.

If the major challenge for the coming decades is to reconnect with the values of integrity and authenticity, and ensure that everyone can be fully responsible, in all fields where citizenship can be expressed, Compilatio will be there to accompany the educators.

Bibliography

Bujokas, E., 'Creating Word Embeddings: Coding the Word2Vec Algorithm in Python using Deep Learning' *Towards Data Science,* 5 March 2020.

https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

Compilatio, ‘Comprendre la nouvelle réforme européenne du droit d’auteur’, *Compilatio,* 29 May 2019. https://www.compilatio.net/blog/nouvelle-reforme-europeenne-du-droit-dauteur

Compilatio, ‘Student Plagiarism: Create an Educational Learning Opportunity’, *Compilatio,* 9 December 2019. https://www.compilatio.net/en/blog/educational-learning-following-plagiarism

Compilatio, ‘Rules of the Road, Plagiarism Prevention | 1 Goal: Responsible Behaviour’, *Compilatio,* 9 April 2020. https://www.compilatio.net/en/blog/plagiarism-prevention

Compilatio, ‘Why Is Plagiarism Prohibited? What Are My Incentives to Respect Copyright?’, *Compilatio,* 27 July 2021. https://www.compilatio.net/en/blog/my-reasons-to-respect-copyright

DeepAI, ‘N-Grams’, *DeepAI,* 17 May 2019. https://deepai.org/machine-learning-glossary-and-terms/n-gram

Desalmand, P., *S.O.S. Citations* (Paris: Leduc.s Éditions, 2008).

Fall, A., ‘Machine Learning: Décryptage d’une révolution en marche’, *Content Shaker,* 4 September 2018. https://contentshaker.Webedia-group.com/article/machine-learning-decryptage-d-une-revolution-en-marche_a377/1

Ferrero, J., ‘Similarités textuelles sémantiques translingues: vers la détection automatique du plagiat par traduction’ (unpublished thesis, Université Grenoble Alpes, 2017). https://tel.archives-ouvertes.fr/tel-01721390

Ferrero, J., and others, ‘CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity’, in *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)* (2017). https://hal.archives-ouvertes.fr/hal-01531330

Maurice, B., ‘Comprendre overfitting et underfitting’, *Deeply Learning,* 15 September 2018. https://deeplylearning.fr/tag/sur-apprentissage/

Peters, M., and S. Gervais, ‘Littératies et créacollage numérique’, *Language and Literacy,* 18(2) (2016). https://doi.org/10.20360/G21W2H