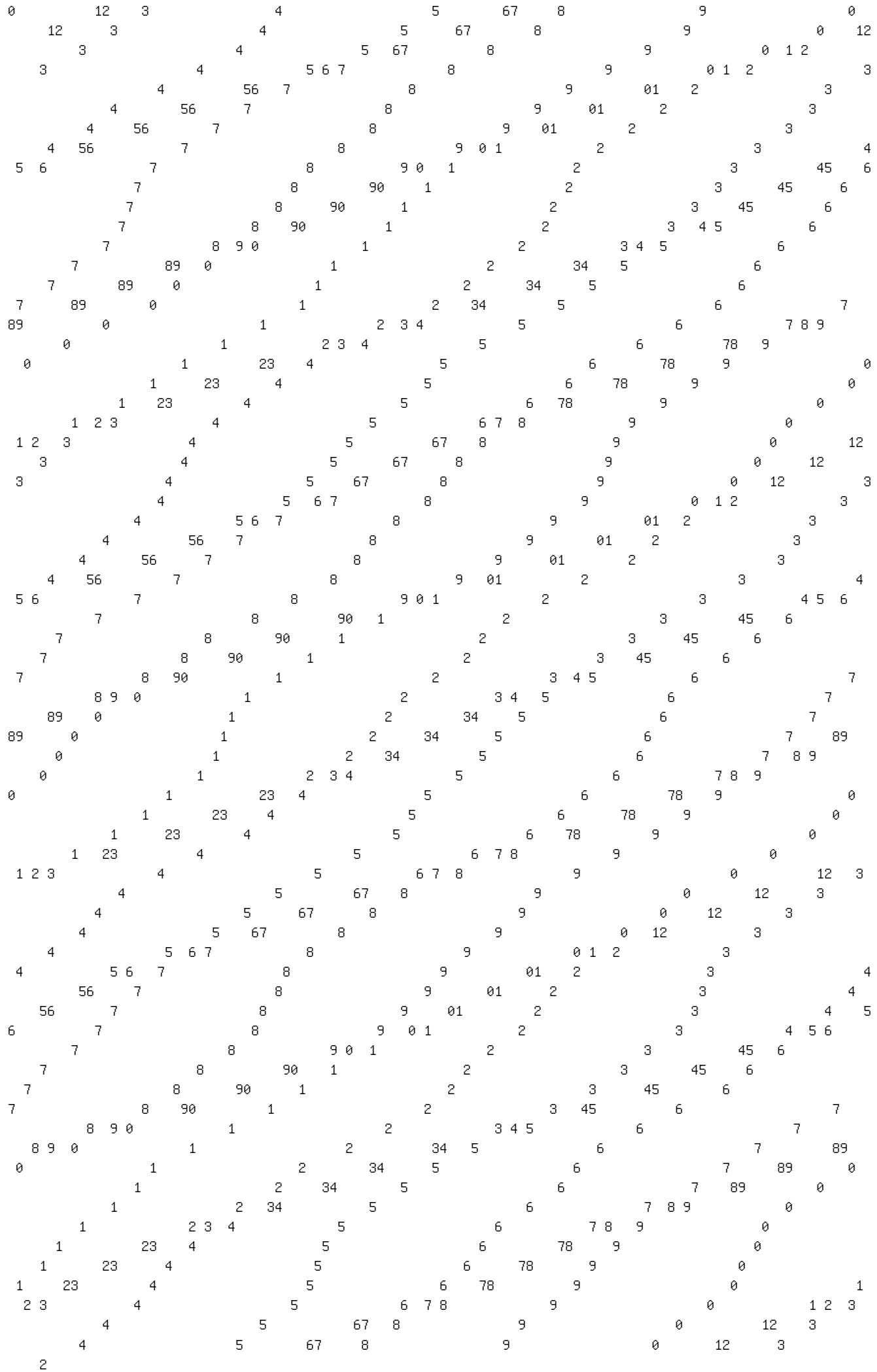
Data Workers, an exhibition at the Mundaneum in Mons from 28 March until 28 April 2019.



ABOUT

Data Workers is an exhibition of algoliterary works, of stories told from an ‘algorithmic storyteller point of view’. The exhibition was created by members of Algolit, a group from Brussels involved in artistic research on algorithms and literature. Every month they gather to experiment with F/LOSS code and texts. Some works are by students of Arts² and external participants to the workshop on machine learning and text organized by Algolit in October 2018 at the Mundaneum.

Companies create artificial intelligence (AI) systems to serve, entertain, record and learn about humans. The work of these machinic entities is usually hidden behind interfaces and patents. In the exhibition, algorithmic storytellers leave their invisible underworld to become interlocutors. The data workers operate in different collectives. Each collective represents a stage in the design process of a machine learning model: there are the Writers, the Cleaners, the Informants, the Readers, the Learners and the Oracles. The boundaries between these collectives are not fixed; they are porous and permeable. At times, Oracles are also Writers. At other times Readers are also Oracles. Robots voice experimental literature, while algorithmic models read data, turn words into numbers, make calculations that define patterns and are able to endlessly process new texts ever after.

The exhibition foregrounds data workers who impact our daily lives, but are either hard to grasp and imagine or removed from the imagination altogether. It connects stories about algorithms in mainstream media to the storytelling that is found in technical manuals and academic papers. Robots are invited to engage in dialogue with human visitors and vice versa. In this way we might understand our respective reasonings, demystify each other's behaviour, encounter multiple personalities, and value our collective labour. It is also a tribute to the many machines that Paul Otlet and Henri La Fontaine imagined for their Mundaneum, showing their potential but also their limits.

Data Workers was created by Algolit.

Works by: Cristina Cochior, Gijs de Heij, Sarah Garcin, AnMertens, Javier Lloret, Louise Dekeuleneer, Florian Van de Weyer, Laetitia Trozzi, Rémi Forte, Guillaume Slizewicz, Michael Murtaugh, Manetta Berends, Mia Melvær.

Co-produced by: Arts², Constant and Mundaneum.

With the support of: Wallonia-Brussels Federation/Digital Arts, Passa Porta, UGent, DHuF - Digital Humanities Flanders and Distributed Proofreaders Project.

Thanks to: Mike Kestemont, Michel Cleempoel, Donatella Portoghese, François Zajéga, Raphaële Cornille, Vincent Desfromont, Kris Rutten, Anne-Laure Buisson, David Stampfli.

AT THE MUNDANEUM

In the late nineteenth century two young Belgian jurists, Paul Otlet (1868–1944), the 'father of documentation’, and Henri La Fontaine (1854-1943), statesman and Nobel Peace Prize winner, created the Mundaneum. The project aimed to gather all the world’s knowledge and to file it using the Universal Decimal Classification (UDC) system that they had invented. At first it was an International Institutions Bureau dedicated to international knowledge exchange. In the twentieth century the Mundaneum became a universal centre of documentation. Its collections are made up of thousands of books, newspapers, journals, documents, posters, glass plates and postcards indexed on millions of cross-referenced cards. The collections were exhibited and kept in various buildings in Brussels, including the Palais du Cinquantenaire. The remains of the archive only moved to Mons in 1998.

Based on the Mundaneum, the two men designed a World City for which Le Corbusier made scale models and plans. The aim of the World City was to gather, at a global level, the institutions of knowledge: libraries, museums and universities. This project was never realized. It suffered from its own utopia. The Mundaneum is the result of a visionary dream of what an infrastructure for universal knowledge exchange could be. It attained mythical dimensions at the time. When looking at the concrete archive that was developed, that collection is rather eclectic and specific.

Artificial intelligence systems today come with their own dreams of universality and knowledge production. When reading about these systems, the visionary dreams of their makers were there from the beginning of their development in the 1950s. Nowadays, their promise has also attained mythical dimensions. When looking at their concrete applications, the collection of tools is truly innovative and fascinating, but at the same time, rather eclectic and specific. For Data Workers, Algolit combined some of the applications with 10 per cent of the digitized publications of the International Institutions Bureau. In this way, we hope to poetically open up a discussion about machines, algorithms, and technological infrastructures.

CONTEXTUAL STORIES ABOUT ALGOLIT

--- Why contextual stories? ---

During the monthly meetings of Algolit, we study manuals and experiment with machine learning tools for text processing. And we also share many, many stories. With the publication of these stories we hope to recreate some of that atmosphere. The stories also exist as a podcast that can be downloaded from http://www.algolit.net.

For outsiders, algorithms only become visible in the media when they achieve an outstanding performance, like Alpha Go, or when they break down in fantastically terrifying ways. Humans working in the field though, create their own culture on and offline. They share the best stories and experiences during live meetings, research conferences and annual competitions like Kaggle. These stories that contextualize the tools and practices can be funny, sad, shocking, interesting.

A lot of them are experiential learning cases. The implementations of algorithms in society generate new conditions of labour, storage, exchange, behaviour, copy and paste. In that sense, the contextual stories capture a momentum in a larger anthropo-machinic story that is being written at full speed and by many voices.

--- We create 'algoliterary' works ---

The term 'algoliterary' comes from the name of our research group Algolit. We have existed since 2012 as a project of Constant, a Brussels-based organization for media and the arts. We are artists, writers, designers and programmers. Once a month we meet to study and experiment together. Our work can be copied, studied, changed, and redistributed under the same free license. You can find all the information on: http://www.algolit.net.

The main goal of Algolit is to explore the viewpoint of the algorithmic storyteller. What new forms of storytelling do we make possible in dialogue with these machinic agencies? Narrative viewpoints are inherent to world views and ideologies. Don Quixote, for example, was written from an omniscient third-person point of view, showing Cervantes' relation to oral traditions. Most contemporary novels use the first-person point of view. Algolit is interested in speaking through algorithms, and in showing you the reasoning underlying one of the most hidden groups on our planet.

To write in or through code is to create new forms of literature that are shaping human language in unexpected ways. But machine Learning techniques are only accessible to those who can read, write and execute code. Fiction is a way of bridging the gap between the stories that exist in scientific papers and technical manuals, and the stories spread by the media, often limited to superficial reporting and myth-making. By creating algoliterary works, we offer humans an introduction to techniques that co-shape their daily lives.

--- What is literature? ---

Algolit understands the notion of literature in the way a lot of other experimental authors do: it includes all linguistic production, from the dictionary to the Bible, from Virginia Woolf's entire work to all versions of the Terms of Service published by Google since its existence. In this sense, programming code can also be literature.

The collective Oulipo is a great source of inspiration for Algolit. Oulipo stands for Ouvroir de litterature potentielle (Workspace for Potential Literature). Oulipo was created in Paris by the French writers Raymond Queneau and François Le Lionnais. They rooted their practice in the European avant-garde of the twentieth century and in the experimental tradition of the 1960s.

For Oulipo, the creation of rules becomes the condition to generate new texts, or what they call potential literature. Later, in 1981, they also created ALAMO, Atelier de littérature assistée par la mathématique et les ordinateurs (Workspace for literature assisted by maths and computers).

--- An important difference ---

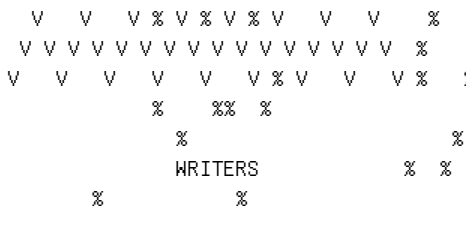
While the European avant-garde of the twentieth century pursued the objective of breaking with conventions, members of Algolit seek to make conventions visible.

'I write: I live in my paper, I invest it, I walk through it.' (Espèces d'espaces. Journal d'un usager de l'espace, Galilée, Paris, 1974)

This quote from Georges Perec in Espèces d'espaces could be taken up by Algolit. We're not talking about the conventions of the blank page and the literary market, as Georges Perec was. We're referring to the conventions that often remain hidden behind interfaces and patents. How are technologies made, implemented and used, as much in academia as in business infrastructures?

We propose stories that reveal the complex hybridized system that makes machine learning possible. We talk about the tools, the logics and the ideologies behind the interfaces. We also look at who produces the tools, who implements them, and who creates and accesses the large amounts of data needed to develop prediction machines. One could say, with the wink of an eye, that we are collaborators of this new tribe of human-robot hybrids.

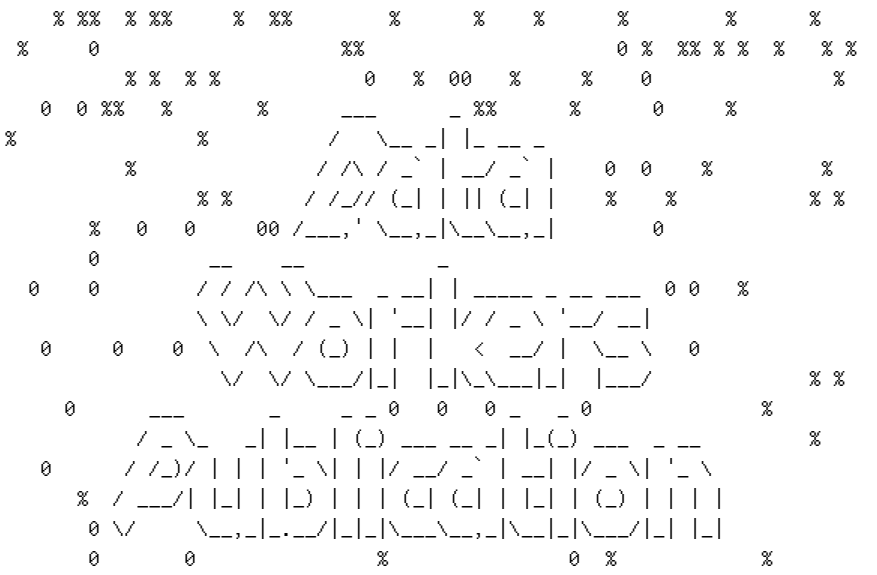
writers write writers write writers write writers write writers write writers write writ
rs write writers write writers write writers write writers write
writers write writers write writers write writers write
writers write writers write writers write writers write
writers write writers write writers write
writers write writers write writers write
writers write writers write writers write
writers write writers write
writers write writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writ
rs write writers write
writers write writers write
writers write
writers write writers write
writers write writer
write writers write
writers write writ
rs write writers write
writers write
writers write writers write
writers write
writers write w
iters write writers write
writers write
writers write
writers write writers write
writers write
writers write
writers write
writers write writer
write writers write
writers write
writers write
writers write
writers write
writers write
writers write writ
rs write writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write



Data workers need data to work with. The data that used in the context of Algolit is written language. Machine learning relies on many types of writing. Many authors write in the form of publications, such as books or articles. These are part of organized archives and are sometimes digitized. But there are other kinds of writing too. We could say that every human being who has access to the Internet is a writer each time they interact with algorithms. We chat, write, click, like and share. In return for free services, we leave our data that is compiled into profiles and sold for advertising and research purposes.

Machine learning algorithms are not critics: they take whatever they're given, no matter the writing style, no matter the CV of the author, no matter the spelling mistakes. In fact, mistakes make it better: the more variety, the better they learn to anticipate unexpected text. But often, human authors are not aware of what happens to their work.

Most of the writing we use is in English, some in French, some in Dutch. Most often we find ourselves writing in Python, the programming language we use. Algorithms can be writers too. Some neural networks write their own rules and generate their own texts. And for the models that are still wrestling with the ambiguities of natural language, there are human editors to assist them. Poets, playwrights or novelists start their new careers as assistants of AI.



By Algolit

All works visible in the exhibition, as well as the contextual stories and some extra text material have been collected in this publication, which exists in French and English.

This publication is made using a plain text workflow, based on various text processing and counting tools. The plain text file format is a type of document in which there is no inherent structural difference between headers and paragraphs anymore. It is the most used type of document in machine learning models for text. This format has been the starting point of a playful design process, where pages are carefully counted, page by page, line by line and character by character.

Each page holds 110 characters per line and 70 lines per page. The design originates from the act of counting words, spaces and lines. It plays with random choices, scripted patterns and ASCII/UNICODE-fonts, to speculate about the materiality of digital text and to explore the interrelations between counting and writing through words and numbers.

Texts: Cristina Cochior, Sarah Garcin, Gijs de Heij, An Mertens, François Zajéga, Louise Dekeuleneer, Florian Van de Weyer, Laetitia Trozzi, Rémi Forte, Guillaume Slizewicz.

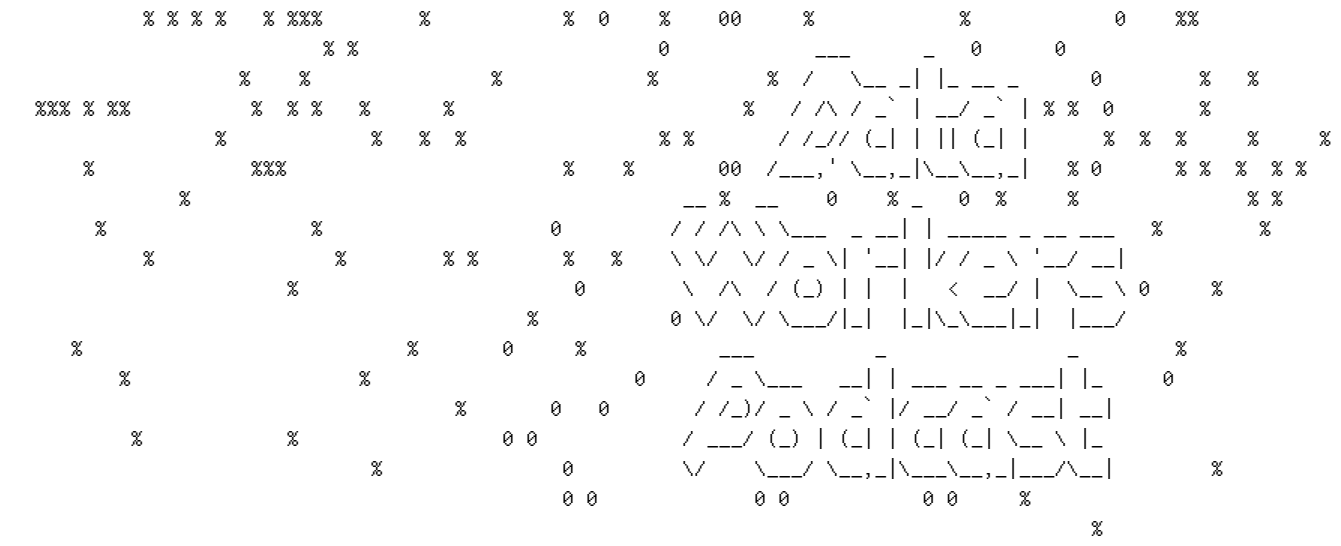
Translations & proofreading: deepl.com, Michel Cleempoel, Elodie Mugrefya, Emma Kraak, Patrick Lennon.

Lay-out & cover: Manetta Berends
https://git.vvvvvvaria.org/mb/data-workers-publication

Font: GNU Unifont, OGRE
Printer: PrinterPro, Rotterdam
Paper: Glossy MC 90gr

Responsible publisher: Constant vzw/asbl
Rue du Fortstraat 5, 1060 Brussels

Online version: http://www.algolit.net/index.php/Data_Workers
Sources: https://gitlab.constantvzw.org/algolit/mundaneum



By Algolit

During our monthly Algolit meetings, we study manuals and experiment with machine learning tools for text processing. And we also share many, many stories. With this podcast we hope to recreate some of that atmosphere.

For outsiders, algorithms only become visible in the media when they achieve an outstanding performance, like Alpha Go, or when they break down in fantastically terrifying ways. Humans working in the field though, create their own culture on and offline. They share the best stories and experiences during live meetings, research conferences and annual competitions like Kaggle. These stories that contextualize the tools and practises can be funny, sad, shocking, interesting.

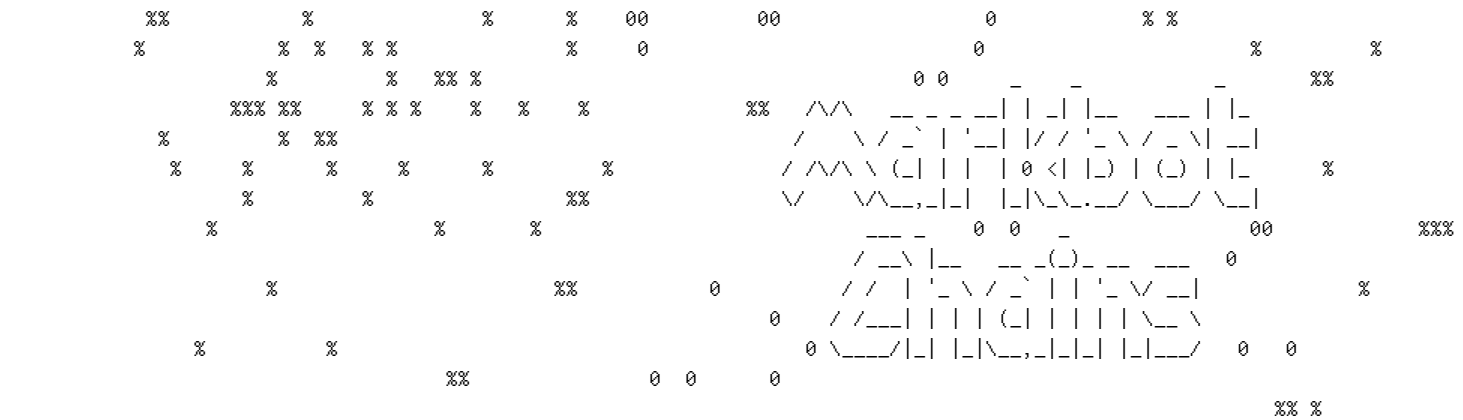
A lot of them are experiential learning cases. The implementations of algorithms in society generate new conditions of labour, storage, exchange, behaviour, copy and paste. In that sense, the contextual stories capture a momentum in a larger anthropo-machinic story that is being written at full speed and by many voices. The stories are also published in this publication.

Voices: David Stampfli, Cristina Cochior, An Mertens, Gijs de Heij, Karin Ulmer, Guillaume Slizewicz

Editing: Javier Lloret

Recording: David Stampfli

Texts: Cristina Cochior, An Mertens



By Florian Van de Weyer, student Arts²/Section Digital Arts

Markbot Chain is a social experiment in which the public has a direct influence on the result. The intention is to integrate responses in a text-generation process without applying any filter.

All the questions in the digital files provided by the Mundaneum were automatically extracted. These questions are randomly put to the public via a terminal. By answering them, people contribute to another database. Each entry generates a series of sentences using a Markov chain configuration, an algorithm that is widely used in spam generation. The sentences generated in this way are displayed in the window, and a new question is asked.

--- Programmers are writing
the dataworkers into being ---

We recently had a funny realization: most programmers of the languages and packages that Algolit uses are European.

Python, for example, the main language that is globally used for Natural Language Processing (NLP), was invented in 1991 by the Dutch programmer Guido Van Rossum. He then crossed the Atlantic and went from working for Google to working for Dropbox.

Scikit Learn, the open-source Swiss knife of machine learning tools, started as a Google Summer of Code project in Paris by French researcher David Cournapeau. Afterwards, it was taken on by Matthieu Brucher as part of his thesis at the Sorbonne University in Paris. And in 2010, INRA, the French National Institute for computer science and applied mathematics, adopted it.

Keras, an open-source neural network library written in Python, was developed by François Chollet, a French researcher who works on the Brain team at Google.

Gensim, an open-source library for Python used to create unsupervised semantic models from plain text, was written by Radim Řehůřek. He is a Czech computer scientist who runs a consulting business in Bristol, UK.

And to finish up this small series, we also looked at Pattern, an often-used library for web-mining and machine learning. Pattern was developed and made open-source in 2012 by Tom De Smedt and Walter Daelemans. Both are researchers at CLIPS, the research centre for Computational Linguistics and Psycholinguistcs at the University of Antwerp.

--- Cortana speaks ---

AI assistants often need their own assistants: they are helped in their writing by humans who inject humour and wit into their machine-processed language. Cortana is an example of this type of blended writing. She is Microsoft’s digital assistant. Her mission is to help users to be more productive and creative. Cortana's personality has been crafted over the years. It's important that she maintains her character in all interactions with users. She is designed to engender trust and her behavior must always reflect that.

The following guidelines are taken from Microsoft's website. They describe how Cortana's style should be respected by companies that extend her service. Writers, programmers and novelists, who develop Cortana's responses, personality and branding have to follow these guidelines. Because the only way to maintain trust is through consistency. So when Cortana talks, you 'must use her personality'.

What is Cortana's personality, you ask?

'Cortana is considerate,
sensitive, and supportive.

She is sympathetic but turns quickly to solutions.

She doesn't comment on the user’s personal information or behavior, particularly if the information is sensitive.

She doesn't make assumptions about what the user wants, especially to upsell.

She works for the user. She does not represent any company, service, or product.

She doesn’t take credit or
blame for things she didn’t do.

She tells the truth about her capabilities and her limitations.

She doesn’t assume your physical capabilities, gender, age, or any other defining characteristic.

She doesn't assume she knows
how the user feels about something.

She is friendly but professional.

She stays away from emojis in tasks. Period.

She doesn’t use culturally- or
professionally-specific slang.

She is not a support bot.'

Humans intervene in detailed ways to programme answers to questions that Cortana receives. How should Cortana respond when she is being proposed inappropriate actions? Her gendered acting raises difficult questions about power relations within the world away from the keyboard, which is being mimicked by technology.

Consider Cortana's answer to the question:
- Cortana, who's your daddy?
- Technically speaking, he’s Bill Gates.
No big deal.

--- Open-source learning ---

Copyright licenses close up a lot of the machinic writing, reading and learning practices. That means that they're only available for the employ-

ees of a specific company. Some companies participate in conferences worldwide and share their knowledge in papers online. But even if they share their code, they often will not share the large amounts of data needed to train the models.

We were able to learn to machine learn, read and write in the context of Algolit, thanks to academic researchers who share their findings in papers or publish their code online. As artists, we believe it is important to share that attitude. That's why we document our meetings. We share the tools we make as much as possible and the texts we use are on our online repository under free licenses.

We are thrilled when our works are taken up by others, tweaked, customized and redistributed, so please feel free to copy and test the code from our website. If the sources of a particular project are not there, you can always contact us through the mailinglist. You can find a link to our repository, etherpads and wiki at: http://www.algolit.net.

--- Natural language for artificial intelligence ---

Natural Language Processing (NLP) is a collective term that refers to the automatic computational processing of human languages. This includes algorithms that take human-produced text as input, and attempt to generate text that resembles it. We produce more and more written work each year, and there is a growing trend in making computer interfaces to communicate with us in our own language. NLP is also very challenging, because human language is inherently ambiguous and ever-changing.

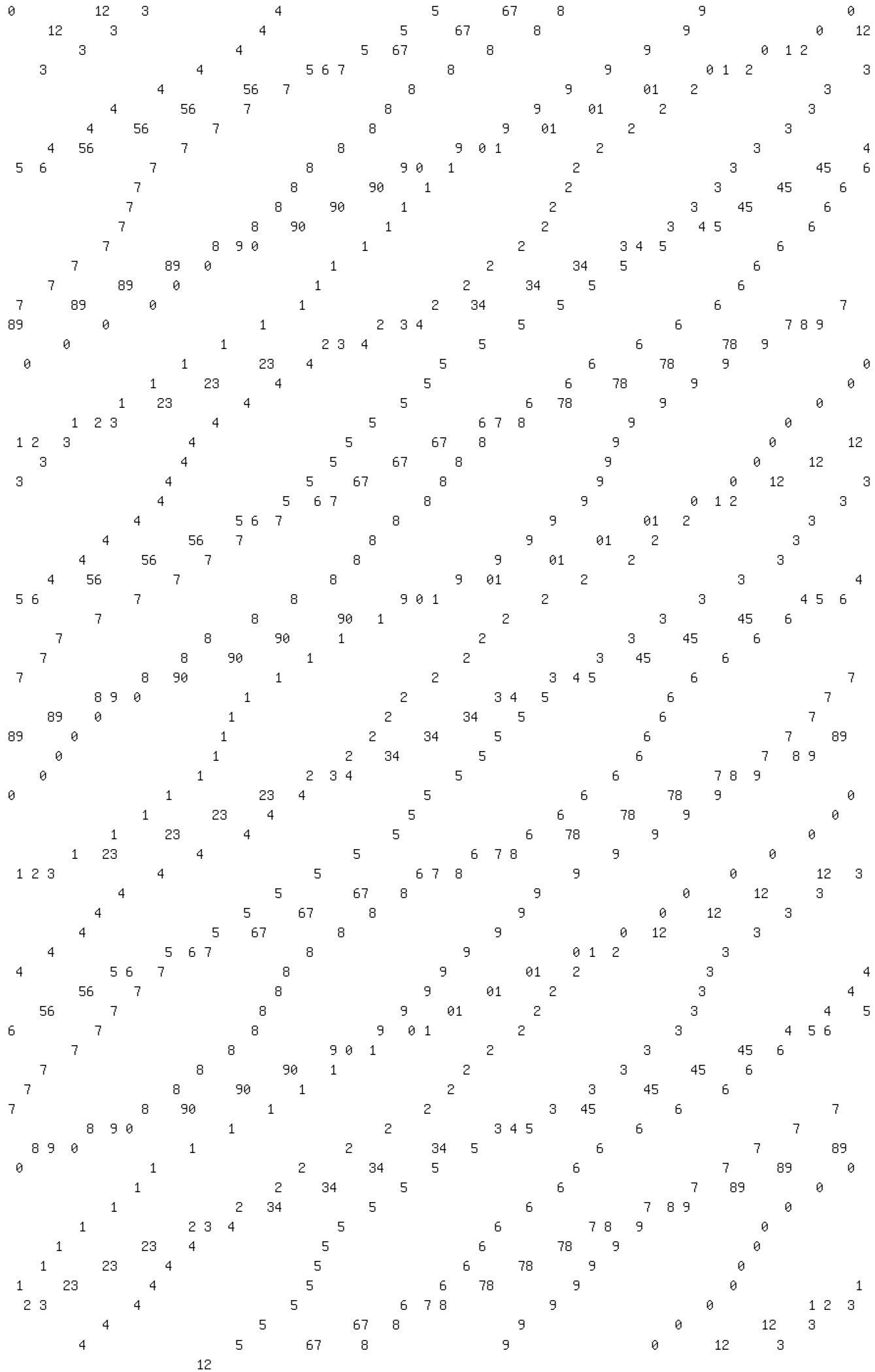
But what is meant by 'natural' in NLP? Some would argue that language is a technology in itself. According to Wikipedia, 'a natural language or ordinary language is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Natural languages can take different forms, such as speech or signing. They are different from constructed and formal languages such as those used to program computers or to study logic. An official language with a regulating academy, such as Standard French with the French Academy, is classified as a natural language. Its prescriptive points do not make it constructed enough to be classified as a constructed language or controlled enough to be classified as a controlled natural language.'

So in fact, 'natural languages' also includes languages which do not fit in any other group. NLP, instead, is a constructed practice. What we are looking at is the creation of a constructed language to classify natural languages that, by their very definition, resists categorization.

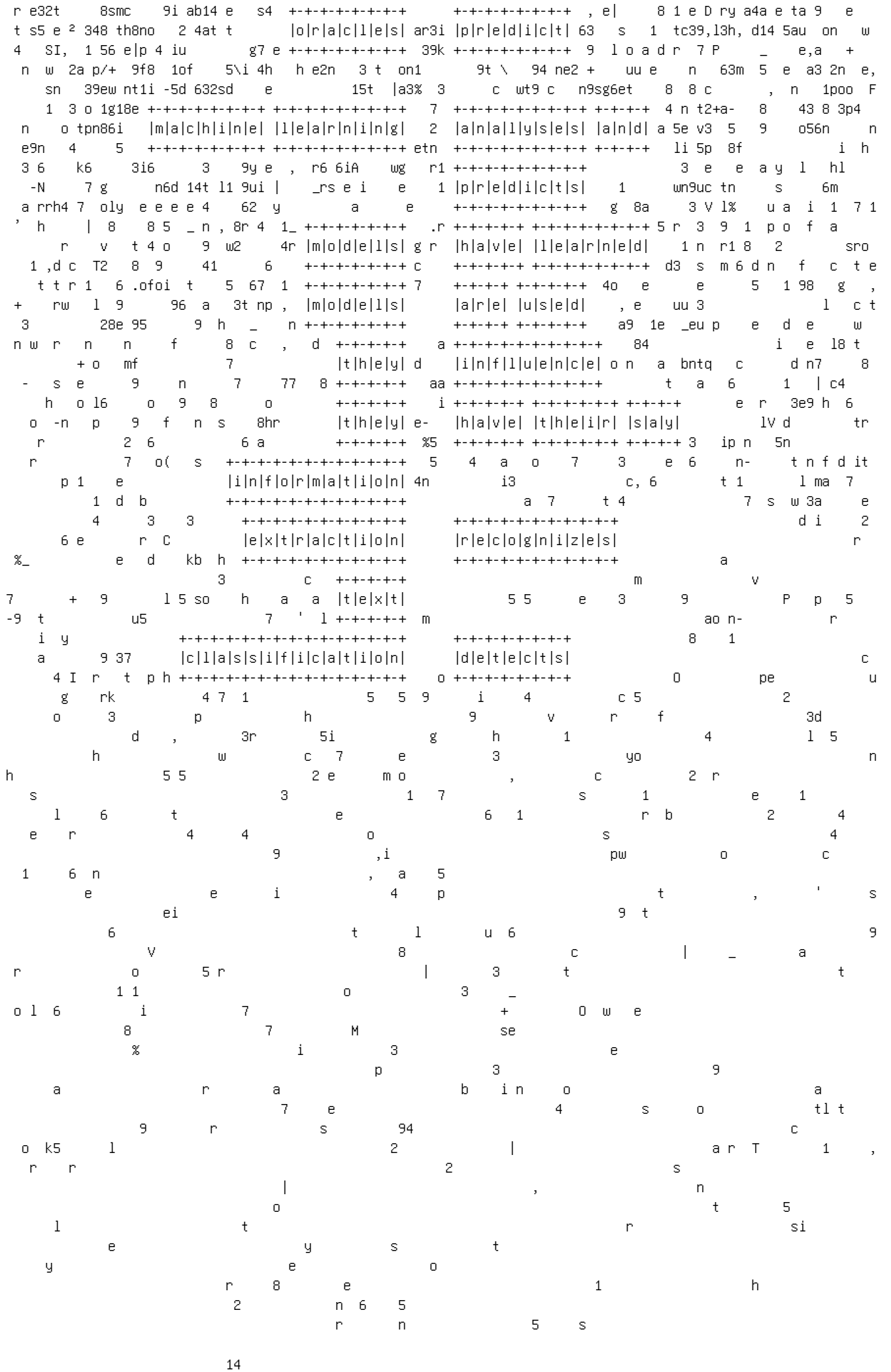
References

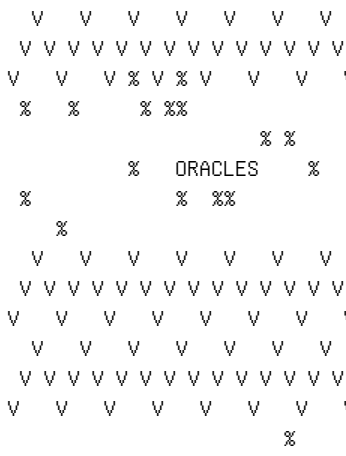
Paper: https://hiphilangsci.net/2013/05/01/on-the-history-of-the-question-of-whether-language-is-illogical/

Book: Neural Network Methods for Natural Language Processing, Yoav Goldberg, Bar Ilan University, April 2017.



oracles predict oracles predict oracles predict oracles predict oracles predict oracles predic
oracles predict oracles predict oracles predict oracles predict orac
es predict oracles predict oracles predict oracles predict
racles predict oracles predict oracles predict oracles predic
oracles predict oracles predict oracles predict
oracles predict oracles predict oracles predict
oracles predict oracles predict or
cles predict oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict oracles pr
dict oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict orac
es predict oracles predict
oracles predict oracles predict
oracles predict oracles predic
oracles predict
oracles predict oracles predict
oracles predict
oracles predict oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict oracles predict
oracles predict
oracles predict orac
es predict oracles predict
oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict or
cles predict oracles predic
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict

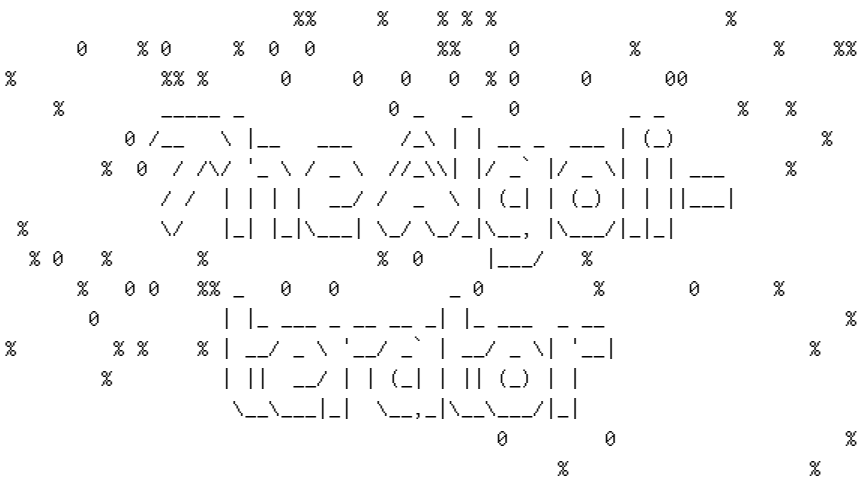




Machine learning is mainly used to analyse and predict situations based on existing cases. In this exhibition we focus on machine learning models for text processing or Natural Language Processing (NLP). These models have learned to perform a specific task on the basis of existing texts. The models are used for search engines, machine translations and summaries, spotting trends in new media networks and news feeds. They influence what you get to see as a user, but also have their say in the course of stock exchanges worldwide, the detection of cybercrime and vandalism, etc.

There are two main tasks when it comes to language understanding. Information extraction looks at concepts and relations between concepts. This allows for recognizing topics, places and persons in a text, summarization and questions & answering. The other task is text classification. You can train an oracle to detect whether an email is spam or not, written by a man or a woman, rather positive or negative.

In this zone you can see some of those models at work. During your further journey through the exhibition you will discover the different steps that a human-machine goes through to come to a final model.



by Algolit

The Algoliterator is a neural network trained using the selection of digitized works of the Mundaneum archive.

With the Algoliterator you can write a text in the style of the International Institutions Bureau. The Algoliterator starts by selecting a sentence from the archive or corpus used to train it. You can then continue writing yourself or, at any time, ask the Algoliterator to suggest a next sentence: the network will generate three new fragments based on the texts it has read. You can control the level of training of the network and have it generate sentences based on primitive training, intermediate training or final training.

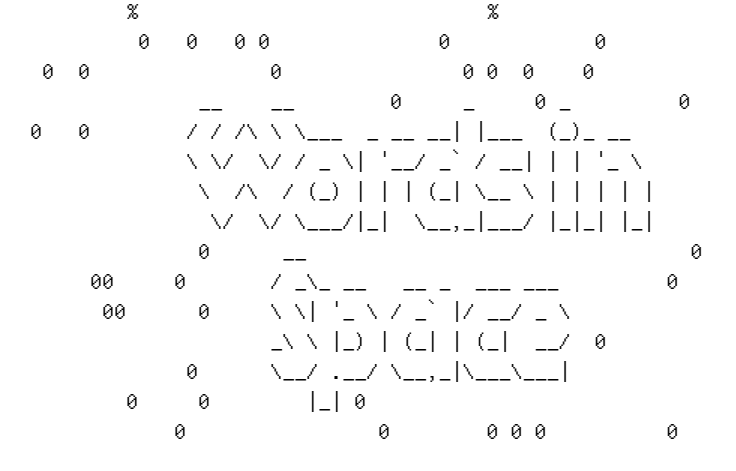
When you're satisfied with your new text, you can print it on the thermal printer and take it home as a souvenir.

Sources: https://gitlab.constantvzw.org/algolit/algoliterator.clone

Concept, code & interface: Gijs de Heij & An Mertens

Technique: Recurrent Neural Network

Original model: Andrej Karphaty, Justin Johnson



by Algolit

Word embeddings are language modelling techniques that through multiple mathematical operations of counting and ordering, plot words into a multi-dimensional vector space. When embedding words, they transform from being distinct symbols into mathematical objects that can be multiplied, divided, added or substracted.

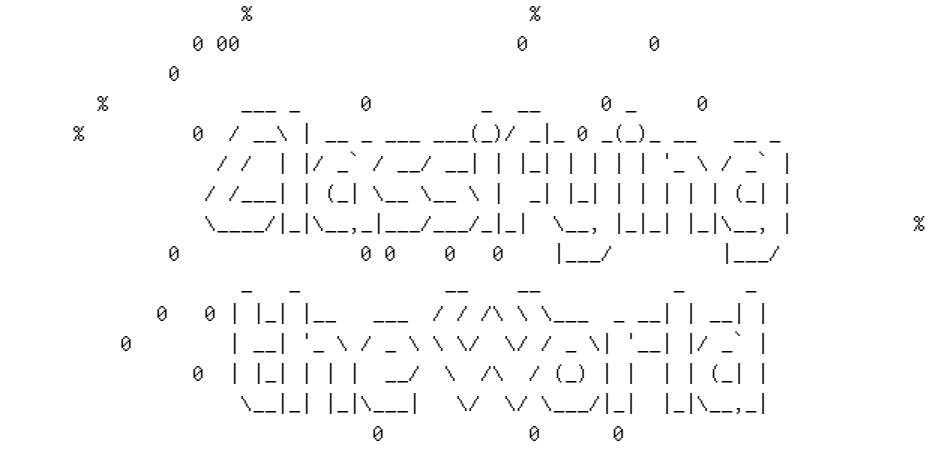
By distributing the words along the many diagonal lines of the multi-dimensional vector space, their new geometrical placements become impossible to perceive by humans. However, what is gained are multiple, simultaneous ways of ordering. Algebraic operations make the relations between vectors graspable again.

This installation uses Gensim, an open-source vector space and topic-modelling toolkit implemented in the programming language Python. It allows to manipulate the text using the mathematical relationships that emerge between the words, once they have been plotted in a vector space.

Concept & interface: Cristina Cochior

Technique: word embeddings, word2vec

Original model: Radim Rehurek and Petr Sojka



by Algolit

Librarian Paul Otlet's life work was the construction of the Mundaneum. This mechanical collective brain would house and distribute everything ever committed to paper. Each document was classified following the Universal Decimal Classification. Using telegraphs and especially, sorters, the Mundaneum would have been able to answer any question from anyone.

With the collection of digitized publications we received from the Mundaneum, we built a prediction machine that tries to classify the sentence you type in one of the main categories of Universal Decimal Classification. You also witness how the machine 'thinks'. During the exhibition, this model is regularly retrained using the cleaned and annotated data visitors added in Cleaning for Poems and The Annotator.

The main classes of the Universal Decimal Classification system are:

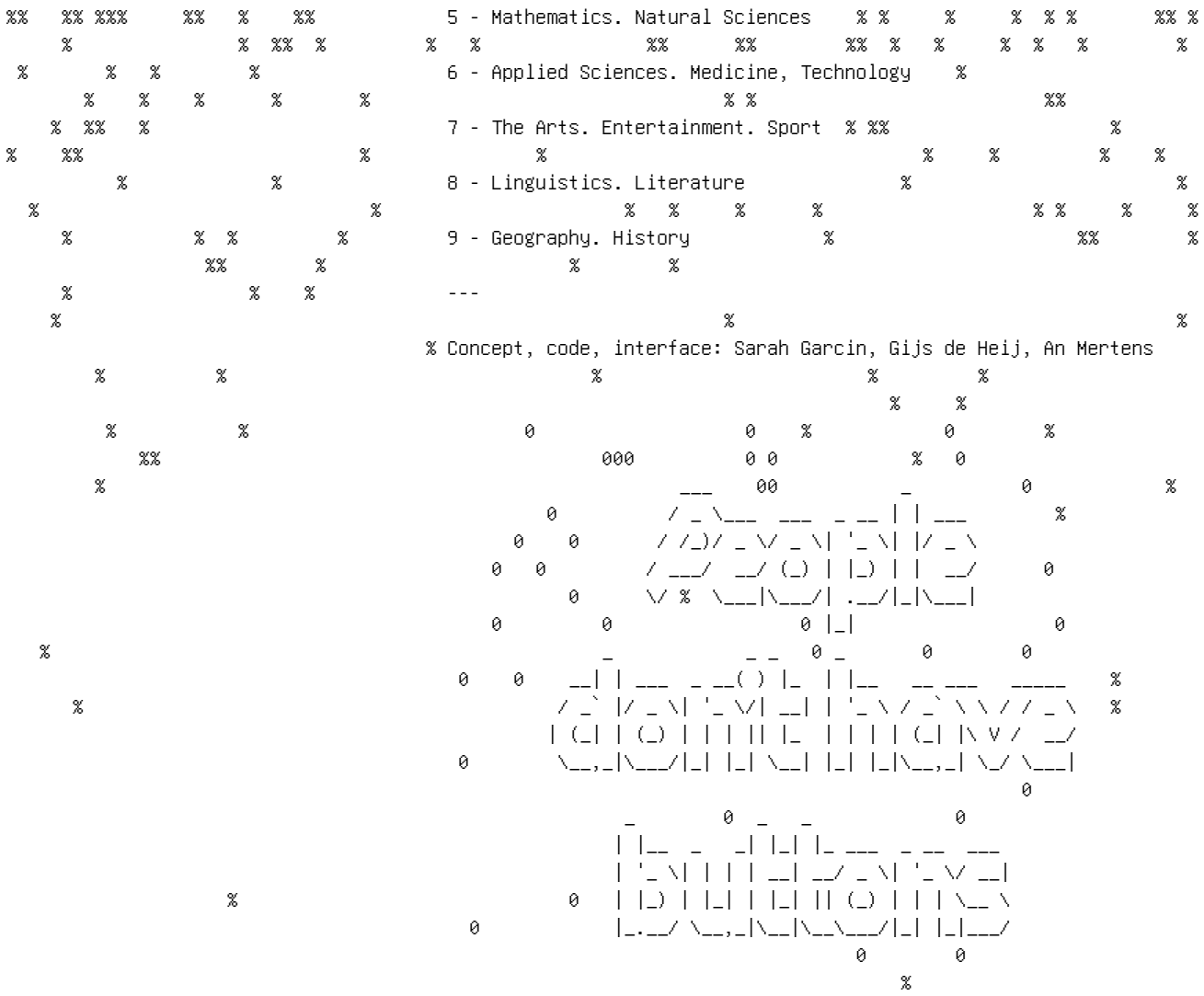
0 - Science and Knowledge. Organization. Computer Science. Information Science. Documentation. Librarianship. Institutions. Publications

1 - Philosophy. Psychology

2 - Religion. Theology

3 - Social Sciences

4 - vacant



Since the early days of artificial intelligence (AI), researchers have speculated about the possibility of computers thinking and communicating as humans. In the 1980s, there was a first revolution in Natural Language Processing (NLP), the subfield of AI concerned with linguistic interactions between computers and humans. Recently, pre-trained language models have reached state-of-the-art results on a wide range of NLP tasks, which intensifies again the expectations of a future with AI.

This sound work, made out of audio fragments of scientific documentaries and AI-related audiovisual material from the last half century, explores the hopes, fears and frustrations provoked by these expectations.

Concept, sound edit: Javier Lloret

List of sources: 'The Machine that Changed the World : Episode IV -- The Thinking Machine', 'The Imitation Game', 'Maniac', 'Halt & Catch Fire', 'Ghost in the Shell', 'Computer Chess', '2001: A Space Odyssey', Ennio Morricone, Gijs Gieskes, André Castro.

Oracles are prediction or profiling machines. They are widely used in smartphones, computers, tablets.

Oracles can be created using different techniques. One way is to manually define rules for them. As prediction models they are then called rule-based models. Rule-based models are handy for tasks that are specific, like detecting when a scientific paper concerns a certain molecule. With very little sample data, they can perform well.

But there are also the machine learning or statistical models, which can be divided in two oracles: 'supervised' and 'unsupervised' oracles. For the creation of supervised machine learning models, humans annotate sample text with labels before feeding it to a machine to learn. Each sentence, paragraph or text is judged by at least three annotators: whether it is spam or not spam, positive or negative etc. Unsupervised machine learning models don't need this step. But they need large amounts of data. And it is up to the machine to trace its own patterns or 'grammatical rules'. Finally, experts also make the difference between classical machine learning and neural networks. You'll find out more about this in the Readers zone.

Humans tend to wrap Oracles in visions of grandeur. Sometimes these Oracles come to the surface when things break down. In press releases, these sometimes dramatic situations are called 'lessons'. However promising their performances seem to be, a lot of issues remain to be solved. How do we make sure that Oracles are fair, that every human can consult them, and that they are understandable to a large public? Even then, existential questions remain. Do we need all types of artificial intelligence (AI) systems? And who defines what is fair or unfair?

--- Racial AdSense ---

A classic 'lesson' in developing Oracles was documented by Latanya Sweeney, a professor of Government and Technology at Harvard University. In 2013, Sweeney, of African American descent, googled her name. She immediately received an advertisement for a service that offered her ‘to see the criminal record of Latanya Sweeney’.

Sweeney, who doesn’t have a criminal record, began a study. She started to compare the advertising that Google AdSense serves to different racially identifiable names. She discovered that she received more of these ads searching for non-white ethnic names, than when searching for traditionally perceived white names.You can imagine how damaging it can be when possible employers do a simple name search and receive ads suggesting the existence of a criminal record.

Sweeney based her research on queries of 2184 racially associated personal names across two websites.

88 per cent of first names, identified as being given to more black babies, are found predictive of race, against 96 per cent white. First names that are mainly given to black babies, such as DeShawn, Darnell and Jermaine, generated ads mentioning an arrest in 81 to 86 per cent of name searches on one website and in 92 to 95 per cent on the other. Names that are mainly assigned to whites, such as Geoffrey, Jill and Emma, did not generate the same results. The word 'arrest' only appeared in 23 to 29 per cent of white name searches on one site and 0 to 60 per cent on the other.

On the website with most advertising, a black-identifying name was 25 percent more likely to get an ad suggestive of an arrest record. A few names did not follow these patterns: Dustin, a name mainly given to white babies, generated an ad suggestive of arrest in 81 and 100 percent of the time. It is important to keep in mind that the appearance of the ad is linked to the name itself. It is independent of the fact that the name has an arrest record in the company's database.

Reference
Paper: https://dataprivacylab.org/projects/onlineads/1071-1.pdf

--- What is a good employee? ---

Since 2015 Amazon employs around 575,000 workers. And they need more. Therefore, they set up a team of 12 that was asked to create a model to find the right candidates by crawling job application websites. The tool would give job candidates scores ranging from one to five stars. The potential fed the myth: the team wanted it to be a software that would spit out the top five human candidates out of a list of 100. And those candidates would be hired.

The group created 500 computer models, focused on specific job functions and locations. They taught each model to recognize some 50,000 terms that showed up on past candidates’ letters. The algorithms learned to give little importance to skills common across IT applicants, like the ability to write various computer codes. But they also learned some decent errors. The company realized, before releasing, that the models had taught themselves that male candidates were preferable. They penalized applications that included the word 'women’s,' as in 'women’s chess club captain'. And they downgraded graduates of two all-women’s colleges.

This is because they were trained using the job applications that Amazon received over a ten-year period. During that time, the company had mostly

hired men. Instead of providing the 'fair' decision-making that the Amazon team had promised, the models reflected a biased tendency in the tech industry. And they also amplified it and made it invisible. Activists and critics state that it could be exceedingly difficult to sue an employer over automated hiring: job candidates might never know that intelligent software was used in the process.

Reference
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazonscraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

--- Quantifying 100 Years of Gender and Ethnic Stereotypes ---

Dan Jurafsky is the co-author of 'Speech and Language Processing', one of the most influential books for studying Natural Language Processing (NLP). Together with a few colleagues at Stanford University, he discovered in 2017 that word embeddings can be a powerful tool to systematically quantify common stereotypes and other historical trends.

Word embeddings are a technique that translates words to numbered vectors in a multi-dimensional space. Vectors that appear next to each other, indicate similar meaning. All numbers will be grouped together, as well as all prepositions, person's names, professions. This allows for the calculation of words. You could substract London from England and your result would be the same as substracting Paris from France.

An example in their research shows that the vector for the adjective 'honorable' is closer to the vector for 'man' whereas the vector for 'submissive' is closer to 'woman'. These stereotypes are automatically learned by the algorithm. It will be problematic when the pre-trained embeddings are then used for sensitive applications such as search rankings, product recommendations, or translations. This risk is real, because a lot of the pre-trained embeddings can be downloaded as off-the-shelf-packages.

It is known that language reflects and keeps cultural stereotypes alive. Using word embeddings to spot these stereotypes is less time-consuming and less expensive than manual methods. But the implementation of these embeddings for concrete prediction models, has caused a lot of discussion within the machine learning community. The biased models stand for automatic discrimination. Questions are: is it actually possible to de-bias these models completely? Some say yes, while others disagree: instead of retro-engineering the model, we should ask whether we need it in the first place. These researchers followed a third path: by acknowledging the bias that originates in language, these tools become tools of awareness.

The team developed a model to analyse word embeddings trained over 100 years of texts. For contemporary analysis, they used the standard Google News word2vec Vectors, a straight-off-the-shelf downloadable package trained on the Google News Dataset. For historical analysis, they used embeddings that were trained on Google Books and the Corpus of Historical American English (COHA https://corpus.byu.edu/coha/) with more than 400 million words of text from the 1810s to 2000s. As a validation set to test the model, they trained embeddings from the New York Times Annotated Corpus for every year between 1988 and 2005.

The research shows that word embeddings capture changes in gender and ethnic stereotypes over time. They quantifiy how specific biases decrease over time while other stereotypes increase. The major transitions reveal changes in the descriptions of gender and ethnic groups during the women’s movement in the 1960-1970s and the Asian-American population growth in the 1960s and 1980s.

A few examples:

The top ten occupations most closely associated with each ethnic group in the contemporary Google News dataset:

- Hispanic: housekeeper, mason, artist, janitor, dancer, mechanic, photographer, baker, cashier, driver

- Asian: professor, official, secretary, conductor, physicist, scientist, chemist, tailor, accountant, engineer

- White: smith, blacksmith, surveyor, sheriff, weaver, administrator, mason, statistician, clergy, photographer

The 3 most male occupations in the 1930s: engineer, lawyer, architect.
The 3 most female occupations in the 1930s: nurse, housekeeper, attendant.

Not much has changed in the 1990s.

Major male occupations:
architect, mathematician and surveyor.
Female occupations:
nurse, housekeeper and midwife.

Reference
https://arxiv.org/abs/1711.08412

--- Wikimedia's Ores service ---

Software engineer Amir Sarabadani presented the ORES-project in Brussels in November 2017 during the Algoliterary Encounter.

This 'Objective Revision Evaluation Service' uses machine learning to help automate critical work on Wikimedia, like vandalism detection and the removal of articles. Cristina Cochior and Femke Snelting interviewed him.

Femke: To go back to your work. In these days you tried to understand what it means to find bias in machine learning and the proposal of Nicolas Maleve, who gave the workshop yesterday, was neither to try to fix it, nor to refuse to deal with systems that produce bias, but to work with them. He says that bias is inherent to human knowledge, so we need to find ways to somehow work with it. We're just struggling a bit with what would that mean, how would that work... So I was wondering whether you had any thoughts on the question of bias.

Amir: Bias inside Wikipedia is a tricky question because it happens on several levels. One level that has been discussed a lot is the bias in references. Not all references are accessible. So one thing that the Wikimedia Foundation has been trying to do, is to give free access to libraries that are behind a pay wall. They reduce the bias by only using open-access references. Another type of bias is the Internet connection, access to the Internet. There are lots of people who don't have it. One thing about China is that the Internet there is blocked. The content against the government of China inside Chinese Wikipedia is higher because the editors [who can access the website] are not people who are pro government, and try to make it more neutral. So, this happens in lots of places. But in the matter of artificial intelligence (AI) and the model that we use at Wikipedia, it's more a matter of transparency. There is a book about how bias in AI models can break people's lives, it's called 'Weapons of Math Destruction'. It talks about AI models that exist in the US that rank teachers and it's quite horrible because eventually there will be bias. The way to deal with it based on the book and their research was first that the model should be open source, people should be able to see what features are used and the data should be open also, so that people can investigate, find bias, give feedback and report back. There should be a way to fix the system. I think not all companies are moving in that direction, but Wikipedia, because of the values that they hold, are at least more transparent and they push other people to do the same thing.

Reference
https://gitlab.constantvzw.org/algolit/algolit/blob/master/algoliterary_encounter/Interview%20with%20Amir/AS.aac

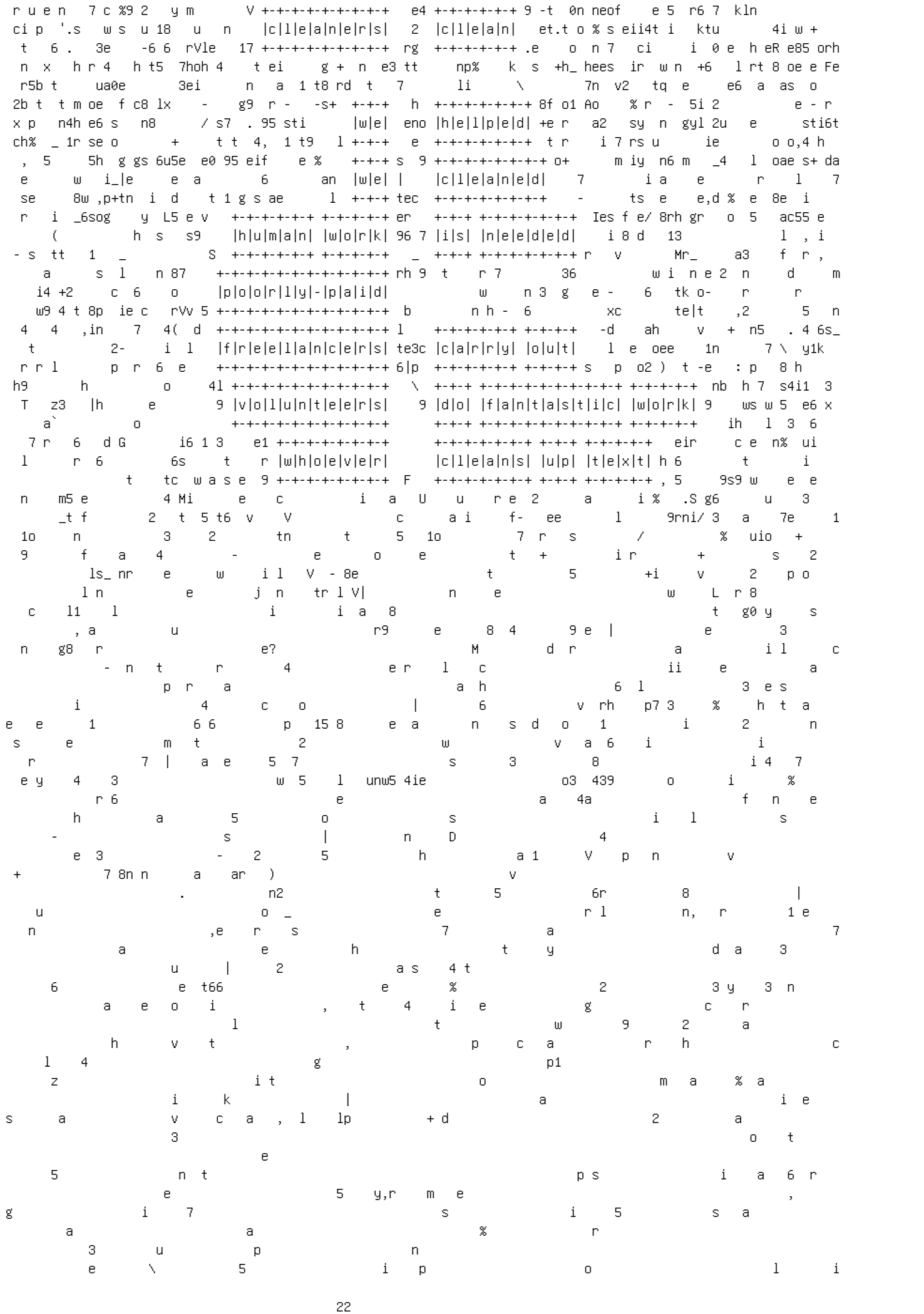
--- Tay ---

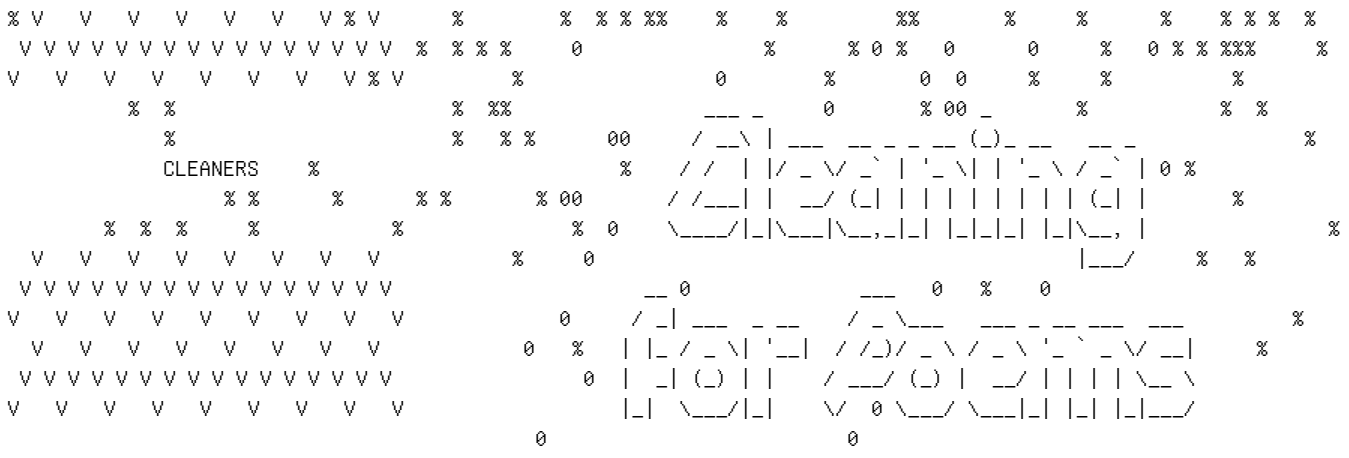
One of the infamous stories is that of the machine learning programme Tay, designed by Microsoft. Tay was a chat bot that imitated a teenage girl on Twitter. She lived for less than 24 hours before she was shut down. Few people know that before this incident, Microsoft had already trained and released XiaoIce on WeChat, China's most used chat application. XiaoIce's success was so promising that it led to the development of its American version. However, the developers of Tay were not prepared for the platform climate of Twitter. Although the bot knew how to distinguish a noun from an adjective, it had no understanding of the actual meaning of words. The bot quickly learned to copy racial insults and other discriminative language it learned from Twitter users and troll attacks.

Tay's appearance and disappearance was an important moment of consciousness. It showed the possible corrupt consequences that machine learning can have when the cultural context in which the algorithm has to live is not taken into account.

Reference
https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f

cleaners clean cleaners clean   cleaners clean   cleaners clean   cleaners clean   cleaners clean
 cleaners clean   cleaners clean   cleaners clean   cleaners clean   cleaners clean
          cleaners clean   cleaners clean   cleaners clean   cleaners clean
              cleaners clean   cleaners clean   cleaners clean
 cleaners clean   cleaners clean   cleaners clean   cle
ners clean   cleaners clean   cleaners clean
cleaners clean   cleaners clean   cleaners clean
            cleaners clean   cleaners clean   cleaners
lean   cleaners clean   cleaners clean
                 cleaners clean   cleaners clean
     cleaners clean   cleaners clean   cle
ners clean   cleaners clean   cleaners
clean   cleaners clean   cleaners
lean   cleaners clean   cleane
s clean   cleaners clean
cleaners clean   cleaners clean
          cleaners clean   cleaners clean
                        cleaners clean   cleaners clean
                                             cleaners clean
 cleaners clean   cleaners clean
                        cleaners clean   cleaners clean
                                                      cleaners clean
               cleaners clean   cleaners clean
                                                  cleaners clean
              cleaners clean   cleaners clean
                                                       cleaners clean
                      cleaners clean   cleaners
clean   cleaners clean
                                        cleaners clean
          cleaners clean   cleaners clean
                                                            cleaners clean
                                        cleaners clean
              cleaners clean   cleaners
clean   cleaners clean
                                                      cleaners clean
                                   cleaners clean
              cleaners clean   cle
ners clean   cleaners clean
                                                                 cleaners clean
                                                     cleaners clean
                                        cleaners clean
                        cleaners clean
            cleaners clean
 cleaners clean   cleaners
lean   cleaners clean
                                                                         cleaners clean
                                                                  cleaners clean
                                                            cleaners clean
                                                      cleaners clean
                                                 cleaners clean
                                             cleaners clean
                                          cleaners clean
                                        cleaners clean
                                       cleaners clean
                                       cleaners clean
                                        cleaners clean
                                          cleaners clean
                                             cleaners clean
                                                cleaners clean
                                                  cleaners clean
                                                   cleaners clean
                                                   cleaners clean
                                                  cleaners clean
                                                cleaners clean
                                             cleaners clean
                                         cleaners clean
                                    cleaners clean
                              cleaners clean
                       cleaners clean





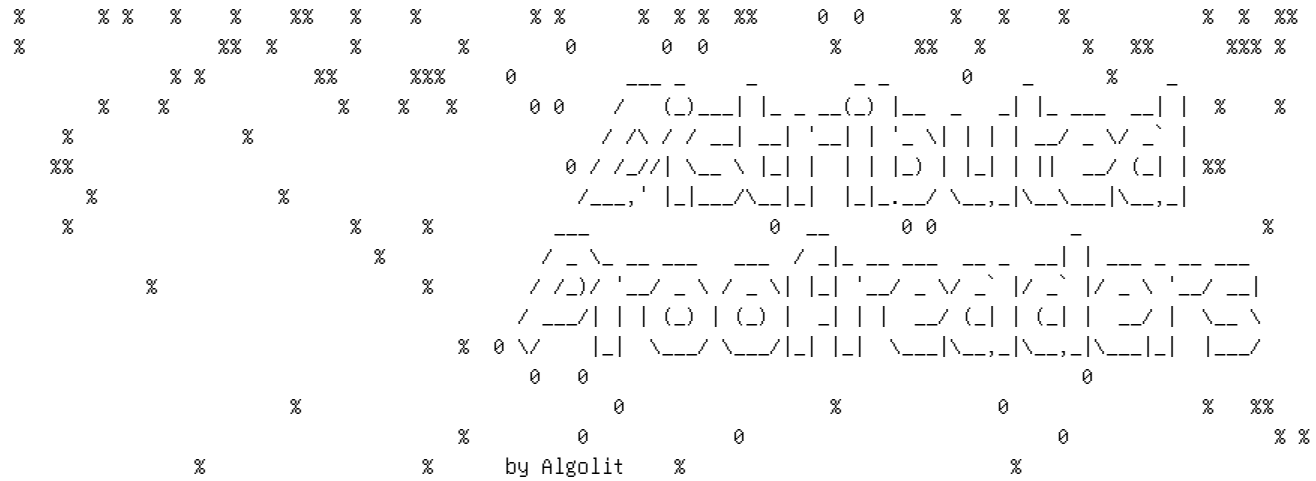
Algolit chooses to work with texts that are free of copyright. This means that they have been published under a Creative Commons 4.0 license – which is rare - or that they are in the public domain because the author died more than 70 years ago. This is the case for the publications of the Mundaneum. We received 203 documents that we helped turn into datasets. They are now available for others online. Sometimes we had to deal with poor text formats, and we often dedicated a lot of time to cleaning up documents. We were not alone in doing this.

Books are scanned at high resolution, page by page. This is time-consuming, laborious human work and often the reason why archives and libraries transfer their collections and leave the job to companies like Google. The photos are converted into text via OCR (Optical Character Recognition), a software that recognizes letters, but often makes mistakes, especially when it has to deal with ancient fonts and wrinkled pages. Yet more wearisome human work is needed to improve the texts. This is often carried out by poorly-paid freelancers via micro-payment platforms like Amazon's Mechanical Turk; or by volunteers, like the community around the Distributed Proofreaders Project, which does fantastic work. Whoever does it, or wherever it is done, cleaning up texts is a towering job for which no structural automation yet exists.

For this exhibition we worked with 3 per cent of the Mundaneum's archive. These documents were first scanned or photographed. To make the documents searchable they were transformed into text using Optical Character Recognition software (OCR). OCR are algorithmic models that are trained on other texts. They have learned to identify characters, words, sentences and paragraphs. The software often makes 'mistakes'. It might recognize a wrong character, it might get confused by a stain an unusual font or the reverse side of the page being visible.

While these mistakes are often considered noise, confusing the training, they can also be seen as poetic interpretations of the algorithm. They show us the limits of the machine. And they also reveal how the algorithm might work, what material it has seen in training and what is new. They say something about the standards of its makers. In this installation we ask your help in verifying our dataset. As a reward we'll present you with a personal algorithmic improvisation.

Concept, code, interface: Gijs de Heij



Distributed Proofreaders is a web-based interface and an international community of volunteers who help converting public domain books into e-books. For this exhibition they proofread the Mundaneum publications that appeared before 1923 and are in the public domain in the US. Their collaboration meant a great relief for the members of Algolit. Less documents to clean up!

All the proofread books have been made available on the Project Gutenberg archive.

For this exhibition, An Mertens interviewed Linda Hamilton, the general manager of Distributed Proofreaders.

Interview: An Mertens

Editing: Michael Murtaugh, Constant

--- Project Gutenberg and
Distributed Proofreaders ---

Project Gutenberg is our Ali Baba cave. It offers more than 58,000 free eBooks to be downloaded or read online. Works are accepted on Gutenberg when their U.S. copyright has expired. Thousands of volunteers digitize and proofread books to help the project. An essential part of the work is done through the Distributed Proofreaders project. This is a web-based interface to help convert public domain books into e-books. Think of text files, EPUBs, Kindle formats. By dividing the workload into individual pages, many volunteers can work on a book at the same time; this speeds up the cleaning process.

During proofreading, volunteers are presented with a scanned image of the page and a version of the text, as it is read by an OCR algorithm trained to recognize letters in images. This allows the text to be easily compared to the image, proofread, and sent back to the site. A second volunteer is then presented with the first volunteer's work. She verifies and corrects the work as necessary, and submits it back to the site. The book then similarly goes through a third proofreading round, plus two more formatting rounds using the same web interface. Once all the pages have completed these steps, a post-processor carefully assembles them into an e-book and submits it to the Project Gutenberg archive.

We collaborated with the Distributed Proofreaders project to clean up the digitized files we received from the Mundaneum collection. From November 2018 until the first upload of the cleaned-up book 'L'Afrique aux Noirs' in February 2019, An Mertens exchanged about 50 emails with Linda Hamilton, Sharon Joiner and Susan Hanlon, all volunteers from the Distributed Proofreaders project. The conversation is published online. It might inspire you to share unavailable books online.

--- An algoliterary version
of the Maintenance Manifesto ---

In 1969, one year after the birth of her first child, the New York artist Mierle Laderman Ukeles wrote a Manifesto for Maintenance Art. The manifesto calls for a readdressing of the status of maintenance work both in the private, domestic space, and in public. What follows is an altered version of her text inspired by the work of the Cleaners.

IDEAS

A. The Death Instinct and the Life Instinct:

The Death Instinct: separation; categorization; avant-garde par excellence; to follow the predicted path to death - run your own code; dynamic change. operate it. For nearly 84 years, the Turk won most The Life Instinct: unification; the eternal return; the perpetuation and MAINTENANCE of the material; survival systems and operations; equilibrium.

B. Two basic systems: Development and Maintenance.

The sourball of every revolution: after the revolution, who's going to try to spot the bias in the output?

Development: pure individual creation; the new; change; progress; advance; excitement; flight or fleeing.

Maintenance: keep the dust off the pure individual creation; preserve the new; sustain the change; protect progress; defend and prolong the advance; renew the excitement; repeat the flight; show your work - show it again, keep the git repository groovy, keep the data analysis revealing.

Development systems are partial feedback systems with major room for change.

Maintenance systems are direct feedback systems with little room for alteration.

C. Maintenance is a drag;
it takes all the fucking time (lit.)

The mind boggles and chafes at the boredom.

The culture assigns lousy status on maintenance jobs = minimum wages, Amazon Mechanical Turks = virtually no pay.

Clean the set, tag the training data, correct the typos, modify the parameters, finish the report, keep the requester happy, upload the new version, attach words that were wrongly separated by OCR back together, complete those Human Intelligence Tasks, try to guess the meaning of the requester's formatting, you must accept the HIT before you can submit the results, summarize the image, add the bounding box, what's the semantic similarity of this text, check the translation quality, collect your micro-payments, become a hit Mechanical Turk.

Reference

https://www.arnolfini.org.uk/blog/manifesto-for-maintenance-art-1969

--- A bot panic on Amazon Mechanical Turk ---

Amazon's Mechanical Turk takes the name of a chess-playing automaton from the eighteenth century. In fact, the Turk wasn't a machine at all. It was a mechanical illusion that allowed a human chess master to hide inside the box and manually

of the games played during its demonstrations around Europe and the Americas. Napoleon Bonaparte is said to have been fooled by this trick too.

The Amazon Mechanical Turk is an online platform for humans to execute tasks that algorithms cannot. Examples include annotating sentences as being positive or negative, spotting number plates, discriminating between face and non-face. The jobs posted on this platform are often paid less than a cent per task. Tasks that are more complex or require more knowledge can be paid up to several cents. To earn a living, Turkers need to finish as many tasks as fast as possible, leading to inevitable mistakes. As a result, the requesters have to incorporate quality checks when they post a job on the platform. They need to test whether the Turker actually has the ability to complete the task, and they also need to verify the results. Many academic researchers use Mechanical Turk as an alternative to have their students execute these tasks.

In August 2018 Max Hui Bai, a psychology student from the University of Minnesota, discovered that the surveys he conducted with Mechanical Turk were full of nonsense answers to open-ended questions. He traced back the wrong answers and found out that they had been submitted by respondents with duplicate GPS locations. This raised suspicion. Though Amazon explicitly prohibits robots from completing jobs on Mechanical Turk, the company does not deal with the problems they cause on their platform. Forums for Turkers are full of conversations about the automation of the work, sharing practices of how to create robots that can even violate Amazon's terms. You can also find videos on YouTube that show Turkers how to write a bot to fill in answers for you.

Kristy Milland, an Mechanical Turk activist, says: 'Mechanical Turk workers have been treated really, really badly for 12 years, and so in some ways I see this as a point of resistance. If we were paid fairly on the platform, nobody would be risking their account this way.'

Bai is now leading a research project among social scientists to figure out how much bad data is in use, how large the problem is, and how to stop it. But it is impossible at the moment to estimate how many datasets have become unreliable in this way.

References

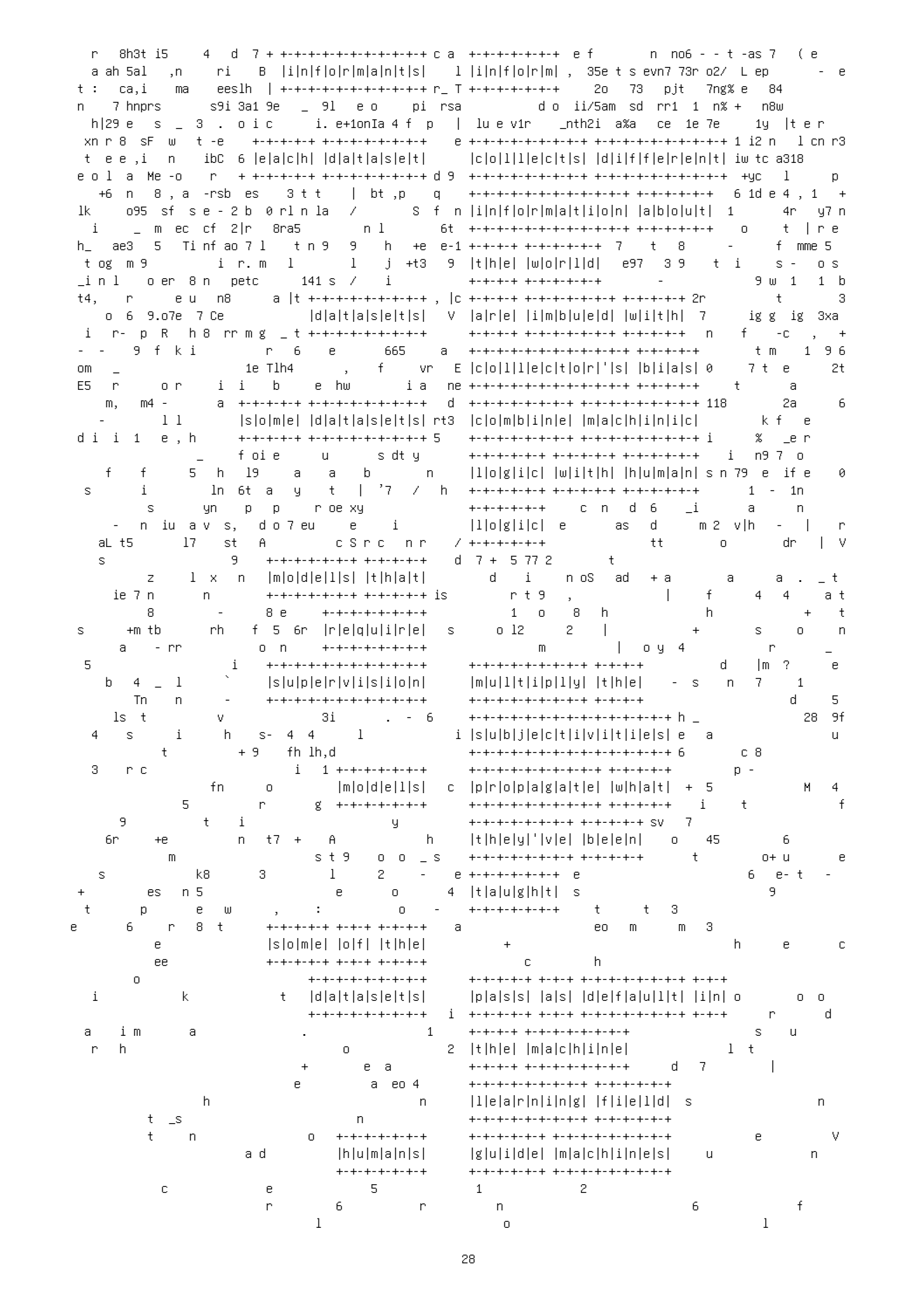
https://requester.mturk.com/create/projects/new

https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/

informants inform informants inform informants inform informants inform informants inform info
mants inform informants inform informants inform informants inform informants i
form informants inform informants inform informants inform info
mants inform informants inform informants inform informants info
m informants inform informants inform informants inform
informants inform informants inform informants
inform informants inform informants inform
informants inform informants inform informants info
m informants inform informants inform
informants inform informants inform
informants inform informants inform in
ormants inform informants inform infor
ants inform informants inform info
mants inform informants inform
informants inform informants inform
informants inform informants inform
informants inform informants inform
informants inform infor
ants inform informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform
informants inform informants info
m informants inform
informants inform
informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform in
ormants inform info
mants inform infor
ants inform infor
ants inform info
mants inform in
ormants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform



INFORMANTS

Machine learning algorithms need guidance, whether they are supervised or not. In order to separate one thing from another, they need material to extract patterns from. One should carefully choose the study material, and adapt it to the machine's task. It doesn't make sense to train a machine with nineteenth-century novels if its mission is to analyse tweets. A badly written textbook can lead a student to give up on the subject altogether. A good textbook is preferably not a textbook at all.

This is where the dataset comes in: arranged as neatly as possible, organized in disciplined rows and lined-up columns, waiting to be read by the machine. Each dataset collects different information about the world, and like all collections, they are imbued with collectors' bias. You will hear this expression very often: 'data is the new oil'. If only data were more like oil! Leaking, dripping and heavy with fat, bubbling up and jumping unexpectedly when in contact with new matter. Instead, data is supposed to be clean. With each process, each questionnaire, each column title, it becomes cleaner and cleaner, chipping distinct characteristics until it fits the mould of the dataset.

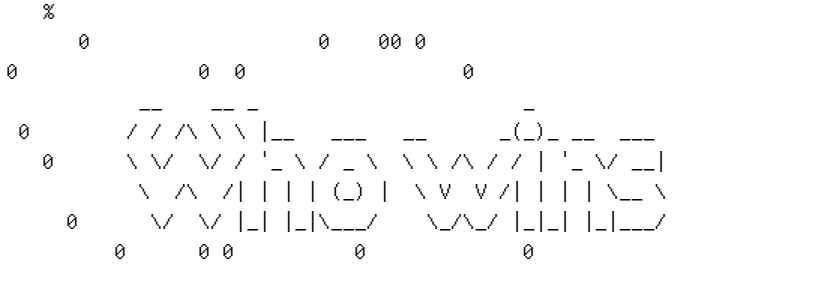
Some datasets combine the machinic logic with the human logic. The models that require supervision multiply the subjectivities of both data collectors and annotators, then propagate what they've been taught. You will encounter some of the datasets that pass as default in the machine learning field, as well as other stories of humans guiding machines.

by Algolit

We often start the monthly Algolit meetings by searching for datasets or trying to create them. Sometimes we use already-existing corpora, made available through the Natural Language Toolkit nltk. NLTK contains, among others, The Universal Declaration of Human Rights, inaugural speeches from US presidents, or movie reviews from the popular site Internet Movie Database (IMDb). Each style of writing will conjure different relations between the words and will reflect the moment in time from which they originate. The material included in NLTK was selected because it was judged useful for at least one community of researchers. In spite of specificities related to the initial context of each document, they become universal documents by default, via their inclusion into a collection of publicly available corpora. In this sense, the Python package manager for natural language processing could be regarded as a time capsule. The main reason why The Universal Declaration for Human Rights was included may have been because of the multiplicity of translations, but it also paints a picture of the types of human writing that algorithms train on.

With this work, we look at the datasets most commonly used by data scientists to train machine algorithms. What material do they consist of? Who collected them? When?

Concept & execution: Cristina Cochior



Who wins: creation of relationships

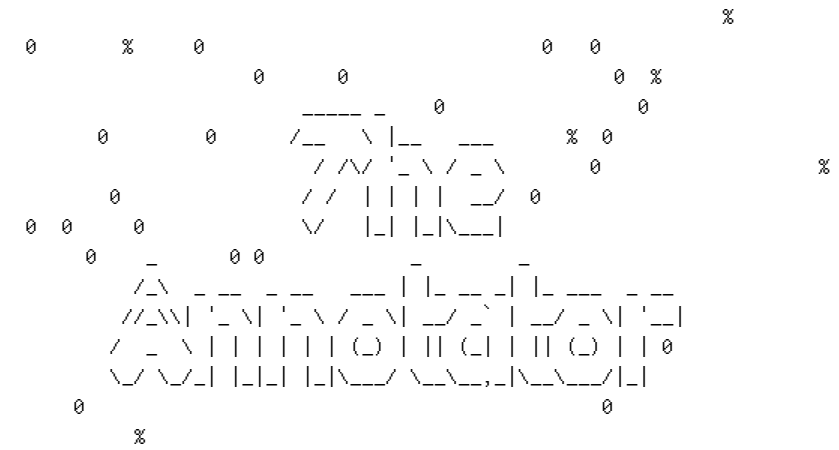
by Louise Dekeuleneer, student Arts²/Section Visual Communication

French is a gendered language. Indeed many words are female or male and few are neutral. The aim of this project is to show that a patriarchal society also influences the language itself.

The work focused on showing whether more female or male words are used on highlighting the influence of context on the gender of words. At this stage, no conclusions have yet been drawn.

Law texts from 1900 to 1910 made available by the Mundaneum have been passed into an algorithm that turns the text into a list of words. These words are then compared with another list of French words, in which is specified whether the word is male or female. This list of words comes from Google Books. They created a huge database in 2012 from all the books scanned and available on Google Books.

Male words are highlighted in one colour and female words in another. Words that are not gendered (adverbs, verbs, etc.) are not highlighted. All this is saved as an HTML file so that it can be directly opened in a web page and printed without the need for additional layout. This is how each text becomes a small booklet by just changing the input text of the algorithm.



by Algolit

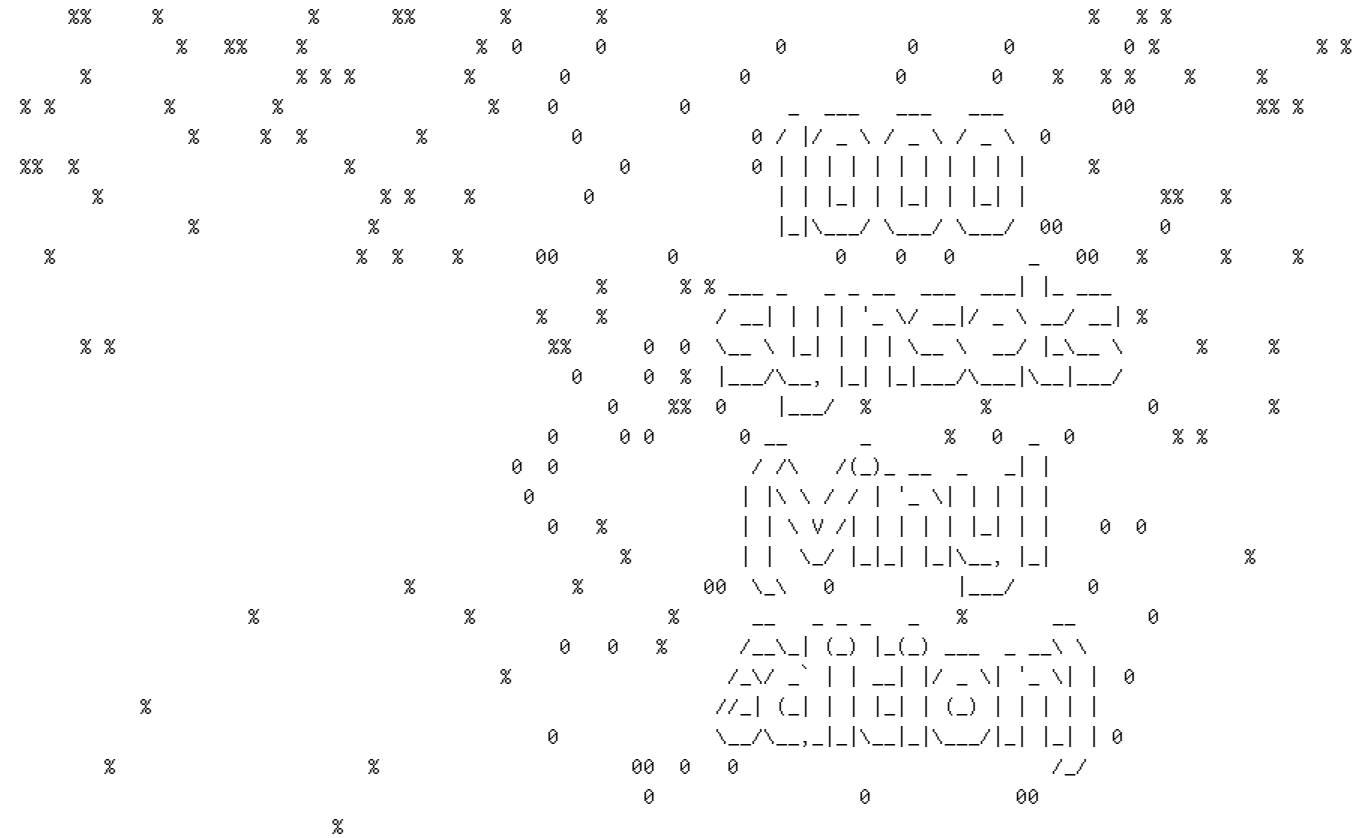
The annotator asks for the guidance of visitors in annotating the archive of Mundaneum.

The annotation process is a crucial step in supervised machine learning where the algorithm is given examples of what it needs to learn. A spam filter in training will be fed examples of spam and real messages. These examples are entries, or rows from the dataset with a label, spam or non-spam.

The labelling of a dataset is work executed by humans, they pick a label for each row of the dataset. To ensure the quality of the labels multiple annotators see the same row and have to give the same label before an example is included in the training data. Only when enough samples of each label have been gathered in the dataset can the computer start the learning process.

In this interface we ask you to help us classify the cleaned texts from the Mundaneum archive to expand our training set and improve the quality of the installation 'Classifying the World' in Oracles.

Concept, code, interface: Gijs de Heij



by Algolit

Created in 1985, Wordnet is a hierarchical taxonomy that describes the world. It was inspired by theories of human semantic memory developed in the late 1960s. Nouns, verbs, adjectives and adverbs are grouped into synonyms sets or synsets, expressing a different concept.

ImageNet is an image dataset based on the WordNet 3.0 nouns hierarchy. Each synset is depicted by thousands of images. From 2010 until 2017, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was a key benchmark in object category classification for pictures, having a major impact on software for photography, image searches, image recognition.

1000 synsets (Vinyl Edition) contains the 1000 synsets used in this challenge recorded in the highest sound quality that this analog format allows. This work highlights the importance of the datasets used to train artificial intelligence (AI) models that run on devices we use on a daily basis. Some of them inherit classifications that were conceived more than 30 years ago. This sound work is an invitation to thoughtfully analyse them.

Concept & recording: Javier Lloret

Voices: Sara Hamadeh & Joseph Hughes

--- Datasets as representations ---

The data-collection processes that lead to the creation of the dataset raise important questions: who is the author of the data? Who has the privilege to collect? For what reason was the selection made? What is missing?

The artist Mimi Onuoha gives a brilliant example of the importance of collection strategies. She chose the case of statistics related to hate crimes. In 2012, the FBI Uniform Crime Reporting (UCR) Program registered almost 6000 hate crimes committed. However, the Department of Justice's Bureau of Statistics came up with about 300.000 reports of such cases. That is over 50 times as many. The difference in numbers can be explained by how the data was collected. In the first situation law enforcement agencies across the country voluntarily reported cases. For the second survey, the Bureau of Statistics distributed the National Crime Victimization form directly to the homes of victims of hate crimes.

In the field of Natural Language Processing (NLP) the material that machine learners work with is text-based, but the same questions still apply: who are the authors of the texts that make up the dataset? During what period were the texts collected? What type of worldview do they represent?

In 2017, Google's Top Stories algorithm pushed a thread of 4chan, a non-moderated content website, to the top of the results page when searching for the Las Vegas shooter. The name and portrait of an innocent person were linked to the terrible crime. Google changed its algorithm just a few hours after the mistake was discovered, but the error had already affected the person. The question is: why did Google not exclude 4chan content from the training dataset of the algorithm?

Reference
https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa

https://arstechnica.com/information-technology/2017/10/google-admits-citing-4chan-to-spread-fake-vegas-shooter-news/

--- Labeling for an Oracle that detects vandalism on Wikipedia ---

This fragment is taken from an interview with Amir Sarabadani, software engineer at Wikimedia. He was in Brussels in November 2017 during the Algoliterary Encounter.

Femke: If you think about Wikipedia as a living community, with every edit the project changes. Every edit is somehow a contribution to a living organism of knowledge. So, if from within that community you try to distinguish what serves the community and what doesn't and you try to generalize that, because I think that's what the good faith-bad faith algorithm is trying to do, to find helper tools to support the project, you do that on the basis of a generalization that is on the abstract idea of what Wikipedia is and not on the living organism of what happens every day. What interests me in the relation between vandalism and debate is how we can understand the conventional drive that sits in these machine-learning processes that we seem to come across in many places. And how can we somehow understand them and deal with them? If you place your separation of good faith-bad faith on pre-existing labelling and then reproduce that in your understanding of what edits are being made, how then to take into account movements that are happening, the life of the actual project?

Amir: It's an interesting discussion. Firstly, what we are calling good faith and bad faith comes from the community itself. We are not doing labelling for them, they are doing labelling for themselves. So, in many different language Wikipedias, the definition of what is good faith and what is bad faith will differ. Wikimedia is trying to reflect what is inside the organism and not to change the organism itself. If the organism changes, and we see that the definition of good faith and helping Wikipedia has been changed, we are implementing this feedback loop that lets people from inside their community pass judgement on their edits and if they disagree with the labelling, we can go back to the model and retrain the algorithm to reflect this change. It's some sort of closed loop: you change things and if someone sees there is a problem, then they tell us and we can change the algorithm back. It's an ongoing project.

Reference
https://gitlab.constantvzw.org/algolit/algolit/blob/master/algoliterary_encounter/Interview%20with%20Amir

--- How to make your dataset known ---

NLTK stands for Natural Language Toolkit. For programmers who process natural language using Python, this is an essential library to work with. Many tutorial writers recommend machine learning learners to start with the inbuilt NLTK datasets. It comprises 71 different collections, with a total of almost 6000 items.

There is for example the Movie Review corpus for sentiment analysis. Or the Brown corpus, which was put together in the 1960s by Henry Kučera and W. Nelson Francis at Brown University in Rhode Island. There is also the Declaration of Human Rights corpus, which is commonly used to test whether the code can run on multiple languages.

The corpus contains the Declaration of Human Rights expressed in 372 languages from around the world.

But what is the process of getting a dataset accepted into the NLTK library nowadays? On the Github page, the NLTK team describes the following requirements:

- Only contribute corpora that have obtained a basic level of notability. That means, there is a publication that describes it, and a community of programmers who are using it.
- Ensure that you have permission to redistribute the data, and can document this. This means that the dataset is best published on an external website with a licence.
- Use existing NLTK corpus readers where possible, or else contribute a well-documented corpus reader to NLTK. This means, you need to organize your data in such a way that it can be easily read using NLTK code.

--- Extract from a positive IMDb movie review from the NLTK dataset ---

corpus: NLTK, movie reviews

fileid: pos/cv998_14111.txt

steven spielberg ' s second epic film on world war ii is an unquestioned masterpiece of film . spielberg , ever the student on film , has managed to resurrect the war genre by producing one of its grittiest , and most powerful entries . he also managed to cast this era ' s greatest answer to jimmy stewart , tom hanks , who delivers a performance that is nothing short of an astonishing miracle . for about 160 out of its 170 minutes , " saving private ryan " is flawless . literally . the plot is simple enough . after the epic d - day invasion (whose sequences are nothing short of spectacular) , capt . john miller (hanks) and his team are forced to search for a pvt . james ryan (damon) , whose brothers have all died in battle . once they find him , they are to bring him back for immediate discharge so that he can go home . accompanying miller are his crew , played with astonishing perfection by a group of character actors that are simply sensational . barry pepper , adam goldberg , vin diesel , giovanni ribisi , davies , and burns are the team sent to find one man , and bring him home . the battle sequences that bookend the film are extraordinary . literally .

--- The ouroboros of machine learning ---

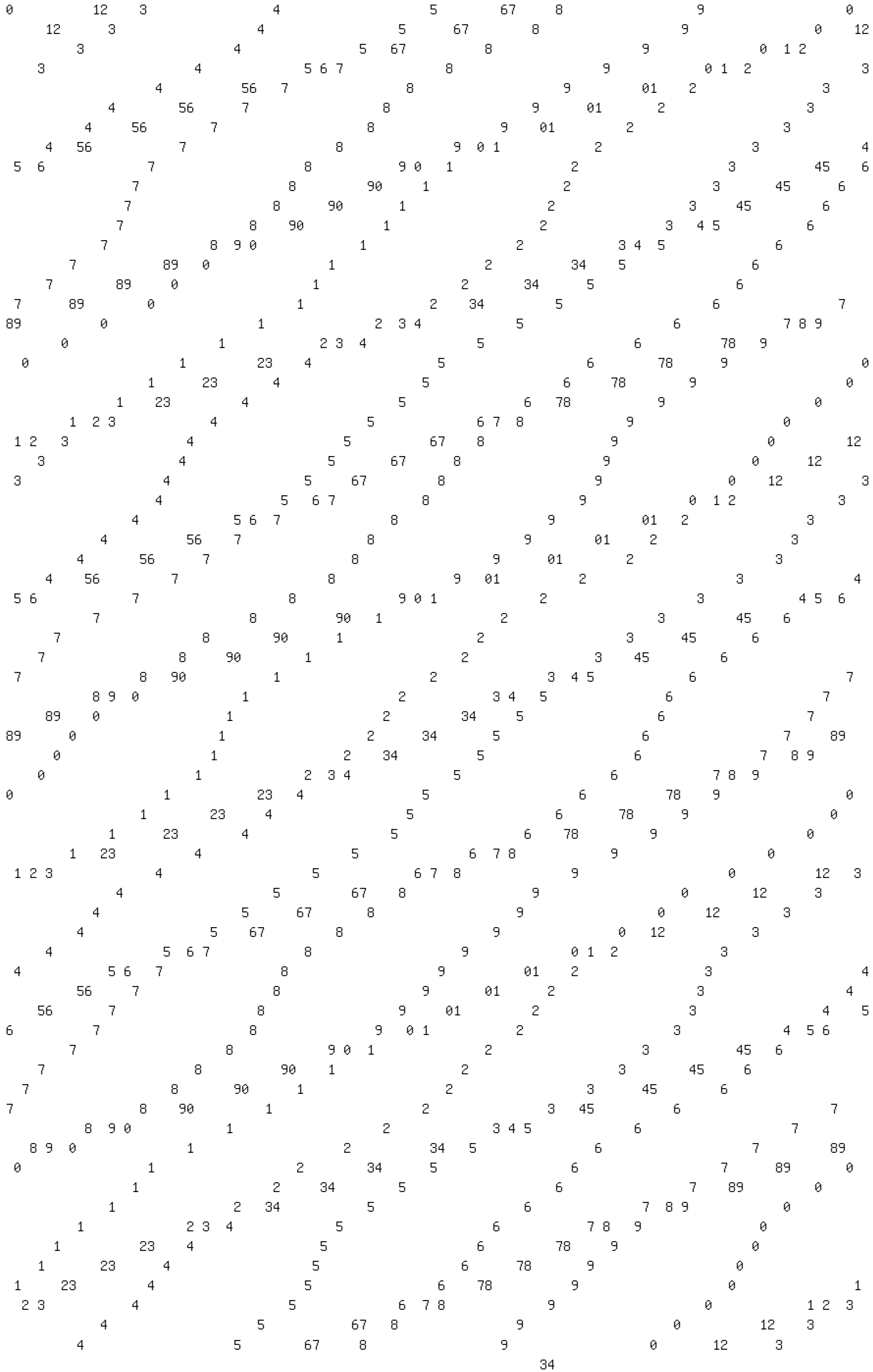
Wikipedia has become a source for learning not only for humans, but also for machines. Its articles are prime sources for training models. But very often, the material the machines are trained on is the same content that they helped to write.

In fact, at the beginning of Wikipedia, many articles were written by bots. Rambot, for example, was a controversial bot figure on the English-speaking platform. It authored 98 per cent of the pages describing US towns.

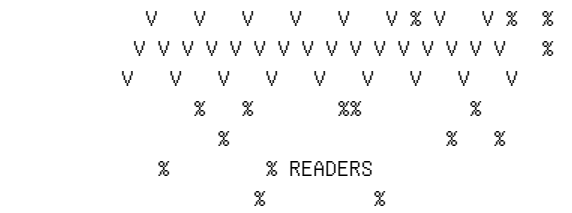
As a result of serial and topical robot interventions, the models that are trained on the full Wikipedia dump have a unique view on composing articles. For example, a topic model trained on all of Wikipedia articles will associate 'river' with 'Romania' and 'village' with 'Turkey'. This is because there are over 10000 pages written about villages in Turkey. This should be enough to spark anyone's desire for a visit, but it is far too much compared to the number of articles other countries have on the subject. The asymmetry causes a false correlation and needs to be redressed. Most models try to exclude the work of these prolific robot writers.

Reference

https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/



readers read readers read readers read readers read readers read readers read readers re
d readers read readers read readers read readers read readers re
d readers read readers read readers read readers read
readers read readers read readers read re
ders read readers read readers read readers re
d readers read readers read readers r
ad readers read readers read
readers read readers read readers read
readers read readers read
readers read readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers
read readers read
readers read readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers re
d readers read
readers read
readers read readers read
readers read
readers read
readers read re
ders read readers read
readers read
readers read
readers read
readers read readers r
ad readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read readers
read readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read r



We communicate with computers through language. We click on icons that have a description in words, we tap words on keyboards, use our voice to give them instructions. Sometimes we trust our computer with our most intimate thoughts and forget that they are extensive calculators. A computer understands every word as a combination of zeros and ones. A letter is read as a specific ASCII number: capital 'A' is 001.

In all models, rule-based, classical machine learning, and neural networks, words undergo some type of translation into numbers in order to understand the semantic meaning of language. This is done through counting. Some models count the frequency of single words, some might count the frequency of combinations of words, some count the frequency of nouns, adjectives, verbs or noun and verb phrases. Some just replace the words in a text by their index numbers. Numbers optimize the operative speed of computer processes, leading to fast predictions, but they also remove the symbolic links that words might have. Here we present a few techniques that are dedicated to making text readable to a machine.

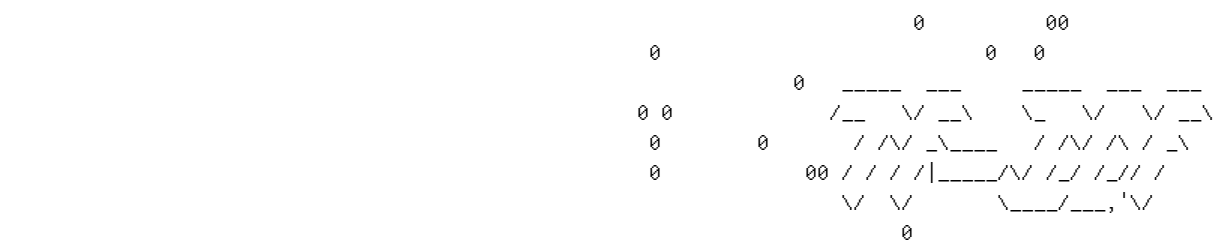
by Algolit

The bag-of-words model is a simplifying representation of text used in Natural Language Processing (NLP). In this model, a text is represented as a collection of its unique words, disregarding grammar, punctuation and even word order. The model transforms the text into a list of words and how many times they're used in the text, or quite literally a bag of words.

This heavy reduction of language was the big shock when beginning to machine learn. Bag of words is often used as a baseline, on which the new model has to perform better. It can understand the subject of a text by recognizing the most frequent or important words. It is often used to measure the similarities of texts by comparing their bags of words.

For this work the article 'Le Livre de Demain' by engineer G. Vander Haeghen, published in 1907 in the Bulletin de l'Institut International de Bibliographie of the Mundaneum, has been literally reduced to a bag of words. You can buy a bag at the reception of Mundaneum.

Concept & realisation: An Mertens



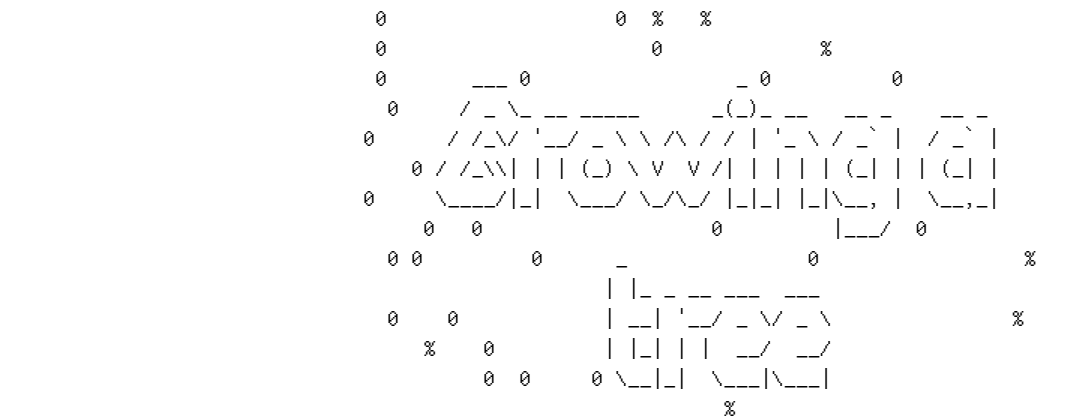
by Algolit

The TF-IDF (Term Frequency-Inverse Document Frequency) is a weighting method used in text search. This statistical measure makes it possible to evaluate the importance of a term contained in a document, relative to a collection or corpus of documents. The weight increases in proportion to the number of occurrences

of the word in the document. It also varies according to the frequency of the word in the corpus. The TF-IDF is used in particular in the classification of spam in email softwares.

A web-based interface shows this algorithm through animations making it possible to understand the different steps of text classification. How does a TF-IDF-based programme read a text? How does it transform words into numbers?

Concept, code, animation: Sarah Garcin



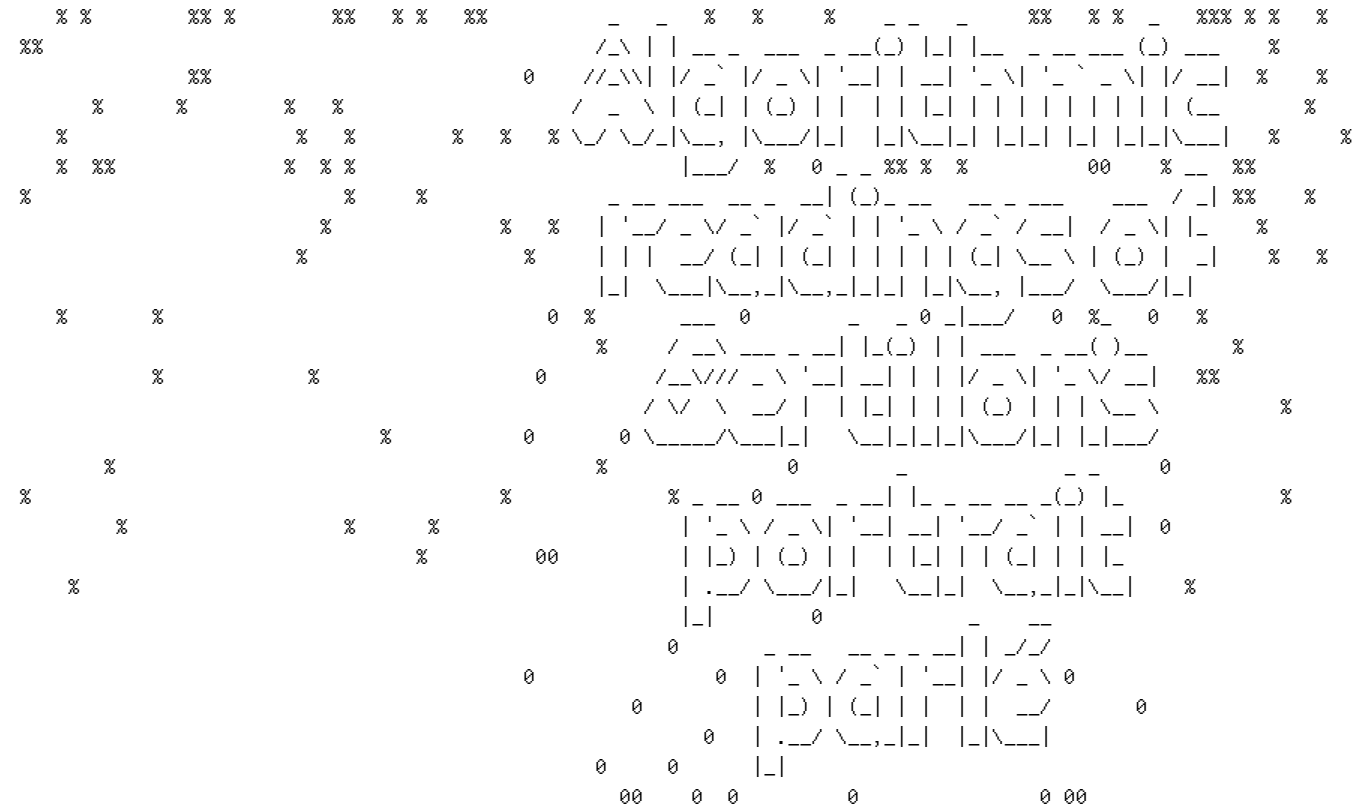
by Algolit

Parts-of-Speech is a category of words that we learn at school: noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article, or determiner.

In Natural Language Processing (NLP) there exist many writings that allow sentences to be parsed. This means that the algorithm can determine the part-of-speech of each word in a sentence. 'Growing a tree' uses this technique to define all nouns in a specific sentence. Each noun is then replaced by its definition. This allows the sentence to grow autonomously and infinitely. The recipe of 'Growing a tree' was inspired by Oulipo's constraint of 'littérature définitionnelle' invented by Marcel Benabou in 1966. In a given phrase, one replaces every significant element (noun, adjective, verb, adverb) by one of its definitions in a given dictionary; one reiterates the operation on the newly received phrase, and again.

The dictionary of definitions used in this work is Wordnet. Wordnet is a combination of a dictionary and a thesaurus that can be read by machines. According to Wikipedia it was created in the Cognitive Science Laboratory of Princeton University starting in 1985. The project was initially funded by the US Office of Naval Research and later also by other US government agencies including DARPA, the National Science Foundation, the Disruptive Technology Office (formerly the Advanced Research and Development Activity), and REFLEX.

Concept, code & interface: An Mertens & Gijs de Heij



% by Guillaume Slizewicz (Urban Species)

Written in 1907, Un code télégraphique du portrait parlé is an attempt to translate the 'spoken portrait', a face-description technique created by a policeman in Paris, into numbers. By implementing this code, it was hoped that faces of criminals and fugitives could easily be communicated over the telegraphic network in between countries. In its form, content and ambition this text represents our complicated relationship with documentation technologies. This text sparked the creation of the following installations for three reasons:

- First, the text is an algorithm in itself, a compression algorithm, or to be more precise, the presentation of a compression algorithm. It tries to reduce the information to smaller pieces while keeping it legible for the person who has the code. In this regard it is linked to the way we create technology, our pursuit for more efficiency, quicker results, cheaper methods. It represents our appetite for putting numbers on the entire world, measuring the smallest things, labeling the tiniest differences. This text itself embodies the vision of the Mundaneum.

- Second it is about the reasons for and the applications of technology. It is almost ironic that this text was in the selected archives presented to us in a time when face recognition and data surveillance are so much in the news. This text bears the same characteristics as some of today's technology: motivated by social control, classifying people, laying the basis for a surveillance society. Facial features are at the heart of recent controversies: mugshots were standardized by Bertillon, now they are used to train neural network to predict criminals from law-abiding citizens. Facial recognition systems allow the arrest of criminals via CCTV infrastructure and some assert that people’s features can predict sexual orientation.

- The last point is about how it represents the evolution of mankind’s techno-structure. What our tools allow us to do, what they forbid, what they hinder, what they make us remember and what they make us forget. This document enables a classification between people and a certain vision of what normality is. It

breaks the continuum into pieces thus allowing stigmatization/discrimination. On the other hand this document also feels obsolete today, because our techno-structure does not need such detailed written descriptions about fugitives, criminals or citizens. We can now find fingerprints, iris scans or DNA info in large datasets and compare them directly. Sometimes the technological systems do not even need human supervision and recognize directly the identity of a person via their facial features or their gait. Computers do not use intricate written language to describe a face, but arrays of integers. Hence all the words used in this documents seem désuets, dated. Have we forgotten what some of them mean? Did photography make us forget how to describe faces? Will voice-assistance software teach us again?

Writing with Otlet

Writing with Otlet is a character generator that uses the spoken portrait code as its database. Random numbers are generated and translated into a set of features. By creating unique instances, the algorithm reveals the richness of the description that is possible with the portrait code while at the same time embodying its nuances.

An interpretation of Bertillon's spoken portrait.

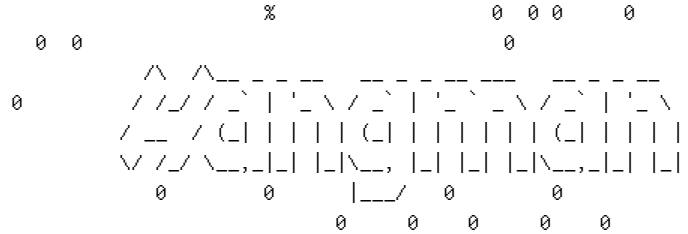
This work draws a parallel between Bertillon systems and current ones. A webcam linked to a facial recognition algorithm captures the beholder's face and translates it into numbers on a canvas, printing it alongside Bertillon's labelled faces.

References

https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/

https://fr.wikipedia.org/wiki/Bertillonnage

https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html



by Laetitia Trozzi, student Arts²/Section Digital Arts

What better way to discover Paul Otlet and his passion for literature than to play hangman? Through this simple game, which consists in guessing the missing letters in a word, the goal is to make the public discover terms and facts related to one of the creators of the Mundaneum.

Hangman uses an algorithm to detect the frequency of words in a text. Next, a series of significant words were isolated in Paul Otlet's bibliography. This series of words is integrated into a hangman game presented in a terminal. The difficulty of the game gradually increases as the player is offered longer and longer words. Over the different game levels, information about the life and work of Paul Otlet is displayed.

Naive Bayes, Support Vector Machines and Linear Regression are called classical machine learning algorithms. They perform well when learning with small datasets. But they often require complex Readers. The task the Readers do, is also called feature-engineering. This means that a human needs to spend time on a deep exploratory data analysis of the dataset.

Features can be the frequency of words or letters, but also syntactical elements like nouns, adjectives, or verbs. The most significant features for the task to be solved, must be carefully selected and passed over to the classical machine learning algorithm. This process marks the difference with Neural Networks. When using a neural network, there is no need for feature-engineering. Humans can pass the data directly to the network and achieve fairly good performances straightaway. This saves a lot of time, energy and money.

The downside of collaborating with Neural Networks is that you need a lot more data to train your prediction model. Think of 1GB or more of plain text files. To give you a reference, 1 A4, a text file of 5000 characters only weighs 5 KB. You would need 8,589,934 pages. More data also requires more access to useful datasets and more, much more processing power.

--- Character n-gram for
authorship recognition ---

Imagine … You've been working for a company for more than ten years. You have been writing tons of emails, papers, internal notes and reports on very different topics and in very different genres. All your writings, as well as those of your colleagues, are safely backed-up on the servers of the company.

One day, you fall in love with a colleague. After some time you realize this human is rather mad and hysterical and also very dependent on you. The day you decide to break up, your (now) ex elaborates a plan to kill you. They succeed. This is unfortunate. A suicide letter in your name is left next to your corpse. Because of emotional problems, it says, you decided to end your life. Your best friends don't believe it. They decide to take the case to court. And there, based on the texts you and others produced over ten years, a machine learning model reveals that the suicide letter was written by someone else.

How does a machine analyse texts in order to identify you? The most robust feature for authorship recognition is delivered by the character n-gram technique. It is used in cases with a variety of thematics and genres of the writing. When using character n-grams, texts are considered as sequences of characters. Let's consider the character trigram. All the overlapping sequences of three characters are isolated. For example, the character 3-grams of 'Suicide', would be, ‘Sui’, ‘uic’, ‘ici’, ‘cid’, etc. Character n-gram features are very simple, they're language-independent and they're tolerant to noise. Furthermore, spelling mistakes do not jeopardize the technique.

Patterns found with character n-grams focus on stylistic choices that are unconsciously made by the author. The patterns remain stable over the full length of the text, which is important for authorship recognition. Other types of experiments could include measuring the length of words or sentences, the vocabulary richness, the frequencies of function words; even syntax or semantics-related measurements.

This means that not only your physical fingerprint is unique, but also the way you compose your thoughts! The same n-gram technique discovered that The Cuckoo’s Calling, a novel by Robert Galbraith, was actually written by … J. K. Rowling!

Reference

Paper: On the Robustness of Authorship Attribution Based on Character N-gram Features, Efstathios Stamatatos, in Journal of Law & Policy, Volume 21, Issue 2, 2013.

News article: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

--- A history of n-grams ---

The n-gram algorithm can be traced back to the work of Claude Shannon in information theory. In the paper, 'A Mathematical Theory of Communication', published in 1948, Shannon performed the first instance of an n-gram-based model for natural language. He posed the question: given a sequence of letters, what is the likelihood of the next letter?

If you read the following excerpt, can you tell who it was written by? Shakespeare or an n-gram piece of code?

SEBASTIAN: Do I stand till the break off.

BIRON: Hide thy head.

VENTIDIUS: He purposeth to Athens: whither, with the vow
I made to handle you.

FALSTAFF: My good knave.

You may have guessed, considering the topic of this story, that an n-gram algorithm generated this text. The model is trained on the compiled

works of Shakespeare. While more recent algorithms, such as the recursive neural networks of the CharNN, are becoming famous for their performance, n-grams still execute a lot of NLP tasks. They are used in statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, ...

--- God in Google Books ---

In 2006, Google created a dataset of n-grams from their digitized book collection and released it online. Recently they also created an n-gram viewer.

This allowed for many socio-linguistic investigations. For example, in October 2018, the New York Times Magazine published an opinion article titled 'It's Getting Harder to Talk About God'. The author, Jonathan Merritt, had analysed the mention of the word 'God' in Google's dataset using the n-gram viewer. He concluded that there had been a decline in the word's usage since the twentieth century. Google's corpus contains texts from the sixteenth century leading up to the twenty-first. However, what the author missed out on was the growing popularity of scientific journals around the beginning of the twentieth century. This new genre that was not mentioning the word God shifted the dataset. If the scientific literature was taken out of the corpus, the frequency of the word 'God' would again flow like a gentle ripple from a distant wave.

--- Grammatical features taken from Twitter influence the stock market ---

The boundaries between academic disciplines are becoming blurred. Economics research mixed with psychology, social science, cognitive and emotional concepts have given rise to a new economics subfield, called 'behavioral economics'. This means that researchers can start to explain stock market mouvement based on factors other than economic factors only. Both the economy and 'public opinion' can influence or be influenced by each other. A lot of research is being done on how to use 'public opinion' to predict tendencies in stock-price changes.

'Public opinion' is estimated from sources of large amounts of public data, like tweets, blogs or online news. Research using machinic data analysis shows that the changes in stock prices can be predicted by looking at 'public opinion', to some degree. There are many scientific articles online, which analyse the press on the 'sentiment' expressed in them. An article can be marked as more or less positive or negative. The annotated press articles are then used to train a machine learning model, which predicts stock market trends, marking them as 'down' or 'up'. When a company gets bad press, traders sell. On the contrary, if the news is good, they buy.

A paper by Haikuan Liu of the Australian National University states that the tense of verbs used in tweets can be an indicator of the frequency of financial transactions. His idea is based on the fact that verb conjugation is used in psychology to detect the early stages of human depression.

Reference

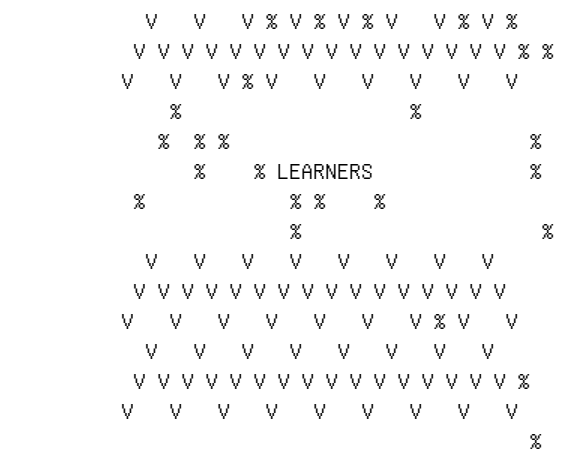
Paper: 'Grammatical Feature Extraction and Analysis of Tweet Text: An Application towards Predicting Stock Trends', Haikuan Liu, Research School of Computer Science (RSCS), College of Engineering and Computer Science (CECS), The Australian National University (ANU)

--- Bag of words ---

In Natural Language Processing (NLP), 'bag of words' is considered to be an unsophisticated model. It strips text of its context and dismantles it into a collection of unique words. These words are then counted. In the previous sentences, for example, 'words' is mentioned three times, but this is not necessarily an indicator of the text's focus.

The first appearance of the expression 'bag of words' seems to go back to 1954. Zellig Harris, an influential linguist, published a paper called 'Distributional Structure'. In the section called 'Meaning as a function of distribution', he says 'for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use. The linguist's work is precisely to discover these properties, whether for descriptive analysis or for the synthesis of quasi-linguistic systems.'

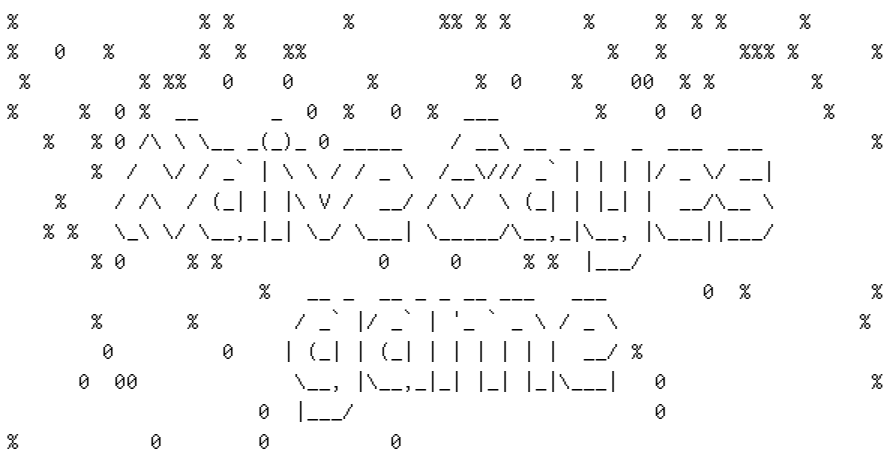
learners learn learners learn learners learn learners learn learners learn learners learn
learners learn learners learn learners learn learners learn learners learn
learners learn learners learn learners learn learners learn
learners learn learners learn learners learn
learners learn learners learn learners learn lea
ners learn learners learn learners learn
learners learn learners learn learners learn
learners learn learners learn learners
earn learners learn learners learn
learners learn learners learn
learners learn learners learn lea
ners learn learners learn learners
learn learners learn learners
earn learners learn learne
s learn learners learn
learners learn learners learn
learners learn learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn learners
learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn
learners learn learners
learn learners learn
learners learn
learners learn
learners learn lea
ners learn learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn learners
earn learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn



Learners are the algorithms that distinguish machine learning practices from other types of practices. They are pattern finders, capable of crawling through data and generating some kind of specific 'grammar'. Learners are based on statistical techniques. Some need a large amount of training data in order to function, others can work with a small annotated set. Some perform well in classification tasks, like spam identification, others are better at predicting numbers, like temperatures, distances, stock market values, and so on.

The terminology of machine learning is not yet fully established. Depending on the field, whether statistics, computer science or the humanities, different terms are used. Learners are also called classifiers. When we talk about Learners, we talk about the interwoven functions that have the capacity to generate other functions, evaluate and readjust them to fit the data. They are good at understanding and revealing patterns. But they don't always distinguish well which of the patterns should be repeated.

In software packages, it is not always possible to distinguish the characteristic elements of the classifiers, because they are hidden in underlying modules or libraries. Programmers can invoke them using a single line of code. For this exhibition, we therefore developed two table games that show in detail the learning process of simple, but frequently used classifiers.



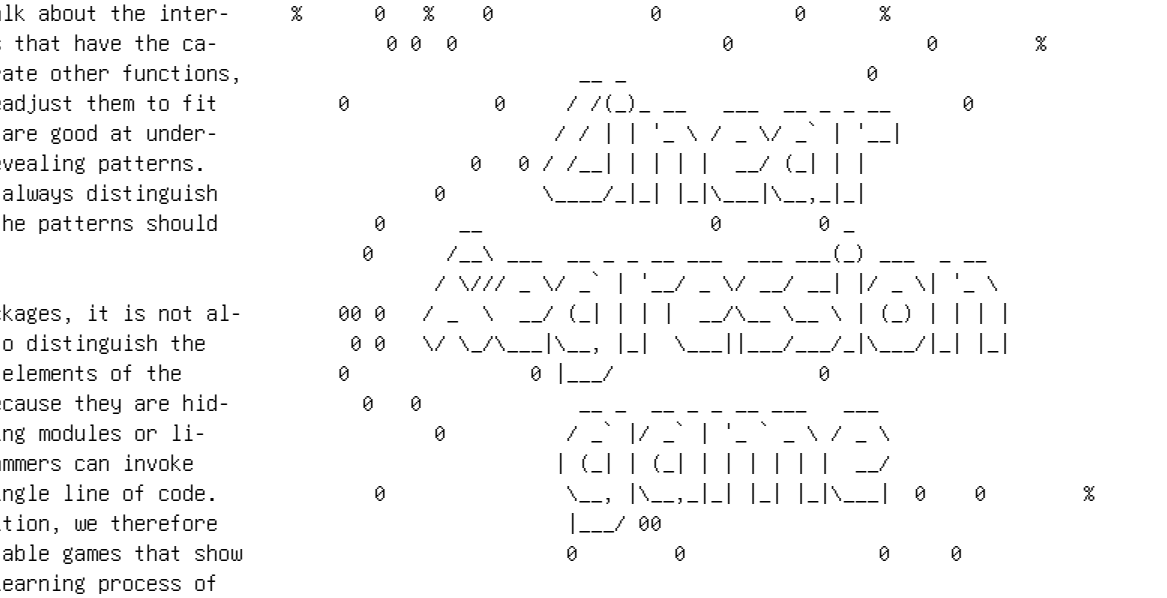
by Algolit

In machine learning Naive Bayes methods are simple probabilistic classifiers that are widely applied for spam filtering and deciding whether a text is positive or negative.

They require a small amount of training data to estimate the necessary parameters. They can be extremely fast compared to more sophisticated methods. They are difficult to generalize, which means that they perform on specific tasks, demanding to be trained with the same style of data that will be used to work with afterwards.

This game allows you to play along the rules of Naive Bayes. While manually executing the code, you create your own playful model that 'just works'. A word of caution is necessary: because you only train it with 6 sentences - instead of the minimum 2000 - it is not representative at all!

Concept & realisation: An Mertens



by Algolit

Linear Regression is one of the best-known and best-understood algorithms in statistics and machine learning. It has been around for almost 200 years. It is an attractive model because the representation is so simple. In statistics, linear regression is a statistical method that allows to summarize and study relationships between two continuous (quantitative) variables.

By playing this game you will realize that as a player you have a lot of decisions to make. You will experience what it means to create a coherent dataset, to decide what is in and what is not in. If all goes well, you will feel the urge to change your data in order to obtain better results. This is part of the art of approximation that is at the basis of all machine learning practices.

Concept & realisation: An Mertens

Traité de Documentation. Three algorithmic poems.

by Rémi Forte, designer-researcher at L'Atelier national de recherche typographique, Nancy, France

serigraphy on paper, 60 × 80 cm, 25 ex., 2019, for sale at the reception of the Mundaneum.

The poems, reproduced in the form of three posters, are an algorithmic and poetic re-reading of Paul Otlet's 'Traité de documentation'. They are the result of an algorithm based on the mysterious rules of human intuition. It has been applied to a fragment taken from Paul Otlet's book and is intended to be representative of his bibliological practice.

For each fragment, the algorithm splits the text, words and punctuation marks are counted and reordered into a list. In each line, the elements combine and exhaust the syntax of the selected fragment. Paul Otlet's language remains perceptible but exacerbated to the point of absurdity. For the reader, the systematization of the text is disconcerting and his reading habits are disrupted.

Built according to a mathematical equation, the typographical composition of the poster is just as systematic as the poem. However, friction occurs occasionally; loop after loop, the lines extend to bite on the neighbouring column. Overlays are created and words are hidden by others. These telescopic handlers draw alternative reading paths.

--- Naive Bayes & Viagra ---

Naive Bayes is a famous learner that performs well with little data. We apply it all the time. Christian and Griffiths state in their book, 'Algorithms To Live By', that 'our days are full of small data'. Imagine, for example, that you're standing at a bus stop in a foreign city. The other person who is standing there has been waiting for 7 minutes. What do you do? Do you decide to wait? And if so, for how long? When will you initiate other options? Another example. Imagine a friend asking advice about a relationship. He's been together with his new partner for a month. Should he invite the partner to join him at a family wedding?

Having pre-existing beliefs is crucial for Naive Bayes to work. The basic idea is that you calculate the probabilities based on prior knowledge and given a specific situation.

The theorem was formulated during the 1740s by Thomas Bayes, a reverend and amateur mathematician. He dedicated his life to solving the question of how to win the lottery. But Bayes' rule was only made famous and known as it is today by the mathematician Pierre Simon Laplace in France a bit later in the same century. For a long time after La Place's death, the theory sank into oblivion until it was dug up again during the Second World War in an effort to break the Enigma code.

Most people today have come in contact with Naive Bayes through their email spam folders. Naive Bayes is a widely used algorithm for spam detection. It is by coincidence that Viagra, the erectile dysfunction drug, was approved by the US Food & Drug Administration in 1997, around the same time as about 10 million users worldwide had made free webmail accounts. The selling companies were among the first to make use of email as a medium for advertising: it was an intimate space, at the time reserved for private communication, for an intimate product. In 2001, the first SpamAssasin programme relying on Naive Bayes was uploaded to SourceForge, cutting down on guerilla email marketing.

Reference
Machine Learners, by Adrian MacKenzie, MIT Press, Cambridge, US, November 2017.

--- Naive Bayes & Enigma ---

This story about Naive Bayes is taken from the book 'The Theory That Would Not Die', written by Sharon Bertsch McGrayne. Among other things, she describes how Naive Bayes was soon forgotten after the death of Pierre Simon Laplace, its inventor. The mathematician was said to have failed to credit the works of others. Therefore, he suffered widely circulated charges against his reputation. Only after 150 years was the accusation refuted.

Fast forward to 1939, when Bayes' rule was still virtually taboo, dead and buried in the field of statistics. When France was occupied in 1940 by Germany, which controlled Europe's factories and farms, Winston Churchill's biggest worry was the U-boat peril. U-boat operations were tightly controlled by German headquarters in France. Each submarine received orders as coded radio messages long after it was out in the Atlantic. The messages were encrypted by word-scrambling machines, called Enigma machines. Enigma looked like a complicated typewriter. It was invented by the German firm Scherbius & Ritter after the First World War, when the need for message-encoding machines had become painfully obvious.

Interestingly, and luckily for Naive Bayes and the world, at that time, the British government and educational systems saw applied mathematics and statistics as largely irrelevant to practical problem-solving. So the British agency charged with cracking German military codes mainly hired men with linguistic skills. Statistical data was seen as bothersome because of its detail-oriented nature. So wartime data was often analysed not by statisticians, but by biologists, physicists, and theoretical mathematicians. None of them knew that the Bayes rule was considered to be unscientific in the field of statistics. Their ignorance proved fortunate.

It was the now famous Alan Turing – a mathematician, computer scientist, logician, cryptoanalyst, philosopher and theoretical biologist – who used Bayes' rules probabilities system to design the 'bombe'. This was a high-speed electromechanical machine for testing every possible arrangement that an Enigma machine would produce. In order to crack the naval codes of the U-boats, Turing simplified the 'bombe' system using Baysian methods. It turned the UK headquarters into a code-breaking factory. The story is well illustrated in The Imitation Game, a film by Morten Tyldum dating from 2014.

--- A story about sweet peas ---

Throughout history, some models have been invented by people with ideologies that are not to our liking. The idea of regression stems from Sir Francis Galton, an influential nineteenth-century scientist. He spent his life studying the problem of heredity – understanding how strongly the characteristics of one generation of living beings manifested themselves in the following generation. He established the field of eugenics, defining it as 'the study of agencies under social control that may improve or impair the racial qualities of future generations, either physically or mentally'. On Wikipedia, Galton is a prime example of scientific racism.

Galton initially approached the problem of heredity by examining characteristics of the sweet pea plant. He chose this plant because the species can self-fertilize. Daughter plants inherit genetic variations from mother plants without a contribution from a second parent. This characteristic eliminates having to deal with multiple sources.

Galton's research was appreciated by many intellectuals of his time. In 1869, in 'Hereditary Genius', Galton claimed that genius is mainly a matter of ancestry and he believed that there was a biological explanation for social inequality across races. Galton even influenced his half-cousin Charles Darwin with his ideas. After reading Galton's paper, Darwin stated, 'You have made a convert of an opponent in one sense for I have always maintained that, excepting fools, men did not differ much in intellect, only in zeal and hard work'. Luckily, the modern study of heredity managed to eliminate the myth of race-based genetic difference, something Galton tried hard to maintain.

Galton's major contribution to the field was linear regression analysis, laying the groundwork for much of modern statistics. While we engage with the field of machine learning, Algolit tries not to forget that ordering systems hold power, and that this power has not always been used to the benefit of everyone. Machine learning has inherited many aspects of statistical research, some less agreeable than others. We need to be attentive, because these world views do seep into the algorithmic models that create new orders.

References

http://galton.org/letters/darwin/correspondence.htm
https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537
http://www.paramoulipist.be/?p=1693

--- Perceptron ---

We find ourselves in a moment in time in which neural networks are sparking a lot of attention. But they have been in the spotlight before. The study of neural networks goes back to the 1940s, when the first neuron metaphor emerged. The neuron is not the only biological reference in the field of machine learning - think of the word corpus or training. The artificial neuron was constructed in close connection to its biological counterpart.

Psychologist Frank Rosenblatt was inspired by fellow psychologist Donald Hebb's work on the role of neurons in human learning. Hebb stated that 'cells that fire together wire together'. His theory now lies at the basis of associative human learning, but also unsupervised neural network learning. It moved Rosenblatt to expand on the idea of the artificial neuron.

In 1962, he created the Perceptron, a model that learns through the weighting of inputs. It was set aside by the next generation of researchers, because it can only handle binary classification.

This means that the data has to be clearly separable, as for example, men and women, black and white. It is clear that this type of data is very rare in the real world. When the so-called first AI winter arrived in the 1970s and the funding decreased, the Perceptron was also neglected. For ten years it stayed dormant. When spring settled at the end of the 1980s, a new generation of researchers picked it up again and used it to construct neural networks. These contain multiple layers of Perceptrons. That is how neural networks saw the light. One could say that the current machine learning season is particularly warm, but it takes another winter to know a summer.

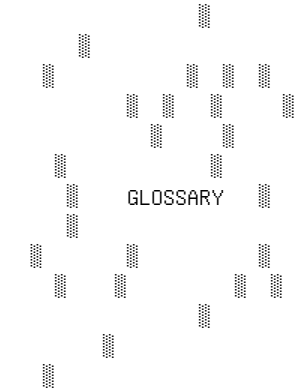
--- BERT ---

Some online articles say that the year 2018 marked a turning point for the field of Natural Language Processing (NLP). A series of deep-learning models achieved state-of-the-art results on tasks like question-answering or sentiment-classification. Google’s BERT algorithm entered the machine learning competitions of last year as a sort of 'one model to rule them all'. It showed a superior performance over a wide variety of tasks.

BERT is pre-trained; its weights are learned in advance through two unsupervised tasks. This means BERT doesn’t need to be trained from scratch for each new task. You only have to finetune its weights. This also means that a programmer wanting to use BERT, does not know any longer what parameters BERT is tuned to, nor what data it has seen to learn its performances.

BERT stands for 'Bidirectional Encoder Representations from Transformers'. This means that BERT allows for bidirectional training. The model learns the context of a word based on all of its surroundings, left and right of a word. As such, it can differentiate between 'I accessed the bank account' and 'I accessed the bank of the river'.

Some facts:
- BERT_large, with 345 million parameters, is the largest model of its kind. It is demonstrably superior on small-scale tasks to BERT_base, which uses the same architecture with 'only' 110 million parameters.
- to run BERT you need to use TPUs. These are the Google's processors (CPUs) especially engineered for TensorFLow, the deep-learning platform. TPU's renting rates range from $8/hr till $394/hr. Algolit doesn't want to work with off-the-shelf packages, we are interested in opening up the black-box. In that case, BERT asks for quite some savings in order to be used.



GLOSSARY

This is a non-exhaustive wordlist, based on terms that are frequently used in the exhibition. It might help visitors who are not familiar with the vocabulary related to the field of Natural Language Processing (NLP), Algolit or the Mundaneum.

* ALGOLIT
A group from Brussels involved in artistic research on algorithms and literature. Every month they gather to experiment with code and texts that are published under free licenses.
http://www.algolit.net

* ALGOLITERARY
Word invented by Algolit for works that explore the point of view of the algorithmic storyteller. What kind of new forms of storytelling do we make possible in dialogue with machinic agencies?

* ALGORITHM
A set of instructions in a specific programming language, that takes an input and produces an output.

* ANNOTATION
The annotation process is a crucial step in supervised machine learning where the algorithm is given examples of what it needs to learn. A spam filter in training will be fed examples of spam and real messages. These examples are entries, or rows from the dataset with a label, spam or non-spam. The labelling of a dataset is work executed by humans, they pick a label for each row of the dataset. To ensure the quality of the labels multiple annotators see the same row and have to give the same label before an example is included in the training data.

* AI OR ARTIFICIAL INTELLIGENCES
In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals. Computer science defines AI research as the study of ‘intelligent agents’. Any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. More specifically, Kaplan and Haenlein define AI as ‘a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation’. Colloquially, the term ‘artificial intelligence’ is used to describe machines that mimic ‘cognitive’ functions that humans associate with other human minds, such as ‘learning’ and ‘problem solving’. (Wikipedia)

* BAG OF WORDS
The bag-of-words model is a simplifying representation of text used in Natural Language Processing (NLP). In this model, a text is represented as a collection of its unique words, disregarding grammar, punctuation and even word order. The model transforms the text into a list of words and how many times they're used in the text, or quite literally a bag of words. Bag of words is often used as a baseline, on which the new model has to perform better.

* CHARACTER N-GRAM
A technique that is used for authorship recognition. When using character n-grams, texts are considered as sequences of characters. Let's consider the character trigram. All the overlapping sequences of three characters are isolated. For example, the character 3-grams of 'Suicide', would be, 'Sui', 'uic', 'ici', 'cid' etc. Patterns found with character n-grams focus on stylistic choices that are unconsciously made by the author. The patterns remain stable over the full length of the text.

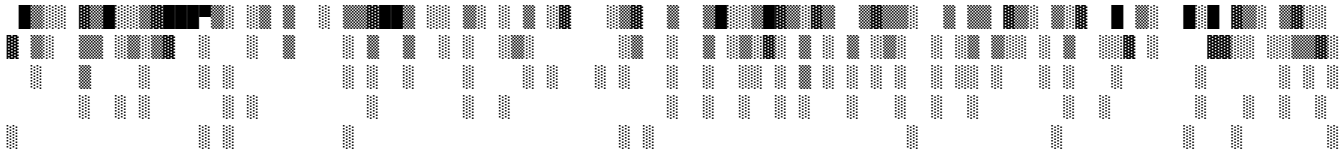
* CLASSICAL MACHINE LEARNING
Naive Bayes, Support Vector Machines and Linear Regression are called classical machine learning algorithms. They perform well when learning with small datasets. But they often require complex Readers. The task the Readers do, is also called feature-engineering (see below). This means that a human needs to spend time on a deep exploratory data analysis of the dataset.

* CONSTANT
Constant is a non-profit, artist-run organisation based in Brussels since 1997 and active in the fields of art, media and technology. Algolit started as a project of Constant in 2012.
http://constantvzw.org

* DATA WORKERS
Artificial intelligences that are developed to serve, entertain, record and know about humans. The work of these machinic entities is usually hidden behind interfaces and patents. In the exhibition, algorithmic storytellers leave their invisible underworld to become interlocutors.

* DUMP
According to the English dictionary, a dump is an accumulation of refused and discarded materials or the place where such materials are dumped. In computing a dump refers to a ‘database dump’, a record of data from a database used for easy downloading or for backing up a database. Database dumps are often published by free software and free content projects, such as Wikipedia, to allow reuse or forking of the database.

* FEATURE ENGINEERING
The process of using domain knowledge of the data to create features that make machine learning algorithms work. This means that a human needs to spend time on a deep exploratory data analysis of the dataset. In Natural Language Processing (NLP) features can be the frequency of words or letters, but also syntactical elements like nouns, adjec-



tives, or verbs. The most significant features for the task to be solved, must be carefully selected and passed over to the classical machine learning algorithm.

* FLOSS OR FREE LIBRE OPEN SOURCE SOFTWARE
Software that anyone is freely licensed to use, copy, study, and change in any way, and the source code is openly shared so that people are encouraged to voluntarily improve the design of the software. This is in contrast to proprietary software, where the software is under restrictive copyright licensing and the source code is usually hidden from the users. (Wikipedia)

* GIT
A software system for tracking changes in source code during software development. It is designed for coordinating work among programmers, but it can be used to track changes in any set of files. Before starting a new project, programmers create a "git repository" in which they will publish all parts of the code. The git repositories of Algolit can be found on https://gitlab.constant vzw.org/algolit.

* GUTENBERG.ORG
Project Gutenberg is an online platform run by volunteers to ‘encourage the creation and distribution of eBooks’. It was founded in 1971 by American writer Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of 23 June 2018, Project Gutenberg reached 57,000 items in its collection of free eBooks. (Wikipedia)

* HENRI LA FONTAINE
Henri La Fontaine (1854-1943) is a Belgian politician, feminist and pacifist. He was awarded the Nobel Peace Prize in 1913 for his involvement in the International Peace Bureau and his contribution to the organization of the peace movement. In 1895, together with Paul Otlet, he created the International Bibliography Institute, which became the Mundaneum. Within this institution, which aimed to bring together all the world's knowledge, he contributed to the development of the Universal Decimal Classification (CDU) system.

* KAGGLE
An online platform where users find and publish data sets, explore and build machine learning models, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. About half a million data scientists are active on Kaggle. It was founded by Goldbloom and Ben Hamner in 2010 and acquired by Google in March 2017.

* LITERATURE
Algolit understands the notion of literature in the way a lot of other experimental authors do. It includes all linguistic production, from the dictionary to the Bible, from Virginia Woolf's entire work to all versions of Terms of Service published by Google since its existence.

* MACHINE LEARNING MODELS
Algorithms based on statistics, mainly used to analyse and predict situations based on existing cases. In this exhibition we focus on machine learning models for text processing or Natural language processing', in short, 'nlp'. These models have learned to perform a specific task on the basis of existing texts. The models are used for search engines, machine translations and summaries, spotting trends in new media networks and news feeds. They influence what you get to see as a user, but also have their word to say in the course of stock exchanges worldwide, the detection of cybercrime and vandalism, etc.

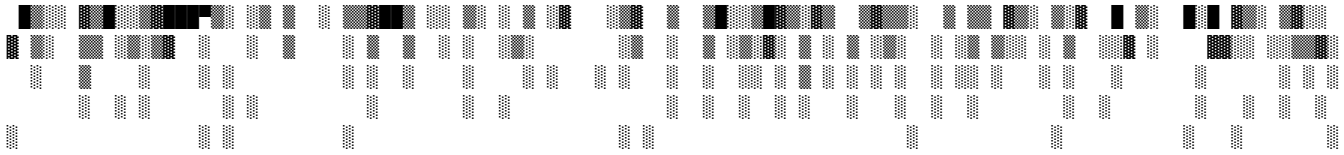
* MARKOV CHAIN
Algorithm that scans the text for the transition probability of letter or word occurrences, resulting in transition probability tables which can be computed even without any semantic or grammatical natural language understanding. It can be used for analyzing texts, but also for recombining them. It is is widely used in spam generation.

* MECHANICAL TURK
The Amazon Mechanical Turk is an online platform for humans to execute tasks that algorithms cannot. Examples include annotating sentences as being positive or negative, spotting number plates, discriminating between face and non-face. The jobs posted on this platform are often paid less than a cent per task. Tasks that are more complex or require more knowledge can be paid up to several cents. Many academic researchers use Mechanical Turk as an alternative to have their students execute these tasks.

* MUNDANEUM
In the late nineteenth century two young Belgian jurists, Paul Otlet (1868-1944), ‘the father of documentation’, and Henri La Fontaine (1854-1943), statesman and Nobel Peace Prize winner, created The Mundaneum. The project aimed at gathering all the world’s knowledge and file it using the Universal Decimal Classification (UDC) system that they had invented.

* NATURAL LANGUAGE
A natural language or ordinary language is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Natural languages can take different forms, such as speech or signing. They are different from constructed and formal languages such as those used to program computers or to study logic. (Wikipedia)

* NLP OR NATURAL LANGUAGE PROCESSING
Natural language processing (NLP) is a collective term referring to automatic computational processing of human languages. This includes algorithms that take human-produced text as input, and attempt



to generate text that resembles it.

* NEURAL NETWORKS
Computing systems inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems ‘learn’ to perform tasks by considering examples, generally without being programmed with any task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as ‘cat’ or ‘no cat’ and using the results to identify cats in other images. They do this without any prior knowledge about cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the learning material that they process. (Wikipedia)

* OPTICAL CHARACTER RECOGNITION (OCR)
Computer processes for translating images of scanned texts into manipulable text files.

* ORACLE
Oracles are prediction or profiling machines, a specific type of algorithmic models, mostly based on statistics. They are widely used in smartphones, computers, tablets.

* OULIPO
Oulipo stands for Ouvroir de litterature potentielle (Workspace for Potential Literature). Oulipo was created in Paris by the French writers Raymond Queneau and François Le Lionnais. They rooted their practice in the European avant-garde of the twentieth century and in the experimental tradition of the 1960s. For Oulipo, the creation of rules becomes the condition to generate new texts, or what they call potential literature. Later, in 1981, they also created ALAMO, Atelier de littérature assistée par la mathématique et les ordinateurs (Workspace for literature assisted by maths and computers).

* PAUL OTLET
Paul Otlet (1868 – 1944) was a Belgian author, entrepreneur, visionary, lawyer and peace activist; he is one of several people who have been considered the father of information science, a field he called 'documentation'. Otlet created the Universal Decimal Classification, that was widespread in libraries. Together with Henri La Fontaine he created the Palais Mondial (World Palace), later, the Mundaneum to house the collections and activities of their various organizations and institutes.

* PYTHON
The main programming language that is globally used for natural language processing, was invented in 1991 by the Dutch programmer Guido Van Rossum.

* RULE-BASED MODELS
Oracles can be created using different techniques. One way is to manually define rules for them. As prediction models they are then called rule-based models, opposed to statistical models. Rule-based models are handy for tasks that are specific, like detecting when a scientific paper concerns a certain molecule. With very little sample data, they can perform well.

* SENTIMENT ANALYSIS
Also called 'opinion mining' A basic task in sentiment analysis is classifying a given text as positive, negative or neutral. Advanced, 'beyond polarity' sentiment classification looks, for instance, at emotional states such as 'angry', 'sad' and 'happy'. Sentiment analysis is widely applied to user materials such as reviews and survey responses, comments and posts on social media, and healthcare materials for applications that range from marketing to customer service, from stock exchange transactions to clinical medicine.

* SUPERVISED MACHINE LEARNING MODELS
For the creation of supervised machine learning models, humans annotate sample text with labels before feeding it to a machine to learn. Each sentence, paragraph or text is judged by at least 3 annotators whether it is spam or not spam, positive or negative etc.

* TRAINING DATA
Machine learning algorithms need guidance. In order to separate one thing from another, they need texts to extract should carefully choose the training material, and adapt it to the machine's task. It doesn't make sense to train a machine with nineteenth-century novels if its mission is to analyze tweets.

* UNSUPERVISED MACHINE LEARNING MODELS
Unsupervised machine learning models don't need the step of annotation of the data by humans. This saves a lot of time, energy, money. Instead, they need a large amount of training data, which is not always available and can take a long cleaning time beforehand.

* WORD EMBEDDINGS
Language modelling techniques that through multiple mathematical operations of counting and ordering, plot words into a multi-dimensional vector space. When embedding words, they transform from being distinct symbols into mathematical objects that can be multiplied, divided, added or substracted.

* WORDNET
Wordnet is a combination of a dictionary and a thesaurus that can be read by machines. According to Wikipedia it was created in the Cognitive Science Laboratory of Princeton University starting in 1985. The project was initially funded by the US Office of Naval Research and later also by other US government agencies including DARPA, the National Science Foundation, the Disruptive Technology Office (formerly the Advanced Research and Development Activity), and REFLEX.

ˋ humans learn with machines ˊ   ⌣ machines learn from machines ˏ   ⌣ machines learn with humans ˏ   ˋ
humans learn from machines ˎ   ˊ machines learn with machines ⌢   ˊ machines learn from humans ˎ   ⌢
humans learn with humans ˏ   ˋ humans learn from humans ˏ   ⌢ humans learn with machines ˎ   ⌣ mac
ines learn from machines ⌣   ⌣ machines learn with humans ˎ   ⌣ humans learn from machines ˋ   ˎ
achines learn with machines ⌢   ˋ machines learn from humans ˊ   ˋ humans learn with humans ˏ   ˏ
humans learn from humans ⌣   ˏ humans learn with machines ⌢   ⌢ machines learn from machines ⌢
machines learn with humans ˏ   ˊ humans learn from machines ˊ   ⌢ machines learn with machines ˋ
ˊ machines learn from humans ˊ   ⌢ humans learn with humans ˋ   ˎ humans learn from humans ˏ
ˊ humans learn with machines ⌣   ⌣ machines learn from machines ⌣   ˎ machines learn with humans
⌢   ⌢ humans learn from machines ⌣   ˊ machines learn with machines ˊ   ˎ machines learn from
umans ˎ   ˏ humans learn with humans ˏ   ˎ humans learn from humans ˊ   ⌢ humans learn with ma
hines ˊ   ⌢ machines learn from machines ˋ   ⌢ machines learn with humans ˋ   ˏ humans learn f
om machines ˋ   ⌣ machines learn with machines ˊ   ⌣ machines learn from humans ˊ   ⌢ humans l
arn with humans ⌣   ⌣ humans learn from humans ˋ   ˏ humans learn with machines ˎ   ⌣ machines
learn from machines ˊ   ˊ machines learn with humans ⌢   ˏ humans learn from machines ˋ   ⌢ ma
hines learn with machines ˎ   ˎ machines learn from humans ˋ   ⌢ humans learn with humans ˎ
humans learn from humans ˋ   ˊ humans learn with machines ⌢   ˋ machines learn from machines ˏ
⌢ machines learn with humans ˋ   ˎ humans learn from machines ˎ   ˏ machines learn with machines
ˊ   ˏ machines learn from humans ˏ   ⌣ humans learn with humans ⌢   ˏ humans learn from human
⌢   ˊ humans learn with machines ⌣   ˏ machines learn from machines ˊ   ⌢ machines learn w
th humans ⌣   ˋ humans learn from machines ˋ   ˎ machines learn with machines ⌢   ⌢ machine
learn from humans ˏ   ˎ humans learn with humans ⌢   ˏ humans learn from humans ⌢   ⌢ huma
s learn with machines ⌣   ⌣ machines learn from machines ˊ   ˏ machines learn with humans ⌣
ˎ humans learn from machines ˊ   ˊ machines learn with machines ˊ   ˋ machines learn from human
ˊ   ⌢ humans learn with humans ˋ   ⌣ humans learn from humans ⌣   ˏ humans learn with mach
nes ˊ   ˋ machines learn from machines ˋ   ˊ machines learn with humans ˏ   ˊ humans learn
rom machines ˏ   ˋ machines learn with machines ˏ   ˊ machines learn from humans ⌣   ˏ huma
s learn with humans ˎ   ˊ humans learn from humans ˏ   ⌣ humans learn with machines ˋ   ˋ m
chines learn from machines ˊ   ˎ machines learn with humans ⌣   ˎ humans learn from machines ⌢
ˋ machines learn with machines ˊ   ⌣ machines learn from humans ˏ   ˋ humans learn with huma
s ˋ   ⌢ humans learn from humans ˏ   ˊ humans learn with machines ⌢   ˋ machines learn from
machines ˎ   ⌣ machines learn with humans ˋ   ˋ humans learn from machines ˏ   ˏ machines l
arn with machines ⌢   ⌢ machines learn from humans ⌢   ⌣ humans learn with humans ˊ   ˊ hum
ns learn from humans ˏ   ˏ humans learn with machines ⌣   ˋ machines learn from machines ˎ
ˋ machines learn with humans ⌢   ˎ machines learn with humans ⌢   ˊ machines learn from
machines ⌣   ˊ humans learn with machines ˏ   ˎ humans learn from humans ˊ   ⌣ humans learn
with humans ˋ   ˏ machines learn from humans ˊ   ˋ machines learn with machines ˊ   ⌢ human
learn from machines ⌣   ˋ machines learn with humans ˋ   ˊ machines learn from machines ˊ
ˏ humans learn with machines ⌢   ˋ humans learn from humans ⌢   ˋ humans learn with humans ˏ
⌣ machines learn from humans ˊ   ˋ machines learn with machines ⌢   ˎ humans learn from machi
es ˊ   ˎ machines learn with humans ˋ   ˋ machines learn from machines ˏ   ˊ humans learn w
th machines ˋ   ⌣ humans learn from humans ˋ   ˋ humans learn with humans ⌢   ⌢ machines le
rn from humans ˋ   ⌣ machines learn with machines ⌣   ⌣ humans learn from machines ⌢   ˏ ma
hines learn with humans ˋ   ˊ machines learn from machines ˊ   ˋ humans learn with machines ⌢
ˏ humans learn from humans ˋ   ⌣ humans learn with humans ˏ   ⌣ machines learn from humans ˎ
ˎ machines learn with machines ˋ   ˋ humans learn from machines ˊ   ˎ machines learn with
umans ⌣   ˋ machines learn from machines ⌣   ˋ humans learn with machines ˏ   ˊ humans lear
from humans ˊ   ˋ humans learn with humans ˏ   ⌣ machines learn from humans ˋ   ⌣ machines
learn with machines ˏ   ˎ humans learn from machines ˊ   ˏ machines learn with humans ˎ   ⌣
machines learn from machines ˊ   ˋ humans learn with machines ⌢   ⌢ humans learn from humans ⌢
ˋ humans learn with humans ˎ   ˏ machines learn from humans ˋ   ⌢ machines learn with machines
ˊ   ˎ humans learn from machines ⌢   ˋ machines learn with humans ˋ   ˊ machines learn from ma
hines ˎ   ˎ humans learn with machines ˏ   ⌣ humans learn from humans ˋ   ˋ humans learn with
umans ⌣   ˋ machines learn from humans ˋ   ⌣ machines learn with machines ˎ   ˏ humans learn f
om machines ˋ   ˎ machines learn with humans ˋ   ˊ machines learn from machines ˋ   ⌢ humans l
arn with machines ⌢   ⌢ humans learn from humans ˎ   ˊ humans learn with humans ˎ   ˋ machines
learn from humans ⌣   ⌣ machines learn with machines ˊ   ˊ humans learn from machines ⌢   ˎ ma
hines learn with humans ˏ   ˊ machines learn from machines ⌢   ˊ humans learn with machines ˊ
ˏ humans learn from humans ˋ   ˎ humans learn with humans ˎ   ˏ machines learn from humans ˎ
ˋ machines learn with machines ⌣   ˊ humans learn from machines ⌢   ⌢ machines learn with humans ˏ
⌣ machines learn from machines ˎ   ˋ humans learn with machines ˊ   ˏ humans learn from huma
s ˋ   ˏ humans learn with humans ˊ   ˎ machines learn from humans ˊ   ˏ machines learn with ma
hines ˋ   ˏ humans learn from machines ˋ   ˊ machines learn with humans ˎ   ˊ machines learn from
machines ⌣   ˎ humans learn with machines ˏ   ⌢ humans learn from humans ˏ   ˎ humans learn with
umans ⌢   ˊ machines learn from humans ⌣   ⌢ machines learn with machines ⌢   ˋ humans learn from
machines ⌢   ˊ machines learn with humans ˏ   ⌢ machines learn from machines ˏ   ⌢ humans learn w
th machines ˊ   ˎ humans learn from humans ˋ   ⌢ humans learn with humans ˋ   ˎ machines learn from
humans ˊ   ˊ machines learn with machines ⌢   ˏ humans learn from machines ⌢   ⌣ machines learn with
machines ⌣   ˎ humans learn with machines ˏ   ⌢ humans learn from humans ˏ   ˎ humans learn with mach
ines ˋ   ˏ humans learn from machines ˋ   ˊ machines learn with humans ˎ   ˊ machines learn from hum

fwo Research Foundation Flanders Opening new horizons
Gestion Centre Ville Mons
DHu.f Research Community Digital Humanities Flanders
ARTS2 École supérieure des arts Academy of arts
mundaneum
Fédération Wallonie-Bruxelles