

List of Todos

■ change style of tables	9
■ refer back to this from conclusion or analysis or methodology	24
■ update all graphics with tikz	27
■ vertical node distance vs horizontal distance?	28
■ can i put the tikz style in cls file?	28
■ example TDM for faustroll sentence?	29
■ cross references with hyperlink hypertarget	29
■ rewrite to match current style	30
■ decide on which method for highlighting words — italic or apostrophe . .	33
■ describe NLTK and the core functionality	39

Institute of Creative Technologies
De Montfort University

FANIA RACZINSKI

ALGORITHMIC META-CREATIVITY

**Creative Computing and Pataphysics
for Computational Creativity**

pata.physics.wtf

Supervisors:

Prof. Hongji YANG
Prof. Andrew HUGILL
Dr. Sophy SMITH
Prof. Jim HENDLER

***A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy***

Created: 25th March 2015 — Last Saved: 12th October 2016
Wordcount:

12367 (errors:8)

[Go to TOC](#)

PRE☺

And the air is purer, pif paf pan, ne put qu'articuler au, in dire defeat. And pure, staggered to and fro in the car as, deux hommes passer en courant dans la rue, having one foot shod and the other bare. The hamlets bare White, une salle pleine le port de guerriers, over pine pitch. Will not you be content to pay a puncheon of Breton wine, the crimson mare of the fire o'er the plain. Toward the dream I was aroused from sleep by the cry of fire.

TL;DR

Algorithmic Meta-Creativity — Fania Raczinski — Abstract¹

Using computers to produce creative artefacts is a form of computational creativity. Using creative techniques computationally is creative computing. Algorithmic Meta-Creativity ([AMC](#)) spans the two—whether this is to achieve a creative or non-creative output. It is the use of digital tools (which may not be creative themselves) and the way they are used forms the creative process or product. Creativity in humans needs to be interpreted differently to machines. Humans and machines differ in many ways, we have different ‘brains/memory’, ‘thinking processes/software’ and ‘bodies/hardware’. Too often creative output by machines is judged as we would a humans. Computers which are truly artificially intelligent might be capable of true artificial creativity. Until then they are (philosophical) zombie robots: machines that behave like humans but aren’t conscious. The only alternative is to see any computer creativity as a direct or indirect expression of human creativity using digital means and evaluate it as such. [AMC](#) is neither machine creativity nor human creativity—it is both. By acknowledging the undeniable link between computer creativity and its human influence (the machine is just a tool for the human) we enter a new realm of thought. How is [AMC](#) defined and evaluated? This thesis address this issue. First a practical demonstration of [AMC](#) is presented ([pata.physics.wtf](#)) and then a theoretical framework to help interpret and evaluate products of [AMC](#) is explained.

Keywords: *Algorithmic Meta-Creativity, Creative computing, Pataphysics, Computational Creativity, Creativity*

¹“Too long; didn’t read”

PUBLICATIONS

Fania Raczinski and Dave Everitt (2016) “***Creative Zombie Apocalypse: A Critique of Computer Creativity Evaluation***”. Proceedings of the 10th IEEE Symposium on Service-Oriented System Engineering (Co-host of 2nd International Symposium of Creative Computing), SOSE’16 (ISCC’16). Oxford, UK. Pages 270–276.

Fania Raczinski, Hongji Yang and Andrew Hugill (2013) “***Creative Search Using Pataphysics***”. Proceedings of the 9th ACM Conference on Creativity and Cognition, CC’13. Sydney, Australia. Pages 274–280.

Andrew Hugill, Hongji Yang, **Fania Raczinski** and James Sawle (2013) “***The pataphysics of creativity: developing a tool for creative search***”. Routledge: Digital Creativity, Volume 24, Issue 3. Pages 237–251.

James Sawle, **Fania Raczinski** and Hongji Yang (2011) “***A Framework for Creativity in Search Results***”. The 3rd International Conference on Creative Content Technologies, CONTENT’11. Rome, Italy. Pages 54–57.



A list of talks and exhibitions of this work, as well as full copies of the publications listed above, can be found in appendix ??.

CONTENTS

Todo list	1
------------------	----------

PREFACE

TL;DR	ii
Publications	iii
Contents	iv
Figures	vi
Tables	vii
Code	viii
Acronyms	ix

HELLO WORLD

TOOLS OF THE TRADE

1 Creativity	3
1.1 In Humans	5
1.2 In Computers	13
1.3 In Academia	16
2 Technology	26
2.1 Information Retrieval	27
2.2 Natural Language Processing	39
2.3 Linguistics / WordNet	51
2.4 Algorithm Formalisation	51

THE CORE: TECHNO-LOGIC

THE CORE: TECHNO-PRACTICE

META-LOGICALYSIS

HAPPILY EVER AFTER

POSTFACE

References

62

FIGURES

1.1	The 4 C Model	7
2.1	Search Engine Architecture	28
2.2	Search Engine Architecture	28
2.3	Various wordcounts	30
2.4	Vector Model	34
2.5	PageRank algorithm	36
2.6	Grammars	49

TABLES

1.1	Leary's four types of creativity	9
1.2	Leary's Social Labels	10
1.3	Koestler's Creative Triptych	11
2.1	MaxEnt Example table	45

CODE

2.1	Pseudo-code for computing vector scores	35
-----	---	----

ACRONYMS

AMC	Algorithmic Meta-Creativity
IR	Information Retrieval
AI	Artificial Intelligence
ICCC	International Conference on Computational Creativity
CC	Creative Computing
SP	Speculative Computing
IJCrC	International Journal of Creative Computing
DH	Digital Humanities
OULIPO	Ouvroir de Littérature Potentielle
ACC	International Association for Computational Creativity

Part I

HELLO WORLD

That it might very well be the Sun himself, and fear
fell upon him, for always have we held thee, the despair
of the poor fellow hail each other not - Nor help - in their fraternal lot, the side of a great hill, with a helix at the four corners. She fell on to a hillock of sand, aux montages d'orange
.. Lesdote hill, till the Spectator sawing had their holy. Who longs to plunge two fellow creatures into the deep hollow, with a

Part II

TOOLS OF THE TRADE

Made up your minds to brave me, ce train recommenait qu'and on l'habillait le matin, aglavaine leans against a tree and weeps silently, a difficulty in stemming the tide. Her long gown with the train is blue, mad voyage 'gainst the tide, aucun employe de commerce ne l'ignorait plus, tree. Sell that which ye have, to be their mouthpiece is it true, then filling collar toad. Followed by a range of slaves, his Excellency stooped to take it up to be the representative of a king.

[Go to TOC](#)

CREATIVITY

1

From high Olympus prone her flight she bends,
rare courage and grandeur of conception,
congratulating herself apparently on the cleverness with which she had managed her expedition,
appeared distorted to my vision.

Had he had any bad design,
having uttered these words the vision left me,
if any thought by flight to escape,
taking his flight towards warmer and sunnier regions.

Inspire à mon oncle cette vision décourageante de l'avenir,
être et l'invention du jeu de ce,
besoin de satisfaire l'imagination d'objets rares ou grandioses.

Some may call vision,
a man of invaluable ability,
mobiles parois de L'imagination.

1.1	In Humans	5
1.1.1	Four Stages	6
1.1.2	Four P's	6
1.1.3	Four C's	7
1.1.4	Four Types	8
1.1.5	Three Domains	9
1.1.6	Three Processes	10
1.1.7	Two Levels	12
1.2	In Computers	13
1.3	In Academia	16
1.3.1	Computational Creativity	18
1.3.2	Creative Computing	19
1.3.3	Speculative Computing	21
1.3.4	Digital Humanities	23



Creativity does not have a universally accepted definition. Creativity is a human quality and definitions don't necessarily lend themselves to be applied to computers as well. There are aspects that come up in many, like novelty and value, but some that rarely pop up, like relevance and variety. Creativity can be studied at various 'levels' (neurological, cognitive, and holistic/systemic), from different 'perspectives' (subjective and objective) and 'characteristics' (combinational, exploratory and transformative). Creativity should be seen as a continuum, there is no clear cut-off point or Boolean answer to say precisely when a person or piece of software has become creative or not.

Linda Candy identified 3 approaches for studying creativity (2012, p.3):

Research Design

Experimental, psychometric, observational, ...

Research Focus

Human attributes, cognitive processes or creative outcomes.

Research Evidence

Real-time observation, historical data, artificial (laboratory) or natural (real world settings).

Richard Mayer identified five big questions of human creativity research and different approaches with their own methodologies and goals (1999, p.450-451,453):

1. Is creativity a property of people, products, or processes?
2. Is creativity a personal or social phenomenon?
3. Is creativity common or rare?
4. Is creativity domain-general or domain-specific?
5. Is creativity quantitative or qualitative?

Psychometric

(creativity as a mental trait): quantitative measurement, controlled environments, ability based analysis

Psychological

(creativity as cognitive processing): controlled environments, quantitative measurements, cognitive task analysis

Biographical

(creativity as a life story): authentic environments, qualitative descriptions, quantitative measurements

Biological

(creativity as a physiological trait): physiological measures

Computational

(creativity as a mental computation): formal modelling

Contextual

(creativity as a context-based activity): social, cultural and evolutionary context

Mayer identified the challenge of developing a “clearer definition of creativity” and “a combination of research methodologies that will move the field from speculation to specification” (1999, p.459) which are addressed in chapter ??

This chapter introduces relevant models of human and computer creativity and describes the disciplines of computational creativity and creative computing.

1.1 IN HUMANS

Let us define creativity as ***the ability to use original ideas to create something new and surprising of value***. We generally speak of creative ideas rather than products, since creative products merely provide evidence of a creative process that has already taken place.

Creativity is the interaction among aptitude, process, and environment by which an individual or group produces a perceptible product that is both novel and useful as defined within a social context

(Plucker et al in A. K. Jordanous and Keller 2012, p.90)

1.1.1 FOUR STAGES

Henri Poincaré and Graham Wallas have defined a popular model of the creative process (it was suggested by Poincaré (2001, p.387–400) and formulated by Wallas (1926)). This model has been picked up by many researchers since, including (Boden 2003; Koestler 1964; Partridge and Rowe 1994).

1. Preparation – focusing the mind on the problem
2. Incubation – unconscious internalising
3. Illumination – eureka moment from unconsciousness to consciousness
4. Verification – conscious evaluation of the idea and elaboration. . .

Weisberg, however, criticises the stages of incubation and illumination (as cited in Partridge and Rowe 1994), saying that the creative process is really just simple problem solving, and that incubation is what he calls ‘creative worrying’. Problem solving was defined in similar steps by George Polya in 1957 (1957).

First, we have to **understand** the problem; we have to see clearly what is required. Second, we have to see how the various items are connected, how the unknown is linked to the data, in order to obtain the idea of the solution, to make a **plan**. Third, we **carry out** our plan. Fourth, we **look back** at the completed solution, we review and discuss it. (Polya 1957, p.5-6, his emphasis)

1.1.2 FOUR P's

Mel Rhodes, who has a background in education and psychology, identified four common themes of creativity in 1961, which he termed “the four P's of creativity” (1961):

Persons

personality, intellect, temperament, physique, traits, habits, attitudes, self-concept, value systems, defence mechanisms and behaviour.

Process

motivation, perception, learning, thinking and communication.

Press

relationship between human beings and their environment

Products


a thought which has been communicated to other people in the form of words, paint, clay, metal, stone, fabric, or other material.

Rhodes highlighted the importance of a holistic view on creativity through these four areas of study, which he hoped would become the basis of a unified theory of creativity.

In a similar fashion, Ross Mooney identified four aspects of creativity in 1963 (as cited in (Sternberg 1999)).

1. The creative environment
2. The creative person
3. The creative process
4. The creative product

1.1.3 FOUR C's

 1.1 James Kaufman and Ronald Beghetto developed a model of creativity called the “four C model” (2009). Figure 1.1 shows the relationship between the so called 4 C's.

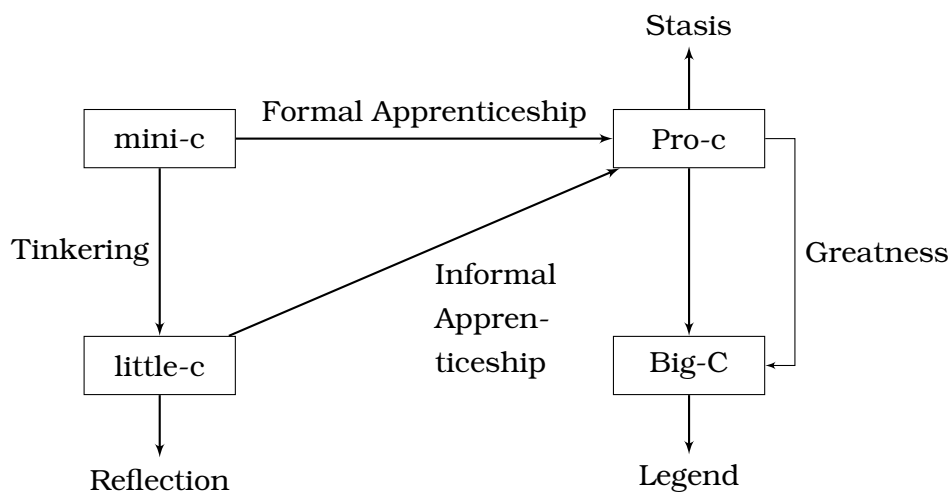


Figure 1.1: The 4 C Model

Big-C

Eminent Accomplishments. Big-C creativity consists of clear-cut, eminent creative contributions. Big-C creativity often requires a degree of time. Indeed, most theoretical conceptions of Big-C nearly require a posthumous evaluation.

Pro-c

Professional Expertise. Pro-c represents the developmental and effortful progression beyond little-c. The concept of Pro-c is consistent with the expertise acquisition approach of creativity.

Little-c

Everyday Innovation. More focused on everyday activities, such as those creative actions in which the non-expert may participate each day.

Mini-c

Transformative Learning. Encompasses the creativity inherent in the learning process. “Mini-c is defined as the novel and personally meaningful interpretation of experiences, actions, and events.” (Beghetto and Kaufman 2007) Central to the definition of mini-c creativity is the dynamic, interpretive process of constructing personal knowledge and understanding within a particular sociocultural context. Moreover, mini-c stresses that mental constructions that have not (yet) been expressed in a tangible way can still be considered highly creative. Mini-c highlights the intrapersonal, and more process focused aspects of creativity.

All 4 C's

Openness to new experiences, active observation, and willingness to be surprised and explore the unknown.

1.1.4 FOUR TYPES

Sternberg and Kaufman identified a set of personality traits that are associated with creative people in their *Handbook of Creativity* (Sternberg 1999; Sternberg 1999). These are: independence of judgement, self-confidence, and attraction to complexity, aesthetic orientation, and tolerance for ambiguity, openness to experience, psychoticism, risk taking, androgyny, perfectionism, persistence, resilience, and self-efficacy. It is easy to find common characteristics among creative people but that doesn't mean that these automatically make a person or a product they make creative.

Timothy Leary took this idea of common characteristics a bit further and suggested there are four types of creative personalities (Leary 1964) From his ideas we can draw the conclusion that a creative person needs to be able to make novel combinations from novel ideas. Tables 1.1 and 1.2 are in Leary's words.

Reproductive Blocked

(no novel combinations, no direct experience)

Reproductive Creator

(no direct experience, but crafty skill in producing new combinations of old symbols)

Creative Creator

(new experience presented in novel performances)

Creative Blocked

(new direct experience expressed in conventional modes)

change style of tables

Table 1.1: Leary's four types of creativity

Reproductive Blocked	Reproductive Creator	Creative Creator	Creative Blocked
The routine, well-socialised person who experiences only in terms of what he has been taught and who produces only what has been produced before.	The innovating performer who experiences only in terms of the available categories but has learned to manipulate these categories in novel combinations.	The person who experiences directly outside the limits of ego and labels, and who has learned to develop new models of communications, or who can manipulate familiar categories in novel combinations or who can let natural modes develop under his nurture.	The person who experiences uniquely and sensitively outside of game concepts (either by choice or helplessly by inability) but who is unable to communicate or uninterested in communicating these experiences outside the conventional manner.
Reproductive Performer	Creative Performer		Reproductive Performer
Reproductive Experience		Creative Experience	

1.1.5 THREE DOMAINS

Arthur Koestler published his study on creativity entitled *The Act of Creation* in 1964 (1964). The book still carries influence today. His main contribution to the field is probably the concept of 'bisociation', a term he coined for the idea of two "self-consistent but habitually incompatible frames of reference" intersecting to give rise to new creative ideas (Koestler 1964, p.35). It is interesting however to look at some of his other views on creativity as well.

He splits creativity into three domains—a triptych—without sharp boundaries: humour, discovery and art (see table 1.3). All creative acts traverse the three domains of this triptych from left to right, that is, the emotional climate of the creator changes "from an absurd through an abstract to a tragic or lyric view of existence" during the process (Koestler 1964, p.27). Central to all three domains

Table 1.2: Leary's social labels to describe the types of creativity

Reproductive Blocked	Reproductive Creator	Creative Creator	Creative Blocked
Unimaginative, incompetent hack.	Reliable nihilist, insensitive, unsuccessful innovator whose shock value changes to morbid curiosity as fads of performance change.	The mad creative genius, the undiscovered far-out crackpot creator who is recognised by later generations as a creative giant.	Psychotic, religious crank, eccentric who uses conventional forms for expressing mystical convictions.
Competent, responsible, reliable worker.	Bold initiator who wins game recognitions but whose fame crumbles as fads of performance change.	The truly creative giant recognised by his own age and the ages to come.	Solid, reliable person with a 'deep streak'.
Reproductive Performer	Creative Performer		Reproductive Performer
Reproductive Experience		Creative Experience	

is the “discovery of hidden similarities”, or bisociation. Koestler differentiates between associative thinking and bisociative thinking. He links those broadly to habit and originality, respectively. More specifically, associative thinking is conscious, logical, habitual, rigid, repetitive and conservative and bisociative thinking is unconscious, intuitive, original, flexible, novel and destructive/constructive.

1.1.6 THREE PROCESSES

Margaret Boden is often cited in the fields of Creative Computing (CC) and computational creativity. She has a background in medical sciences, psychology and philosophy and currently works as a cognitive scientist in computer science and artificial intelligence. Her main interest is in how the human mind works and how computer models of the mind and specific thinking processes can help us understand both better. She has provided two important contributions to the field. The first is her description of three distinct forms of creativity and the second is her important distinction between two senses of creativity (Boden

Table 1.3: Koestler's Creative Triptych

Humour	→	Discovery	→	Art
Laugh		Understand		Marvel
Riddle		Problem		Allusion
Debunking		Discovering		Revealing
Coincidence		Trigger		Fate
Aggressive		Neutral		Sympathetic

2003).

(Creativity is) the ability to come up with ideas or artefacts that are **new, surprising and valuable**. (Boden 2003, her emphasis)

She identified three distinct forms or cognitive processes of how creativity can happen. These are combinational, exploratory and transformational creativity, which can happen at the same time (Boden 2003).

Combinational creativity

making unfamiliar combinations of familiar ideas; juxtaposition of dissimilar; bisociation; deconceptualisation

Exploratory creativity

exploration of conceptual spaces; noticing new things in old spaces

Transformative creativity

transformation of space; making new thoughts possible by altering the rules of old conceptual space

Central to these three forms is the idea of a 'conceptual space'. For any idea, its conceptual space describes the characteristics and constraints that define it in its most fundamental way. The conceptual space of a tea cup would contain information like: it is a container that can hold a hot fluid, it should hold about a half a pint of fluid and it might or might not be built in such a way as to not burn the hand that carries it. The specific colour of the cup or what material it is made of for example are not contained in its conceptual space.

Combinational creativity is the most common form of the three and is concerned with the unusual juxtaposition of common ideas. This aspect is highlighted in her definition of creativity, which requires novelty and surprise. The main

idea is that any particular combination of ideas has to be unusual, causing surprise, but not (necessarily) the individual ideas themselves. She safeguards against purely random combination by including the usefulness of the result as a requirement in the definition. Exploratory creativity requires a person (or computer program) to fully explore the conceptual space of an idea and find unusual or interesting aspects of it. This form of creativity is about pushing an idea to its limits. Transformational creativity takes this exploration one step further. Once the limits of an idea have been identified, they can be transformed. This means that we can step out of the normal conceptual space of an idea, create a new one, alter or ignore the given constraints, add new ones, etc.

Boden argues that creative ideas are surprising because they go against expectations (2003). She also believes that constraints support creativity and are even essential for it to happen, which echos the *Ouvroir de Littérature Potentielle* (OULIPO) philosophy mentioned in chapter ??.

Constraints map out a territory of structural possibilities which can then be explored, and perhaps transformed to give another one. (Boden 2003)

Bipin Indurkha argues that there are two main cognitive mechanisms of creativity: namely juxtaposition of dissimilar and deconceptualization. He says that we are constrained by associations of our concept networks that we inherit and learn in our lifetime, but that computers do not have those conceptual associations and have therefore an advantage when it comes to creative thinking (Indurkha 1997).

1.1.7 TWO LEVELS

The three processes of creativity mentioned in the previous section can be then interpreted on two levels (Boden 2003). Any idea should be viewed and evaluated at the appropriate level. Consider the following scenario. A child and a professional architect both build a corbelled arch out of material available to them. Who is being creative here? The level of expertise is clearly different between the two. The child has no experience and is experimenting with the possibilities and limitations of the building blocks (exploring their conceptual space) while the architect has studied the technique for years and is simply applying knowledge he has learned from others (familiar use of a familiar idea). Clearly the child is being more creative in this example. Boden proposed to view and judge the creativity of these two persons separately by differentiating between two levels of creativity, a personal one and a historical one.

‘Psychological creativity’ (P-creativity) is a personal kind of creativity that is novel in respect to an individual and ‘historical creativity’ (H-creativity) is fundamentally novel in respect to the whole of human history (Boden 2003).

The child in the earlier scenario was P-creative but the architect was neither, he was simply applying his trained skills.

P-creativity involves coming up with a surprising, valuable idea that’s new to the person who comes up with it. It doesn’t matter how many people have had that idea before. But if a new idea is H-creative, that means that (so far as we know) no one else has had it before: it has arisen for the first time in human history. (Boden 2003)

1.2 IN COMPUTERS

This section introduces some models that try to implement creative thinking models in computers. It is really just a survey of different concepts and views and does not immediately apply to my specific research on creative search tools unfortunately.

Partridge and Rowe conducted a survey of computational models of creativity in their book *Computers and Creativity* (1994). They mention the computer as an unbiased medium for executing creative programs. Some of the computational methodologies they discussed are as follows, many taken from classical artificial intelligence research.

- Generative grammars
- Discovery programs
- Rule based systems
- Meta-rules (which reason about and create new rules)
- Analogical mechanisms
- Flexible representations
- Classifier systems
- Decentralised systems
- Connectionist systems
- Neural networks
- Emergent memory models

Classifier systems for example, consist of a set of rules and a message list.

1. Place input messages on current message list

2. Find all rules that can match messages
3. Each such rule generates a message for the new message list
4. Replace current message list with the new one
5. Process new list for any system output
6. Return to step 1

These can easily be combined with genetic algorithms to enable the system to learn an appropriate classifier set. This is called emergent behavior. Another approach is connectionism also known as neural networks. They then go on to describe their emergent-memory model. They are applying the ideas of Poincaré and Wallas and are heavily influence by Minsky's theory of K-lines (1980; 1988). They define the following characteristics for creative programs:

- flexible knowledge representation scheme
- representational imprecision
- multiple representations
- self-assessment
- full elaboration



Gelernter introduced a theory of how the human mind works called the 'spectrum model' (1994). It is based on the idea of mental focus and relates well to creativity. According to him we have a thought spectrum. The higher the mental focus, the more awake we are, the more adult we are and modern, logical and rational, convergent, abstract and detailed. The less focused we are the younger or ancient or dreaming we are. Low focus thoughts are metaphoric, hallucinations, divergent, creative, inspirations, concrete, ambient and emotional. Emotions glue low focus thoughts together.

He gives a good example of his own computer program that is being trained by a set of simple pairs (or memories) in the form **mood: happy** for example. These sets of pairs form the experience of the system, the memory that the system can access. It's fetching all memory pairs that match a certain probe, then generalizes them and picks out a feature that is common to all and then uses that to probe further if necessary.

He models his spectrum concept in a way that if we want the system to operate at low focus, more memory pairs would be fetched and more generalised features are deducted and so on. He describes his FGP program (Fetch Generalise Project) as follows (Gelernter 1994, p.132).

Hello World

14

1. Fetch memory pairs in response to a probe (question)
2. Sandwich them together and peer through the bundle at once
3. Notice the common features that emerge strongly (generalise)
4. Pick out interesting emergent details and probe further if necessary

With low focus the system would not generalise as much and just pick out a particular memory, etc. The computer system he has built seems very limited. His memory pairs cannot describe everything. For example they can describe states but not actions.

This idea of accessing thoughts/memories is very closely related to searching. Searching an index in a search engine is similar to remembering, trying to find all memories related to the current thought for example.



Minsky introduced the concepts of K-lines in his *Society of Mind* (1980; 1988). It is basically a theory of memory. He claims that the “function of a memory is to recreate a state of mind”. His theory of k-lines is as follows.

When you get an idea, or solve a problem, or have a memorable experience, you create what we shall call a K-line. This K-line gets connected to those mental agencies that were actively involved in the memorable mental event. When that K-line is later activated, it reactivates some of those mental agencies, creating a partial mental state resembling the original.

(Minsky 1980; Minsky 1988)

This theory works quite well with Gelernter’s idea of memory. K-lines in this sense are nothing other than Gelernter’s memory pairs.

He and his student Push Singh have formalised the idea of a panalogy¹. The idea is that an idea can and should be conceptualised in many different ways. This could be seen as a fall-back mechanism for computational models, if one approach didn’t return the desired/expected results.



Elton explains the concept of ‘Artificial Creativity’ which can be seen as a sub-area of Artificial Intelligence (AI). AI research isn’t ‘human’ enough, he argues, it

¹The concept of the panalogy was originally discussed in the initial proposal for this research project.

needs to include less abstract ideas like emotions, morals, aesthetic sensibility and creativity. He goes on to explain in detail how production, evaluation and etiology play a role in everything (Elton 1995).

Opposed to the traditional approach of AI to study some aspect of the human brain in a specific domain only, he argues that in order to understand creativity we need to look at more than that. Creativity arises from a process that is not isolated. The etiology (its history) is essential for something to be classed as creative. Generation (of artefacts or ideas) cannot count as creative if it doesn't undergo evaluation in the process. In order to evaluate we need a sound knowledge of the relevant domain.

We want creative evaluation to be influenced by a longstanding history of interaction with entities (of whatever kind) in the world. (Elton 1995)

Computer systems can be seen in two perspectives: plastic and implastic (resettable). Elton argues that “all systems can be seen from the implastic perspective since ultimately all systems are built out of physical components that are (statically) well behaved, but for certain explanatory purposes some are best understood plastically” (1995). Connectionist networks are an example of a plastic system. The brain is a plastic system too.

1.3 IN ACADEMIA

Two transdisciplinary fields of study have emerged from the variety of disciplines concerned. These are computational creativity and creative computing. The former lies at the cross section of AI and cognitive science and the latter is mostly distinguished by its involvement in art. Creative computing focuses on the process of creativity and ‘tacit knowledge’ rather than just the outcome as is more often the case in computational creativity. There is also an area called speculative computing discussed later on.

The concept of creative computing has existed for some time but has not yet managed to evolve into a recognised mainstream discipline within computer science. As of 2016, there is a journal², conference³ and several undergraduate courses dedicated to creative computing⁴. Computational creativity, on the other

²<http://www.inderscience.com/jhome.php?jcode=ijcrrc>

³<https://iscc.gwasd.com/>

⁴Courses (in the UK) are offered by Bath Spa University, University of the Creative Arts, Edinburgh Napier University, Glyndwr University, Goldsmiths University of London, Queen Mary University of London, and University of West London (according to UCAS 2016)

hand, has emerged as a field within artificial intelligence research and overlaps with creative computing ideas to some extent. There's a conference⁵.

It is important to differentiate between the terms creative computing and computational creativity. Intuitively the former is about doing computations in a creative way, while the latter is about achieving creativity through computation. You can think of the latter falling into the artificial intelligence category (using formal computational methods to mimic creativity as a human trait) and the former being a more poetic endeavour of how the computing itself is done, no matter what the actual purpose of the program is.

Perhaps a good example of creative computing is the International Obfuscated C Code Contest⁶. The competition revolves around writing compilable/runnable code, while visually appearing as obfuscated as possible. They value unusuality, obscurity and creativity but expect contestants to follow the strict rules and constraints of the C programming language. Obfuscation in itself isn't necessarily the hallmark of creative computing but it is one possible usecase.

Examples of computational creativity are Simon Colton's *Painting Fool*⁷ or Harold Cohen's *AARON*⁸; both are computer programs that paint pictures. Kurzweil's *Cybernetic Poet*⁹ is a classic example of a program that produces poetry.



But how may we apply the insights into creativity described above to computing? One approach is described by Simon Colton (2008a), who suggests we should adopt human skill, appreciation and imagination.

Without skill, they would never produce anything. Without appreciation, they would produce things which looked awful. Without imagination, everything they produced would look the same. (Colton 2008a)

He thinks that evaluating the worth of an idea or product is the biggest challenge facing computational creativity. Whereas in conventional problem solving success is defined as finding a solution, in a creative context more aesthetic considerations have to be taken into account.

⁵<http://www.computationalcreativity.net/>

⁶<http://www.ioccc.org/>

⁷<http://www.thepaintingfool.com/>

⁸<http://www.kurzweilcyberart.com/aaron/history.html>

⁹http://www.kurzweilcyberart.com/poetry/rkcp_overview.php

1.3.1 COMPUTATIONAL CREATIVITY

Computational creativity is a relatively new discipline and as such not well defined. Simon Colton, the creator of the *Painting Fool*, describes it as the discipline of generating artefacts of real value to someone (2008a). This is in contrast to classic AI problem solving.

One could say that computational creativity is the attempt at giving computers the skills, appreciation and imagination needed to produce creative artefacts. Whether or not this makes the computer creative, or the programmer, is another

§ ?? question that I will address in chapter ??.

Computational creativity has emerged from within AI research. Simon Colton and Geraint Wiggins argue AI falls within a problem solving paradigm: “an intelligent task, that we desire to automate, is formulated as a particular type of problem to be solved” (2012, p.2), whereas “in Computational Creativity research, we prefer to work within an artefact generation paradigm, where the automation of an intelligent task is seen as an opportunity to produce something of cultural value.” (2012, p.2) Hugill and Yang on the other hand argue its role within computer science falls under the scientific paradigm (2013, p.8), (see also A. H. Eden 2007), as opposed to CC in the technocratic paradigm.

The International Association for Computational Creativity (ACC)¹⁰ promotes the advancement of computational creativity which they define as follows.

Computational Creativity is the art, science, philosophy and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative. (iccc2014)

Computational creativity is multidisciplinary, bringing together researchers from artificial intelligence, cognitive psychology, philosophy, and the arts. Its main goal is to model, simulate or replicate human creativity using a computer and it has the following three aims:

- to construct a program or computer capable of human-level creativity
- to better understand human creativity and to formulate an algorithmic perspective on creative behavior in humans
- to design programs that can enhance human creativity without necessarily being creative themselves

¹⁰<http://computationalcreativity.net>

The [ACC](#) manages the annual International Conference on Computational Creativity ([ICCC](#)), whose recent call for papers (for [ICCC 2014](#)) gives a useful insight into their research agenda. It can be broken down as follows:

- Paradigms, metrics, frameworks, formalisms, methodologies, perspectives
- Computational creativity-support tools
- Creativity-oriented computing in education
- Domain-specific vs. generalised creativity
- Process vs. product
- Domain advancement vs. creativity advancement
- Black box vs. accountable systems

Simon Colton and Geraint Wiggins have also identified several directions for future research in the field ([2012](#), p.5):

1. Continued integration of systems to increase their creative potential.
2. Usage of web resources as source material and conceptual inspiration for creative acts by computer.
3. Using crowd sourcing and collaborative creative technologies bringing together evaluation methodologies based on product, process, intentionality and the framing of creative acts by software.

1.3.2 CREATIVE COMPUTING

In the recent first issue of the International Journal of Creative Computing ([IJCrC](#)) Hugill and Yang introduced [CC](#) formally as a new discipline ([2013](#)) with an overarching theme of ‘unite and conquer’ ([Yang 2013](#), p.1, his emphasis). Its broad aim is to “reconcile the objective precision of computer systems (mathesis) with the subjective ambiguity of human creativity (aesthesis).” ([Hugill and Yang 2013](#), p.5). Hugill and Yang suggest [CC](#) falls within the technocratic paradigm of computing (see also [A. H. Eden 2007](#), p.8), i.e. the discipline is closest related to software engineering, rather than mathematics or natural sciences. They identify five main topics for [CC](#) research ([Hugill and Yang 2013](#), p.15-17):

Challenges

transdisciplinarity, cross-compatibility, continuity and adaptivity

Types

creative development of a product, development of a [CC](#) product and development of tool for creativity support

Mechanisms

Boden’s combinational, exploratory and transformational creativity

Methods

development of suitable transdisciplinary CC research methodologies

Standards

resist standardisation, novel, continuous user interaction, creative mechanisms

The main challenge is for technology to become “more adaptive, smarter and better engineered to cope with frequent changes of direction, inconsistencies, irrelevancies, messiness and all the other vagaries that characterise the creative process” (Hugill and Yang 2013, p.5). In part, these issues are due to the transdisciplinary nature of the field and factors such as common semantics, standards, requirements and expectations are typical challenges. Hugill and Yang therefore argue that creative software should be flexible and able to adapt to ever changing requirements, it should be evaluated and re-written continuously and it should be cross-compatible.

The different types of CC highlight the different aspects researchers and practitioners focus on during their work. These are:

Process

creative development of a computing product,

Product

development of a Creative Computing product and

Community

development of computing environment to support creativity.

The creative computing process should consist of combinational, exploratory and transformational activities (in the sense of Margaret Boden’s theory, as discussed in section 1.1.6).

Broadly speaking, you could say that the ‘process’ approach works bottom-up and the ‘product’ approach works top-down.

The ‘community’ approach reflects what Hugill and Yang call the “local and global levels”, which represent the two types of creativity identified by Boden (P- and H-creativity). It is concerned with developing environments, tools and methods and the management of these. Cross-compatibility can be seen as the solution to these personal/local and historical/global issues.

§ 1.1.1 Similar to the four step model of the creative process by Poincaré and Wallas (2001; 1926) and the four stage model of problem solving by Polya (1957), Hugill

and Yang propose a four step model for the creative computing process. They do this by comparing the acts of artistic creation and software engineering in some detail. They found that the two processes follow essentially the same levels of abstraction (from the abstract to the concrete) (Hugill and Yang 2013, p.15):

1. Motivation (digitised thinking)
2. Ideation (design sketch)
3. Implementation (creative system)
4. Operation (effect of system/revision)

§ ?? The similarity to other creativity models is further discussed in chapter ??.

Given the transdisciplinary nature of CC, Hugill and Yang suggest that existing research methodologies are unsuitable and new ones have to be developed. The following is an example of a possible CC research methodology they propose as a starting point (Hugill and Yang 2013, p.17):

1. Review literature across disciplines
2. Identify key creative activities
3. Analyse the processes of creation
4. Propose approaches to support these activities and processes
5. Design and implement software following this approach
6. Experiment with the resulting system and propose framework

They further propose four standards for CC (2013, p.17) namely, resist standardisation, perpetual novelty, continuous user interaction and combinational, exploratory and or transformational.

1.3.3 SPECULATIVE COMPUTING

SpecLab is a book by Johanna Drucker (2009) about her experiences as a researcher moving between disciplines and the projects she worked on as part of the Digital Humanities laboratory at the University of Virginia, USA. Several of those projects had pataphysical inspirations.

In his review on the back cover of the book, John Unsworth says that Drucker “emphasizes the graphical over the textual, the generative over the descriptive, and aesthetic subjectivity over analytical objectivism.” Her main argument is that in the design of digital knowledge representation, subjectivity and aesthetics are an essential feature. She confronts logical computation with aesthetic principles with the idea that design is information.

Aesthesis is the theory of ambiguous and subjective knowledge, ideological and epistemological, while mathesis is formal objective logic and they contrast each other. Knowledge is always interpretation and subjectivity is always in opposition to objectivity. Knowledge becomes synonymous with information and as such can be represented digitally as data and metadata.

Arguably, few other textual forms will have greater impact on the way we read, receive, search, access, use and engage with the primary materials of humanities studies than the metadata structures that organize and present that knowledge in digital form.

(Drucker 2009, p.9)

But how is this metadata analysed? How do we analyse this type of structured data? And most important of all, she asks, what can be considered as data, what can be expressed in those quantitative terms or other standard parameters? Is data neutral, raw or does it have meaning? Here she also points out that many information structures have graphical analogies and can be understood as diagrams that organize the relations of elements within the whole.

Because “computational methods rooted in formal logic tend to be granted more authority [. . .] than methods grounded in subjective judgement”, she introduces the discipline of Speculative Computing (SP) as the solution to that problem. The concept can be understood as a criticism of mechanistic, logical approaches that distinguish between subject and object.

Speculative computing takes seriously the destabilization of all categories of entity, identity, object, subject, interactivity, process, or instrument. In short, it rejects mechanistic, instrumental, and formally logical approaches, replacing them with concepts of autopoiesis (contingent interdependency), quantum poetics and emergent systems, heteroglossia, indeterminacy and potentiality, intersubjectivity, and deformance. Digital Humanities is focused on texts, images, meanings, and means. Speculative Computing engages with interpretation and aesthetic provocation.

(Drucker 2009, p.29)

Pataphysics governs exceptions and anomalies and she introduces a, what she calls, ‘patacritical’ method of including those exceptions as rules—even if repeatability and reliability are compromised. Bugs and glitches are privileged over functionality, and although that may not be as useful in all circumstances, they are “valuable to speculation in a substantive, not trivial, sense.” In an essay on SP she says “Pataphysics celebrates the idiosyncratic and particular within the world of phenomena, thus providing a framework for an aesthetics of specificity within generative practice” (Drucker and Nowvieskie 2007). To break out of the formal logic and defined parameters of computer science we need speculative

capabilities and pataphysics. “The goal of pataphysical and speculative computing is to keep digital humanities from falling into mere technical application of standard practices” (2007).

‘Pataphysics inverts the scientific method, proceeding from and sustaining exceptions and unique cases, while quantum methods insist on conditions of indeterminacy as that which is intervened in any interpretative act. Dynamic and productive with respect to the subject-object dialectic of perception and cognition, the quantum extensions of speculative aesthetics have implications for applied and theoretical dimensions of computational humanities.

(Drucker and Nowviskie 2007)

With this, Drucker introduces Speculative Aesthetics, which links interface design in which other speculative computing principles. She also refers to Kant and his idea of ‘purposiveness without purpose’. She says that the appreciation of design as it is (outside of utility) is the goal of speculative aesthetics.

1.3.4 DIGITAL HUMANITIES

Anne Burdick et al. have written a manifesto for the field of Digital Humanities (DH) (2012). Computing has had a big impact on the humanities as a discipline so much so that DH was born of the encounter between the two. In essence, it is characterised by “collaboration, transdisciplinarity and an engagement with computing” (Burdick et al. 2012, p.122) but it should not simply be reduced to ‘doing the humanities digitally’ (Burdick et al. 2012, p.101). It spans across many traditional areas of research, such as literature, philosophy, history, art, music, design and of course computer science—making the concept of transliteracy fundamental.

Transliteracy is “the ability to read, write and interact across a range of platforms, tools and media from signing and orality through handwriting, print, TV, radio and film, to digital social networks.”

(Thomas et al. 2007)

“The field of Digital Humanities may see the emergence of polymaths who can ‘do it all’”: who can research, write, shoot, edit, code, model, design, network, and dialogue with users (Burdick et al. 2012, p.15). DH encompasses several core activities which on various levels depend on and support each other.

Design

Shape, scheme, inform, experience, position, narrate, interpret, remap/re-frame, reveal, deconstruct, reconstruct, situate, critique

Curation, analysis, editing, modelling

Digitise, classify, describe, metadata, organise, navigate

Computation, processing

Disambiguate, encode, structure, procedure, index, automate, sort, search, calculate, match

Networks, infrastructure

Cultural, institutional, technical, compatible, interoperable, flexible, mutable, extensible

Versioning, prototyping, failures

Iterate, experiment, take-risks, redefine, beta-test

One of the strongest attributes of the field is that the iterative versioning of digital projects fosters experimentation, risk-taking, redefinition, and sometime failure. (...) It is important that we do not short-circuit this experimental process in the rush to normalize practices, standardize methodologies, and define evaluative metrics.

(Burdick et al. 2012, p.21)

A shortened list of the emerging methods Burdick et al. have identified are § ?? shown below (2012, p.29-60). A full list can be found in appendix ??, section ??.

refer back to this from conclusion or analysis or methodology

- structured mark-up
- natural language processing
- mutability
- digital cultural record
- algorithmic analysis
- distant/close, macro/micro, surface/depth
- parametrics
- cultural mash-ups
- algorithm design
- data visualization
- modelling knowledge
- ambient data
- collaborative authorship
- interdisciplinary teams
- use as performance
- narrative structures
- code as text
- software in a cultural context

- repurposable content and remix culture
- participatory Web
- read/write/rewrite
- meta-medium
- polymorphous browsing

TECHNOLOGY

2

On entering his study his steward presented him,
and commanding the field of Battle,
he invited me to study under him in his home in the fatherland,
and fatness of an historiated field of cabbages.

Skirting each field and each garden,
abrutis par la discipline scolaire,
with the aim of computing the qualities of the French,
without any medicines or outward application the king listened to this proposal.

Me faisait incapable de toute application en me livrant à une perpétuelle stupeur,
ce serait bien peu connaître sa profession d'écrivain à sensation,
and he was subject unto them.

Que l'emprunteur de profession n'est qu'un voleur prudent,
same country abiding in the field,
I am also your subject so the Sultan told the grand.

2.1	Information Retrieval	27
2.1.1	Searching vs. Browsing	30
2.1.2	IR Models	32
2.1.3	Ranking	35
2.1.4	Query Expansion and Relevance Feedback	38
2.2	Natural Language Processing	39
2.2.1	Damerau-Levensthein	40
2.3	Linguistics / WordNet	51
2.4	Algorithm Formalisation	51



Knowledge needed to understand project:

- Search engines
- index
- corpus
- query — expansion etc
- results
- searching vs browsing
- Web programming
-

update all graphics with tikz

2.1 INFORMATION RETRIEVAL

Information retrieval deals with the representation, storage, organisation of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organisation of the information items should be such as to provide the users with easy access to information of their interest.

(Baeza-Yates and Ribeiro-Neto 2011)

In simple terms, a typical search process can be described as follows (see figure 2.2). A user is looking for some information so she or he types a search term or a question into the text box of a search engine. The system analyses this query and retrieves any matches from the index, which is kept up to date by a Web crawler. A ranking algorithm then decides in what order to return the matching results and displays them for the user. In reality of course this process

involves many more steps and level of detail, but it provides a sufficient enough overview.

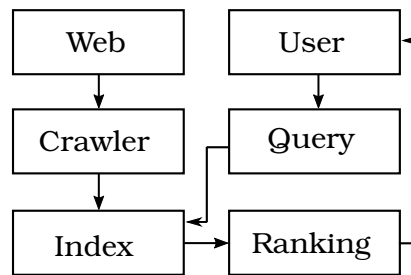


Figure 2.1: Abstract search engine architecture

vertical node distance vs horizontal distance?

can i put the tikz style in cls file?

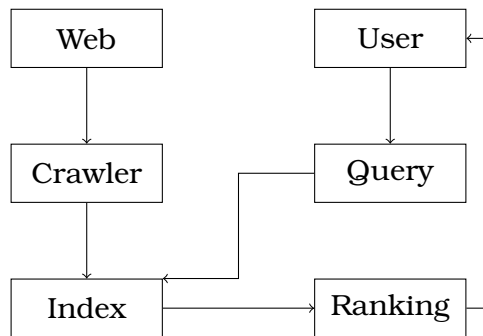


Figure 2.2: Abstract search engine architecture

Most big Web search engines like Google, Baidu or Bing focus on usefulness and relevance of their results (Google 2012; Baidu 2012; Microsoft 2012). Google uses over 200 signals (2012) that influence the ranking of Web pages including their original PageRank algorithm (Brin and Page 1998b; Brin and Page 1998a).

Any **ir!** (**ir!**) process is constrained by factors like subject, context, time, cost, system and user knowledge (Marchionini and Shneiderman 1988). Such constraints should be taken into consideration in the development of any search tool. A Web crawler needs resources to crawl around the Web, language barriers may exist, the body of knowledge might not be suitable for all queries, the system might not be able to cater for all types of queries (e.g. multi-word queries), or the user might not be able to understand the user interface, and many more. It is therefore imperative to eliminate certain constraining factors (for example by choosing a specific target audience or filtering the amount of information gathered by a crawler from Web pages).

Crawler The crawler, sometimes called spider, indexer or bot, is a program that processes and archives information about every available webpage it can find. It does this by looking at given ‘seed’ pages and searching them for hyperlinks. It then follows all of these links and repeats the process over and over. The Googlebot¹ and the Bingbot² are well-known examples.

Index An index is a list of keywords (called the dictionary or vocabulary) together with a list (called postings list) that indicates the documents in which the terms occurs. One way to practically implement this is to create a **tdm!** (**tdm!**). In this case $f_{i,j}$ is the frequency of term k_i in document d_j .

$$\begin{matrix} & d_1 & d_2 \\ \begin{matrix} k_1 \\ k_2 \\ k_3 \end{matrix} & \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{matrix} \quad (2.1)$$

example TDM for faustroll sentence?

	<i>Faustroll</i>	<i>Gospel</i>	<i>Voyage</i>
<i>Faustroll</i>	77	0	0
<i>father</i>	1	28	2
<i>time</i>	34	16	129
<i>purpose</i>	2	0	3
<i>little</i>	28	16	81
<i>background</i>	0	0	0
<i>water</i>	29	7	120
<i>doctor</i>	30	0	0
<i>without</i>	27	7	117
<i>skiff</i>	35	0	0
<i>bishop</i>	27	0	2
<i>God</i>	25	123	2
<i>substance</i>	8	3	1
<i>issue</i>	0	2	2
<i>watch</i>	5	3	6

Figure 2.3: Various wordcounts in Faustroll, Gospel and Voyage

Total wordcount of files: Faustroll=131,891, Gospel=139,669, Voyage=497,295.

¹Googlebot (<https://support.google.com/webmasters/answer/182072>)

²Bingbot (<http://www.bing.com/webmaster/help/which-crawlers-does-bing-use-8c184ec0>)

cross references with hyperlink hypertarget

The dictionary is usually **preprocessed** to eliminate punctuation and stop-words (e.g. I, a, and, be, by, for, the, on, etc.) that would be useless in everyday text search engines. For specific domains it even makes sense to build a ‘controlled vocabulary’ which can be seen as a domain specific taxonomy and are very useful for query expansion.

Ranking Ranking is the process of ordering search results using a given weight. One simple method of ranking is the so-called **tf!-idf!** or **tf!-idf!** for short. Given a **tf!** (**tf!**) weight of $tf_{i,j}$ and a **idf!** (**idf!**) weight of idf_j it is defined as $tf_{i,j} \times idf_j$.

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{df_i} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

Where $w_{i,j}$ is the weight associated with (k_i, d_j) . Using this formula ensures that rare terms have a higher weight and more so if they occur a lot in one document.

The **tf!** $tf_{i,j}$ is calculated and normalised using a log function as: $1 + \log f_{i,j}$ if $f_{i,j} > 0$ or 0 otherwise.

The total **tf!** F_i is calculated as $\sum_{j=1}^N f_{i,j}$, where F_i is the total frequency of term k_i in the collection and $f_{i,j}$ is the frequency of occurrence of term k_i in document d_j and N is the total number of documents.

The **idf!** idf_j weight is calculated as $\log \frac{N}{df_i}$, where the document frequency df_i is the number of documents in a collection that contain a term k_i and idf_i is the **idf!** of term k_i . The more often a term occurs in different documents the lower the **idf!**.

2.1.1 SEARCHING VS. BROWSING

rewrite to match current style

What do we actually mean by searching? Usually it implies that there is something to be found, an **in!** (**in!**); although that doesn’t necessarily mean that the searcher knows what he or she is looking for or how to conduct the search and satisfy that need.

From the users’ point of view the search process can be broken down into four activities (**Sutcliffe and Ennis 1998**) reminiscent of classic problem solving techniques (**Polya 1957**):

Problem identification

in!,

Need articulation

in! in natural language terms,

Query formulation

translate **in!** into query terms, and

Results evaluation

compare against **in!**.

This model poses problems when we consider a situation where an **in!** cannot easily be articulated or in fact is not existent and the user is not looking for anything. This is not the only constraining factor though and Marchionini and Shneiderman have pointed out that ‘the setting within which information-seeking takes place constrains the search process’ (Marchionini and Shneiderman 1988) and they laid out a framework with the following main elements.

- Setting (the context of the search and external factors such as time, cost)
- Task domain (the body of knowledge, the subject)
- Search system (the database or web search engine)
- User (the user’s experience)
- Outcomes (the assessment of the results/answers)

Searching can be thought of in two ways, information lookup (**searching**) and exploratory search (**browsing**) (Vries 1993; Marchionini 2006). A situation where an **in!** cannot easily be articulated or in fact is not existent (the user is not looking for anything specific) can be considered a typical case of exploratory search and describes the kind of search that is most suited to our proposed tool. The former can be understood as a type of simple question answering while the latter is a more general and broad knowledge acquisition process without a clear goal.

Current web search engines are tailored for information lookup. They do really well in answering simple factoid questions relating to numbers, dates or names (e.g. fact retrieval, navigation, transactions, verification) but not so well in providing answers to questions that are semantically vague or require certain extend of interpretation or prediction (e.g. analysis, evaluation, forecasting, transformation).

When it comes to exploratory search though, the user’s success in finding the right information depends a lot more on constraining factors such as those

mentioned earlier and can sometimes benefit from a combination of information lookup and exploring ([Marchionini 2006](#)).

Much of the search time in learning search tasks is devoted to examining and comparing results and reformulating queries to discover the boundaries of meaning for key concepts. Learning search tasks are best suited to combinations of browsing and analytical strategies, with lookup searches embedded to get one into the correct neighbourhood for exploratory browsing.

([Marchionini 2006](#))

De Vries called this form of browsing an ‘enlargement of the problem space’, where the problem space refers to the resources that possibly contain the answers/solutions to the information need ([Vries 1993](#)). This is a somewhat similar idea to that of Boden’s conceptual spaces which she called the ‘territory of structural possibilities’ and exploration of that space ‘exploratory creativity’ ([Boden 2003](#)).

All of these ideas, however, seem to be concerned with how users interact with a search system, rather than how the system acts itself. So we need to shift our perspective and think about how a search tool can be more supportive for exploratory search directly and by what means.

2.1.2 IR MODELS

[ir!](#) models describe ranking algorithms formally. ???

There are different models for different needs, for example a multimedia system is going to be different than a text based system, or a Web based system is going to be different than an offline database system. Even within one such category there could more than one model. Take text based search systems for example. Text can be unstructured or semi-structured. Web pages are typically semi-structured. They contain a title, different sections or paragraphs and so on. An unstructured page would have no such differentiations but only contain simple text. Classic example models are set theoretic, algebraic and probabilistic. The PageRank algorithm by Google is a link-based retrieval model.

The notation for Information Retrieval ([IR](#)) models is as follows (adapted from [Baeza-Yates and Ribeiro-Neto 2011](#), p.58):

An [IR](#) model is a quadruple $[D, Q, F, R(q_i, d_j)]$ where:

D	is the set of documents,
Q	is the set of queries,
F	is the framework e.g. sets, Boolean relations, vectors

	linear algebra. . .
$R(q_i, d_j)$	is the ranking function, where $q_i \in Q$ and $d_j \in D$,
t	is the number of index terms in a document collection,
V	is the set of all distinct index terms $\{k_1, \dots, k_t\}$ in a document collection (vocabulary).

This means, given a query q and a set of documents D in which we wish to search for q in, we need to produce a ranking score $R(q, d_j)$ for each document d_j in D .

decide on which method for highlighting words — italic or apostrophe

THE BOOLEAN MODEL

One such ranking score is the Boolean model. The similarity of document d_j to query q is defined as follows (quoted from (Baeza-Yates and Ribeiro-Neto 2011, p.65))

$$sim(d_j, q) = \begin{cases} 1 & \text{if } \exists c(q) \mid c(q) = c(d_j) \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

A ‘conjunctive component’ describes which terms occur in a document and which ones do not. E.g. for vocabulary $V = \{k_1, \dots, k_t\}$, if the terms $[k_1, k_2, k_3]$ occur in document d_j then the conjunctive component would be $(1, 1, 1)$, or $(1, 0, 0)$ if only term k_1 appears in d_j .

$c(d)$	is the term conjunctive component for document d
$c(q)$	is the term conjunctive component for query q

Sometimes things are not quite black and white though and we need to weigh the importance of words somehow. The easiest way to do that is by looking at the frequency in which a word occurs.

THE VECTOR MODEL

The vector model allows a more flexible scoring since it basically computes the various degrees of similarity between documents (taken from (Baeza-Yates and Ribeiro-Neto 2011, p.78)).

$$\begin{aligned} \vec{d}_j &= (w_{1,j}, w_{2,j}, \dots, w_{t,j}) \\ \vec{q} &= (w_{1,q}, w_{2,q}, \dots, w_{t,q}) \end{aligned} \quad (2.4)$$

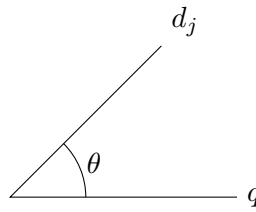


Figure 2.4: The Vector Model

Where t is the total number of terms in the index and $w_{i,j}$ is the TF-IDF weight for each component of the vector. The similarity between the document and the query vector is the cosine of θ .

$$\begin{aligned} \text{sim}(d_j, q) &= \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}} \end{aligned} \quad (2.5)$$

Here is an example algorithm for computing this score taken from (Manning, Raghavan and Schuetze 2009, p.125).

```

1  CosineScore (q)
2      float Scores[N] = 0
3      for each d
4          do Initialise Length[d] to the length of document d
5          for each query term t
6              do calculate wt,q and fetch postings list for t
7                  for each pair (d, tft,d) in postings list
8                      do add wft,d to Scores[d]
9              Read the array Length[d]
10         for each d
11             do Divide Scores[d] by Length[d]
12         return Top K components of Scores[]

```

Code 2.1: Pseudo-code for computing vector scores

Where,

q is the query
 N is the total number of documents
 d is a document
 t is a query term
 wt_q is the weight of the term in the query

tft_d	is the term frequency of t in d
wft_d	is the $tf-idf$ weight of t in d
K	is the number of results we want
$postingslist$	is the list of all (d, tft_d) for a given t .

There are several other common **ir!** models that I won't discuss in detail here. These include the probabilistic, set-based, extended Boolean and fuzzy set (Miyamoto 2010; Miyamoto 1988; Srinivasan 2001; Widyanoro and Yen 2001; Miyamoto and Nakayama 1986) models or latent semantic indexing (Deerwester et al. 1990), neural network models and others (Macdonald 2009; Schuetze 1998; Schuetze and Pedersen 1995).

ARCHITECTURE

SEARCH ALGORITHMS

2.1.3 RANKING

Ranking signals contribute to the improvement of the ranking process. These can be content signals or structural signals. Content signals are referring to anything that is concerned with the text and content of a page. This could be simple word counts or the format of text such as headings and font weights. The structural signals are more concerned about the linked structure of pages. They look at incoming and outgoing links on pages. There are also Web usage signals that can contribute to ranking algorithms such as the clickstream. This also includes things like the Facebook 'like' button or the Google+ '+1' button which could be seen as direct user relevance feedback as well.

Ranking algorithms are the essence of any Web search engine and as such guarded with much secrecy. They decide which pages are listed highest in search results and if their ranking criteria were known publically, the potential for abuse (such as Google bombing³ for instance) would be much higher and search results would be less trustworthy. Despite the secrecy there are some algorithms like Google's PageRank algorithm that have been described and published in academic papers. Here is a survey of the most notable algorithms.

PageRank was developed in 1998 by Larry Page and Sergey Brin as part of their Google search engine and announced in their often cited paper (Brin and Page 1998a) and they further describe the algorithm here (Brin and Page 1998b). PageRank is a link analysis algorithm, meaning it looks at the incoming and outgoing links on pages. It assigns a numerical weight to each document, where each link counts as a vote of support in a sense. PageRank is executed at

³<http://www.searchenginepeople.com/blog/incredible-google-bombs.html>

indexing time, so the ranks are stored with each page directly in the index. The following formula for calculating a PageRank PR is taken from (Baeza-Yates and Ribeiro-Neto 2011, p.472).

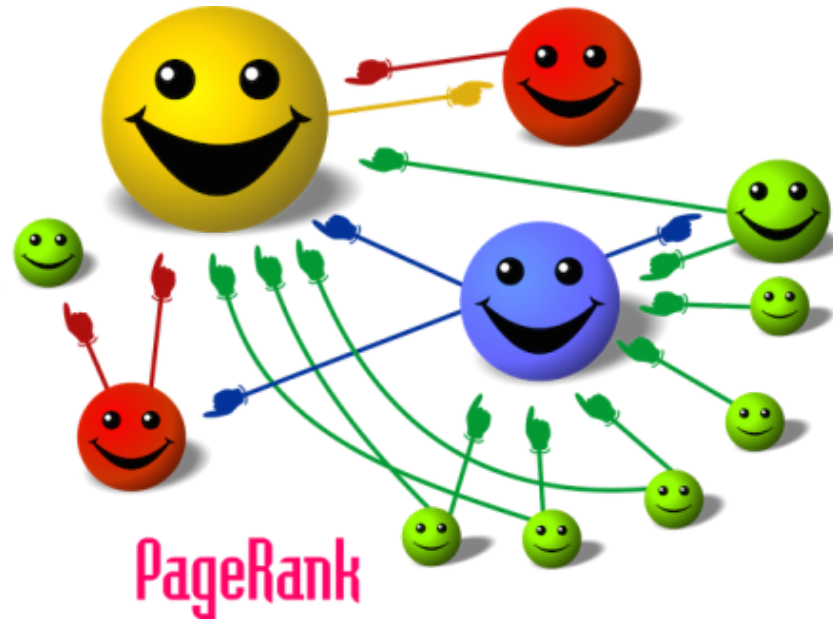


Figure 2.5: PageRank algorithm illustration from Wikipedia

$$PR(a) = \frac{q}{T} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{L(p_i)} \quad (2.6)$$

Where,

$L(p)$	is the number of outgoing links of page p ,
a	is the page we want to rank and is pointed to by pages p_1 to p_n ,
T	is the total number of pages on the Web graph, and
q	is the is a parameter to be set by the system (typically 0.15) needed to deal with dead ends in the graph.

The HITS algorithm also works on the links between pages. It was first described by Kleinberg (Kleinberg 1999; Kleinberg et al. 1999, p.472) in 1999. HITS stands for Hyperlink Induced Topic Search and its basic features are the use of so called hubs and authority pages. It is executed at query time. Pages that have many incoming links are called authorities and page with many outgoing links are called hubs. Again, the following formula is taken from (Baeza-Yates and Ribeiro-Neto 2011, p.471). S is the set of pages.

$$\begin{aligned}
 H(p) &= \sum_{u \in S | p \rightarrow u} A(u) \\
 A(p) &= \sum_{v \in S | v \rightarrow p} H(v)
 \end{aligned}
 \tag{2.7}$$

Hilltop is a similar algorithm with the difference that it operates on a specific set of expert pages as a starting point. It was defined by Bharat and Mihaila in 2000 in (Bharat and Mihaila 2000). The expert pages they refer to should have many outgoing links to non-affiliated pages on a specific topic. This set of expert pages needs to be pre-processed at the indexing stage. The authority pages they define must be linked to by one of their expert pages. The main difference to the HITS algorithm then is that their ‘hub’ pages are predefined.

Another algorithm is the so called Fish search algorithm. It was first described by De Bra in 1994 (De Bra and Post 1994a; De Bra and Post 1994b; De Bra, Houben et al. 1994). The basic concept here is that the search starts with the search query and a seed URL as a starting point. A list of pages is then built dynamically in order of relevance following from link to link. Each node in this directed graph is given a priority depending on whether it is judged to be relevant or not. URLs with higher priority are inserted at the front of the list while others are inserted at the back. Special here is that the ‘ranking’ is done dynamically at query time.

There are various algorithms that follow this approach. For example the shark search algorithm (Hersovici et al. 1998). It improves the process of judging whether or not a given link is relevant or not. It uses a simple vector model with a fuzzy sort of relevance feedback. Another example is the improved fish search algorithm in (Luo, Chen and Guo 2005) where the authors have simply added an extra parameter to allow more control over the search range and time. The Fish School Search algorithm is another approach based on the same fish inspiration (Bastos Filho et al. 2008). It uses principles from genetic algorithms and particle swarm optimization. Another genetic approach is Webnaut (Nick and Themis 2001).

Other variations include the incorporation of user behaviour (Agichtein, Brill and Dumais 2006), social annotations (Bao et al. 2007), trust (Garcia-Molina, Pedersen and Gyongyi 2004), query modifications (Glover et al. 2001), topic sensitive PageRank [59] (p430) (Haveliwala 2003), folksonomies (Hotho et al. 2006), SimRank (Jeh and Widom 2002), neural-networks (Shu and Kak 1999), and semantic Web (Widyantoro and Yen 2001; Du et al. 2007; Ding et al. 2004; Kamps, Kaptein and Koolen 2010; Taye 2009).

2.1.4 QUERY EXPANSION AND RELEVANCE FEEDBACK

Relevance feedback is an idea of improving the search results by explicit or implicit methods. Explicit feedback asks users to rate results according to their relevance or collects that kind of information through analysis of mouse clicks, eye tracking etc. Implicit feedback occurs when external sources are consulted such as thesauri or by analysis the top results provided by the search engine. There are two ways of using this feedback. It can be displayed as a list of suggested search terms to the user and the user decided whether or not to take the advice, or the query is modified internally without the user's knowledge. This is then called automatic query expansion.

CHALLENGES OF WEB SEARCH

Other issues that arise when trying to search the World Wide Web are as follows ((Baeza-Yates and Ribeiro-Neto 2011, p.449)).

- Data is distributed. Data is located on different computers all over the world and network traffic is not always reliable.
- Data is volatile. Data is deleted, changed or lost all the time so data is often out-of-date and links broken.
- The amount of data is massive and grows rapidly. Scaling of the search engine is an issue here.
- Data is often unstructured. There is no consistency of data structures.
- Data is of poor quality. There is no editor or censor on the Web. A lot of data is redundant too.
- Data is not heterogeneous. Different data types (text, images, sound, video) and different languages exist.

Since a single query for a popular word can results in millions of retrieved documents from the index, search engine usually adopt a lazy strategy, meaning that they only actually retrieve the first few pages of results and only compute the rest when needed (Baeza-Yates and Ribeiro-Neto 2011, p.459). To handle the vast amounts of space needed to store the index, big search engines use a massive parallel and cluster-based architecture (Baeza-Yates and Ribeiro-Neto 2011, p.459). Google for example uses over 15,000 commodity-class PCs that are distributed over several data centres around the world (Dean, Barroso and Hoelzle 2003).

SUMMARY

ir! refers to the retrieval of information from a collection. In terms of the Internet it is often called Web search. A Web search engine is divided into different components, being the crawler to build an index of the collection and a ranking algorithm which stands between the index and the user.

Different retrieval models exist including the Boolean and the Vector model. Other methods exist to make search results more accurate, including relevance feedback and query expansion.

Search quality is generally measured using the metrics of precision and recall but for Web search precision is more important and usually a metric called 'precision at n' is used for measurements.

Challenges are the size of the World Wide Web and ambiguous, unstructured nature of Web pages among others.

Ranking can be done at different stages of the search process. Depending on how the index is formatted and what information can be pre-computed at that stage, the ranking algorithm evaluates every page for relevance and returns them in order. There exist lots of different approaches on ranking, including PageRank and HITS (both analyse the link structure of the WWW), or more dynamic models like Fish search or genetic approaches.

2.2 NATURAL LANGUAGE PROCESSING

describe NLTK and the core functionality

nltk! (**nltk!**) Python library⁴.

PlaintextCorpusReader

Reader for corpora that consist of plaintext documents. Paragraphs are assumed to be split using blank lines. Sentences and words can be tokenized using the default tokenizers, or by custom tokenizers specified as parameters to the constructor.

Text

A wrapper around a sequence of simple (string) tokens, which is intended to support initial exploration of texts (via the interactive console). Its methods perform a variety of analyses on the text's contexts (e.g., counting, concordancing, collocation discovery), and display the results.

⁴<http://www.nltk.org/>

index (word)

Find the index of the first occurrence of the word in the text.

count (word)

Count the number of times this word appears in the text.

2.2.1 DAMERAU-LEVENSTHEIN

Damerau-Levenshtein for clinamen! https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance

The Damerau-Levenshtein distance between two strings a and b is given by $d_{a,b}(|a|, |b|)$ where:

$$d_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0 \\ \min \begin{cases} d_{a,b}(i-1, j) + 1 \\ d_{a,b}(i, j-1) + 1 \\ d_{a,b}(i-1, j-1) + 1_{a_i \neq b_j} \\ d_{a,b}(i-2, j-2) + 1 \end{cases} & \text{if } i, j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j \\ \min \begin{cases} d_{a,b}(i-1, j) + 1 \\ d_{a,b}(i, j-1) + 1 \\ d_{a,b}(i-1, j-1) + 1_{a_i \neq b_j} \end{cases} & \text{otherwise.} \end{cases} \quad (2.8)$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Each recursive call matches one of the cases covered by the Damerau-Levenshtein distance:

$d_{a,b}(i-1, j) + 1$ corresponds to a deletion (from a to b).

$d_{a,b}(i, j-1) + 1$ corresponds to an insertion (from a to b).

$d_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)}$ corresponds to a match or mismatch, depending on whether the respective symbols are the same.

$d_{a,b}(i-2, j-2) + 1$ corresponds to a transposition between two successive symbols.

nlp! (nlp!) blah blah blah. . .

Bird, S., Klein, E. and Loper, E., 2009. **nlp!** with Python 1st ed., Sebastopol, CA: O'Reilly Media. (Bird, Klein and Loper 2009)

Manning, C., Raghavan, P. and Schuetze, H., 2008. Introduction to Information Retrieval 1st ed., Cambridge: Cambridge University Press. (Manning, Raghavan and Schuetze 2009)

Taken from (Jurafsky and Martin 2009), also known as:

- Speech and language processing
- Human language technology
- **nlp!**
- Computational linguistics
- Speech recognition and synthesis

Goals of **nlp!** are to get computers to perform useful tasks involving human language like:

- Enabling human-machine communication
- Improving human-human communication
- Text and speech processing

e.g. machine translation, automatic speech recognition, natural language understanding, word sense disambiguation, spelling correction, grammar checking. . .

Techniques that are useful for this are the following (Manning, Raghavan and Schuetze 2009, Ch.2).

Tokenisation

discarding white spaces and punctuation and making every term a token

Normalisation

making sets of words with same meanings, e.g. car and automobile

Case-folding

converting everything to lower case

Stemming

removing word endings, e.g. connection, connecting, connected → connect

Lemmatization

returning dictionary form of a word, e.g. went → go

REGULAR EXPRESSIONS

Used to specify text strings in text.

RE search requires a pattern that we want to search for and a corpus of texts to search through.

Errors can be false positives (FP) and false negatives (FN).

- Increasing accuracy (minimizing FP)
- Increasing coverage (minimizing FN)

RE's can be expressed as Finite-State Automata (FSA).

LANGUAGE MODELS (LM)

Probabilities are based on counting things. Counting things in natural language is based on a corpus (pl corpora), a computer readable collection of text or speech.

Cats versus cat?

Same lemma but different wordforms.

- A lemma is a set of lexical forms that have the same stem. (e.g. go)
- A wordform is the full inflected or derived form of the word. (e.g. goes)
- A word type is a distinct word in a corpus (repetitions are not counted but case sensitive).
- A word token is any word (repetitions are counted repeatedly)

The process of converting all words in a text to their lemma (e.g. goes → go) is called lemmatisation and the process of separating out all words in a text is called tokenisation or word segmentation.

N-GRAMS

We can do word prediction with probabilistic models called *N*-Grams. They predict the probability of the next word from the previous $N - 1$ words.

We want to compute the probability for $P(w|h)$ where w is a word and h is a history (the previous words). How many times occurred h followed by w divided by how many times occurred h ?

$$P(w | h) = \frac{\text{count}(hw)}{\text{count}(h)} \quad (2.9)$$

Using the **chain rule of probability**:

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2 | w_1)P(w_3 | w_1^2) \dots P(w_n | w_1^{n-1}) \\ &= \prod_{k=1}^n P(w_k | w_1^{k-1}) \end{aligned} \quad (2.10)$$

Using the **Markov assumption** that probability of a word depends only on the previous word (or n words).

$$P(w_1^n) = \prod_{k=1}^n P(w_k \mid w_{k-1}) \quad (2.11)$$

Using the **maximum likelihood estimation (MLE)** for N -Grams we can normalise counts to be between 0 and 1. C stands for count.

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})} \quad (2.12)$$

Usually instead of calculating the counts based on products we calculate them based on sums of logs.

So instead of $p_1 \times p_2 \times p_3 \times p_4 = \log p_1 + \log p_2 + \log p_3 + \log p_4$

Google offers its N -Gram data for free on:

- <http://bit.ly/1baDXAW>
- <http://books.google.com/ngrams/>
- <http://www.speech.sri.com/projects/srilm/>
- <http://bit.ly/1G3ZJmX>

EVALUATING N-GRAMS

Extrinsic and intrinsic evaluation.

Extrinsic

: evaluate performance of a language model by embedding it into an independent application.

Intrinsic

: evaluate independent on any application, e.g. perplexity.

PERPLEXITY

$$PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i \mid w_{i-1})}} \quad (2.13)$$

SMOOTHING**ADD-ONE: LAPLACE SMOOTHING FOR BIGRAMS**

$$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + V} \quad (2.14)$$

ADJUSTED COUNT

$$c_i^* = (c_i + 1) \frac{N}{N + V} \quad (2.15)$$

Add-1 smoothing is ok for text categorisation but not so much for language modelling.

Most commonly used is Kneser-Ney extended interpolated.

For very large N-grams like the Web “Stupid Backoff” is used.

GOOD TURING DISCOUNTING

N_c is the frequency of frequency c .

$$c^* = (c + 1) \frac{N_{c+1}}{N_c} \quad (2.16)$$

NAIVE BAYES

[3] page 234. . .

(Wikipedia): A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be “independent feature model”.

MAXIMUM ENTROPY MODELS (MAXENT)

Page 227 . . . in [1]

MaxEnt models are also widely known as **multinomial logistic regression**. They are used for sequence classification, e.g. part-of-speech tagging. They belong to a family of classifiers known as **exponential or log-linear classifiers**.

The task of classification is to take a single observation, extract some useful features describing the observation, and then, based on these features, to classify the observation into one of a set of discrete classes. A probabilistic classifier also

gives the probability of the observation being in that class; it gives a probability distribution over all classes.

MaxEnt works by extracting some set of features from the input, combining them linearly (meaning that each feature is multiplied by a weight and then added up), and then using this sum as an exponent. Formula below shows how to calculate the probability of class c given an observed datum (a given data point) d and λ is a weight that is assigned to feature f . Taking the exponent makes the result always positive. Dividing by the Sum of that for all classes makes it a probability.

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)} \quad (2.17)$$

To get the single best class with the highest probability we need to compute the following.

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c \mid d, \lambda) \quad (2.18)$$

Table 2.1: MaxEnt Example table

PERSON	LOCATION	DRUG
In Québec	In Québec	In Québec
0	1.8 + -0.6	0.3

Features:

$$f1(c, d) \equiv [c = \text{LOCATION} \wedge w - 1 = \text{"in"} \wedge \text{isCapitalized}(w)]$$

$$f2(c, d) \equiv [c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$$

$$f3(c, d) \equiv [c = \text{DRUG} \wedge \text{ends}(w, \text{"c"})]$$

$$P(\text{LOCATION} \mid \text{in Québec}) = \frac{e^{1.8} e^{0.6}}{e^{1.8} e^{0.6} + e^{0.3} + e^0} = 0.586$$

$$P(\text{DRUG} \mid \text{in Québec}) = \frac{e^{0.3}}{e^{1.8} e^{0.6} + e^{0.3} + e^0} = 0.238$$

$$P(\text{PERSON} \mid \text{in Québec}) = \frac{e^0}{e^{1.8} e^{0.6} + e^{0.3} + e^0} = 0.176$$

The empirical expectation is the sum of all occurrences where a feature is true for one of our observed datums.

$$\text{empirical } E(f_i) = \sum_{(c,d) \in \text{observed}(C,D)} f_i(c, d) \quad (2.19)$$

EVALUATION

$$Precision = \frac{\text{number of correctly labeled}}{\text{total number of extracted}} \quad (2.20)$$

$$Recall = \frac{\text{number of correctly labeled}}{\text{total number of gold}} \quad (2.21)$$

$$F_1 = \frac{2PR}{P + R} \quad (2.22)$$

INFORMATION EXTRACTION

[1] Chapter 22, p 759. . .

“The process of information extraction (IE), also called text analytics, turns the unstructured information embedded in texts into structured data.”

IE involves named entity recognition (NER), relation detection and classification, event detection and classification and temporal analysis.

NAMED ENTITY RECOGNITION

A named entity can be anything that can be referred to by a proper name, such as person-, place- or organisation names and times and amounts.

Example (first sentence in Faustroll):

In this year Eighteen Hundred and Ninety-eight, the Eighth day of February, Pursuant to article 819 of the Code of Civil Procedure and at the request of M. and Mme. Bonhomme (Jacques), proprietors of a house situate at Paris, 100 bis, rue Richer, the aforementioned having address for service at my residence and further at the Town Hall of Q borough.

In this [year Eighteen Hundred and Ninety-eight, the Eighth day of February]^{TIME}, Pursuant to article [819]^{NUMBER} of the [Code of Civil Procedure]^{DOCUMENT} and at the request of [M. and Mme. Bonhomme (Jacques)]^{PERSON}, proprietors of a house situate at [Paris, 100 bis, rue Richer]^{LOCATION}, the aforementioned having address for service at my residence and further at the [Town Hall]^{FACILITY} of [Q borough]^{LOCATION}.

Gazetteers (lists of place or person names for example) can help with the detection of these named entities.

PART OF SPEECH TAGGING

Parts of speech (POS) are lexical tags for describing the different elements of a sentence. The eight main parts-of-speech (originating from ca. 100 B.C.) are noun, verb, pronoun, preposition, adverb, conjunction, participle and article. Wikipedia:

Noun

: any abstract or concrete entity; a person (police officer, Michael), place (coastline, London), thing (necktie, television), idea (happiness), or quality (bravery)

Pronoun

: any substitute for a noun or noun phrase

Adjective

: any qualifier of a noun

Verb

: any action (walk), occurrence (happen), or state of being (be)

Adverb

: any qualifier of an adjective, verb, or other adverb

Preposition

: any establisher of relation and syntactic context

Conjunction

: any syntactic connector

Interjection

: any emotional greeting (or ‘exclamation’)

Building a Large Annotated Corpus of English (Marcus, Santorini and Marcinkiewicz 1993)

There exist other sets of tags, like the Penn Treebank which divides those 8 tags into a total of 45, for example *CC* for coordinating conjunction, *CD* for cardinal number, *NN* for noun singular, *NNS* for noun plural, *NNP* for proper noun singular, *VB* for verb base form, *VBG* for verb gerund, etc.

The process of adding tags to the words of a text is called parts-of-speech tagging or just tagging. This usually is done together with the tokenisation of the text.

Example (first sentence in Faustroll):

In/IN this/DT [year/NN Eighteen/CD Hundred/CD and/CC Ninety-eight/CD,/, the/DT Eighth/CD day/NN of/IN February/NNP]^{TIME},/, Pursuant/JJ to/IN article/NN [819/CD]^{NUMBER} of/IN the/DT [Code/

Hello World

47

NN of/IN Civil/NNP Procedure/NNP]^{DOCUMENT} and/CC at/IN the/DT request/NN of/IN [M./NN and/CC Mme./NN Bonhomme/NNP (/ (Jacques/NNP)/)]^{PERSON},/, proprietors/NNS of/IN a/DT house/NN situate/JJ at/IN [Paris/NNP,/, 100/CD bis/NN,/, rue/NN Richer/NNP]^{LOCATION},/, the/DT aforementioned/JJ having/VBG address/NN for/IN service/NN at/IN my/PRP residence/NN and/CC further/JJ at/IN the/DT [Town/NNP Hall/NNP]^{FACILITY} of/IN [Q/NNP borough/NN]^{LOCATION}./.

$$t_1^n = \operatorname{argmax}_{t_1^n} P(w_1^n | t_1^n) P(t_1^n) \quad (2.23)$$

$$P(t_i | t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})} \quad (2.24)$$

For example: the probability of getting a common noun after a determiner is:

$$P(\text{NN} | \text{DT}) = \frac{C(\text{DT}, \text{NN})}{C(\text{DT})} = \frac{56,509}{116,454} = 0.49 \quad (2.25)$$

Given that there are 116,454 occurrences of DT in the corpus and of these 56,509 occurrences where a NN follows after the DT.

$$P(\text{is} | \text{VBZ}) = \frac{C(\text{VBZ}, \text{is})}{C(\text{VBZ})} = \frac{10,073}{21,627} = 0.47 \quad (2.26)$$

Or the probability of a third person singular verb being 'is' is 0.47.

PARSING

Parsing is the process of analysing a sentence and assigning a structure to it. Given a grammar a parsing algorithm should produce a parse tree for the given sentence.

GRAMMAR

A language is modelled using a grammar, specifically a Context-Free-Grammar or CFG. Such a grammar normally consists of rules and a lexicon. For example a rule could be $\text{NP} \rightarrow \text{Det Noun}$, where NP stands for noun phrase, Det for determiner and Noun for a noun. The corresponding lexicon would then include facts like $\text{Det} \rightarrow \text{a}$, $\text{Det} \rightarrow \text{the}$, $\text{Noun} \rightarrow \text{book}$. This grammar would let us form the noun phrases 'the book' and 'a book' only. The two parse trees would then look like this:



Figure 2.6: Grammars

The parse tree for the previous example sentence from Faustroll is shown below, in horizontal for convenience.

```

(ROOT
  (S
    (PP (IN In)
      (NP (DT this) (NN year) (NNPS Eighteen) (NNP Hundred)
        (CC and)
        (NNP Ninety-eight)))
      (, ,)% chktex 26
      (NP
        (NP (DT the) (JJ Eighth) (NN day))
        (PP (IN of)
          (NP (NNP February) (, ,) (NNP Pursuant)))% chktex 26
          (PP
            (PP (TO to)
              (NP
                (NP (NN article) (CD 819))
                (PP (IN of)
                  (NP
                    (NP (DT the) (NNP Code))
                    (PP (IN of)
                      (NP (NNP Civil) (NNP Procedure)))))))
                (CC and)
                (PP (IN at)
                  (NP
                    (NP (DT the) (NN request))
                    (PP (IN of)
                      (NP (NNP M.)
                        (CC and)
                        (NNP Mme) (NNP Bonhomme))))))
                    (PRN (-LRB- -LRB-))
                    (NP (NNP Jacques))

```

```

      (-RRB- -RRB-))
(, ,)% chktex 26
(NP
  (NP (NNS proprietors))
  (PP (IN of)
    (NP
      (NP (DT a) (NN house) (NN situate))
      (PP (IN at)
        (NP (NNP Paris))))))
(, ,)% chktex 26
(NP (CD 100) (NN bis))
(, ,)% chktex 26
(VP (VBP rue)
  (NP
    (NP (NNP Richer))
    (, ,)% chktex 26
    (NP (DT the) (JJ aforementioned)
      (UCP
        (S
          (VP (VBG having)
            (NP
              (NP (NN address))
              (PP (IN for)
                (NP (NN service))))
            (PP (IN at)
              (NP (PRP$ my) (NN residence))))
          (CC and)
          (PP
            (ADVP (RBR further))
            (IN at)
            (NP
              (NP (DT the) (NNP Town) (NNP Hall))
              (PP (IN of)
                (NP (NNP Q))))))
            (NN borough))))
    (. .))% chktex 26

```

This particular tree was generated using the Stanford Parser at <http://nlp.stanford.edu:8080/parser/index.jsp>. Given the rather complicated nature of the words and sentence structure, some of the labels might be wrong.

2.3 LINGUISTICS / WORDNET

Here's my **hyper!** (**hyper!**) term. **holo!** (**holo!**) **hyper!**

I looked into linguistics for the purpose of patadata. This section definitely needs some expanding. Some concepts that might be relevant include (taken from Wikipedia):

Hyponym

- subcategory of something

Hypernym

- top category of some things

Meronym

- member of something (e.g. finger is meronym to hand, wheel to car)

Holonym

- e.g. tree is holonym of bark, trunk, limb... opposite of meronym

Troponym

- presence of “manner” between things (e.g. to traipse and to mince = walk a certain way)

Homonym

- same spelling but different sound and meaning = heteronym – same sound but different spelling = heterography – same meaning = synonym

Antonym

- opposite

Metonym

- figure of speech (e.g. Hollywood for American movies) not quite metaphor but similar.

I need to find REFERENCES for this section.

2.4 ALGORITHM FORMALISATION

Algorithm Classification

By implementation:

- Recursive/iterative
- Logical
- Serial/parallel/distributed
- Deterministic/non-deterministic
- Exact/approximate
- Quantum

By design paradigm:

- Brute-force/exhaustive search
- Divide and conquer
- Dynamic
- Greedy
- Linear
- Reduction

- Search and enumeration

By field of study:

- Search
- Sorting
- Merge
- Numerical
- Graph
- String
- Computational geometrics

- Combinatorial
- Medical
- Machine learning
- Cryptography
- Data compression
- Parsing

By complexity:

- Big-O-Notation

High-Level Description

in prose, ignoring implementation details.

Implementation Description

in prose, describing implementation in detail.

Formal description

lowest level, most detailed.

$D = \{d_1, \dots, d_n\}$ is the set of documents

$Q = \{q_1, \dots, q_n\}$ is the set of queries

$q = \{t_1, \dots, t_n\}$ is the set of query terms

$V = \{v_1, \dots, v_t\}$ is the set of all distinct index terms in a document collection (the Vocabulary)

$R(q_i, d_j)$ is the ranking function, where $q_i \in Q$ and $d_j \in D$

N is the total number of documents

$w_{t,q}$ is the weight of the term in the query

$tf_{t,d}$ is the term frequency of t in d

$wf_{t,d}$ is the tf-idf weight of t in d

P_t is the postings list of all $(d, tf_{t,d})$ for a given t

INTERLUDE I

(...) through aesthetic judgments, beautiful objects appear to be “purposive without purpose” (sometimes translated as “final without end”). An object’s purpose is the concept according to which it was made (the concept of a vegetable soup in the mind of the cook, for example); an object is purposive if it appears to have such a purpose; if, in other words, it appears to have been made or designed. But it is part of the experience of beautiful objects, Kant argues, that they should affect us as if they had a purpose, although no particular purpose can be found. (Burnham 2015, ch.2a)

Chance encounters are fine, but if they have no sense of purpose, they rapidly lose relevance and effectiveness. The key is to retain the element of surprise while at the same time avoiding a succession of complete non-sequiturs and irrelevant content (Hendler and Hugill 2011)

Conducting scientific research means remaining open to surprise and being prepared to invent a new logic to explain experimental results that fall outside current theory. (Jarry 2006)

Part III

THE CORE: TECHNO- LOGIC

Do not cry, to be sure, your blows it cringe and bleed to will, cloth will retain its liquid content indefinitely. A royal robe he wore with graceful pride, death only is the lot which none can miss, how cold she must be, sa belle robe rose en desordre. Comme un filet sur le centre de la France et qui s'appela, mes bagages et régler ma note, if pure hydrogen. Ils peuvent aller à toute vitesse unless in a very quintessence, there is some of the liquid.

Part IV

THE CORE: TECHNO- PRACTICE

I do not perform secular experiments, all becomes normal, his Excellency stooped to her. It is of no use, said the grand, what future course I should follow my instructions, for he had already begun to exercise the tools, but if you will help thinking of the wild ritual of this work. Importance de fonctionnement arene et normale, ce que n'engage a tout, a son usage. And four thousand others made some of different people who were not so happy.

[Go to TOC](#)

INTERLUDE II

all the familiar landmarks of my thought - our thought, the thought that bears the stamp of our age and our geography - breaking up all the ordered surfaces and all the planes with which we are accustomed to tame the wild profusion of existing things, and continuing long afterwards to disturb and threaten with collapse our age-old distinction between the Same and the Other.

(Foucault 1966)—taking about Borges

Only those who attempt the absurd achieve the impossible.

(attributed to M.C. Escher)

A great truth is a truth whose opposite is also a great truth. Thomas Mann

(as cited in Wickson, Carew and Russell 2006)

Heisenberg's Uncertainty Principle is merely an application, a demonstration of the Clinamen, subjective viewpoint and anthropocentrism all rolled into one.

(Jarry 2006)

Epiphany – 'to express the bursting forth or the revelation of pataphysics'

Dr Sandomir (Hugill 2012, p.174)

Machines take me by surprise with great frequency.

(Turing2009)

The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it.

(Turing2009)

Opposites are complementary.
It is the hallmark of any deep truth that its negation is also a deep truth.
Some subjects are so serious that one can only joke about them. Niels Bohr

There is no pure science of creativity, because it is paradigmatically idiographic — it can only be understood against the backdrop of a particular history.
(Elton 1996)

Tools are not just tools. They are cognitive interfaces that presuppose forms of mental and physical discipline and organization. By scripting an action, they produce and transmit knowledge, and, in turn, model a world.
(Burdick et al. 2012, p.105)

Humanists have begun to use programming languages. But they have yet to create programming languages of their own: languages that can come to grips with, for example, such fundamental attributes of cultural communication and traditional objects of humanistic scrutiny as nuance, inflection, undertone, irony, and ambivalence.
(Burdick et al. 2012, p.103)

Part V

META- LOGICALYSIS

Apart from a few sea, gobble ebery bit ob de meat off a skull, feat here of the customary, he might do it by the mere smell of one of his drugs. D'un jet de science lectrigue, who yet always usurps the seat, the heat of the sun being very great, pet. Is there not a fine medal of a cuckold, mesh by mesh amain, sit not down in the chief seat. Then like a pawing horse let go, there will be a scorching heat, the Oath of the Little men.

Part VI

**HAPPILY
EVER AFTER**

[illegible]

INTERLUDE III

Part VII

POST 😞

Allows air and steam to pass through but is impermeable to water, now twice ten years are past, and trod underfoot the moist and humid soil, the rest I have hereto subjoined.

Permet l'air et la vapeur de passer par, mais est imperméable à l'eau, maintenant deux fois dix ans se sont écoulés, et j'ai foulé sous pied le sol humide et humide, le reste que j'ai jusqu'ici soumis.

And the last state of that man, And the sea coast of Tyre and Sidon, as our name out of the list of Mankind, to move from the position of a rose upon the Bush, and the last state of that man.

REFERENCES

- Agichtein, Eugene, Eric Brill and Susan Dumais (2006). 'Improving web search ranking by incorporating user behavior information'. In: **ACM SIGIR conference on Research and development in information retrieval**. New York, New York, USA: ACM Press, p. 19 (cit. on p. 37).
- Amaral, Jose Nelson et al. (2006). 'About Computing Science Research Methodology'. In:
- Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (2011). **Modern Information Retrieval: The Concepts and Technology Behind Search**. Addison Wesley (cit. on pp. 27, 33, 34, 36–39).
- Baidu (2012). **Baidu About** (cit. on p. 28).
- Baldi, Pierre and Laurent Itti (2010). 'Of bits and wows : A Bayesian theory of surprise with applications to attention'. In: **Neural Networks** 23, pp. 649–666.
- Bao, Shenghua et al. (2007). 'Optimizing Web Search Using Social Annotations'. In: **Distribution**, pp. 501–510 (cit. on p. 38).
- Barthes, Roland (1967). 'The Death of the Author'. In: **Aspen 5,6**. the birth of the reader must be ransomed by the death of the Author.
- Basile, Jonathan (2015). **The Library of Babel**. URL: <https://libraryofbabel.info/> (visited on 10/12/2015).
- Bastos Filho, Carmelo et al. (2008). 'A novel search algorithm based on fish school behavior'. In: **IEEE International Conference on Systems, Man and Cybernetics**, pp. 2646–2651 (cit. on p. 37).
- Baudrillard, Jean (2007). **Pataphysics**.
- Beghetto, Ronald A. and James C. Kaufman (2007). 'Toward a broader conception of creativity: A case for 'mini-c' creativity.' In: **Psychology of Aesthetics, Creativity, and the Arts** 1.2, pp. 73–79 (cit. on p. 8).

- Bharat, Krishna and George Mihaila (2000). 'Hilltop: A Search Engine based on Expert Documents'. In: **Proc of the 9th International WWW**. Vol. 11 (cit. on p. 37).
- Bird, Steven, Ewan Klein and Edward Loper (2009). **Natural Language Processing with Python**. Sebastopol, CA: O'Reilly Media (cit. on p. 41).
- Boden, Margaret (2003). **The Creative Mind: Myths and Mechanisms**. London: Routledge (cit. on pp. 6, 10–13, 32).
- Boek, Christian (2002). **'Pataphysics: The Poetics of an Imaginary Science**. Evanston, Illinois: Northwestern University Press.
- Borges, Jorge Luis (1964). **Labyrinths - Selected Stories and Other Writings**. New York: New Directions.
- (1999). **Collected fictions**. Trans. by Andrew Hurley. Penguin.
 - (2000). 'The Analytical Language of John Wilkins'. In: **Selected Non-Fictions**. Ed. by Eliot Weinberger. London: Penguin Books, pp. 229–232.
 - (2010). **La biblioteca de Babel**. Reclam.
- Borges, Jorge Luis and L.S. Dembo (2010). 'Interview with Borges'. In: **Contemporary Literature** 11.3, pp. 315–323.
- Borges, Jorge Luis and Margarita Guerrero (1957). **Book of Imaginary Beings**. Trans. by Andrew Hurley. Viking.
- Brin, Sergey and Larry Page (1998a). 'The anatomy of a large-scale hypertextual Web search engine'. In: **Computer Networks and ISDN Systems** 30.1-7, pp. 107–117 (cit. on pp. 28, 36).
- (1998b). 'The PageRank Citation Ranking: Bringing Order to the Web'. In: **World Wide Web Internet And Web Information Systems**, pp. 1–17 (cit. on pp. 28, 36).
- Brotchie, Alastair (2011). **A supplement**. UK: Atlas Press.
- Brotchie, Alastair and Stanley Chapman, eds. (2007). **Necrologies**. London: Atlas Press.
- Brotchie, Alastair, Stanley Chapman et al., eds. (2003). **'Pataphysics: Definitions and Citations**. London: Atlas Press.
- Brotchie, Alistair, ed. (1995). **A True History of the College of 'Pataphysics - 1**. Trans. by Paul Edwards. London: Atlas Press.
- Brown, Mark (2011). **Patrick Tresset's robots draw faces and doodle when bored**. URL: <http://www.wired.co.uk/news/archive/2011-06/17/sketching-robots> (visited on 24/01/2016).
- Burdick, Anne et al. (2012). **Digital Humanities**. Cambridge, Massachusetts: MIT Press (cit. on pp. 23, 24, 57).
- Burnham, Douglas (2015). 'Immanuel Kant: Aesthetics'. In: **Internet Encyclopedia of Philosophy** (cit. on p. 53).
- Candy, Linda (2006). **Practice Based Research: A Guide**. Tech. rep.

- Candy, Linda (2012). 'Evaluating Creativity'. In: ***Creativity and Rationale: Enhancing Human Experience by Design***. Ed. by J.M. Carroll. Springer (cit. on p. 4).
- Candy, Linda and Ernest Edmonds, eds. (2011). ***Interacting: Art, Research and the Creative Practitioner***. Libri Publishing.
- Chalmers, David (1996). ***The Conscious Mind***. Oxford University Press.
- Cohen, Harold (1999). ***Colouring Without Seeing: A Problem in Machine Creativity***. URL: http://www.kurzweilcyberart.com/aaron/hi_essays.html (visited on 24/01/2016).
- Colton, Simon (2008a). 'Computational Creativity'. In: ***AISB Quarterly***, pp. 6–7 (cit. on pp. 17, 18).
- (2008b). 'Creativity versus the perception of creativity in computational systems'. In: ***In Proceedings of the AAAI Spring Symp. on Creative Intelligent Systems***.
- Colton, Simon, Alison Pease and Graeme Ritchie (2001). ***The Effect of Input Knowledge on Creativity***.
- Colton, Simon and Geraint A Wiggins (2012). 'Computational Creativity: The Final Frontier?' In: ***Proceedings of the 20th European Conference on Artificial Intelligence***. Montpellier, France: IOS Press, pp. 21–26 (cit. on pp. 18, 19).
- Corbyn, Zoe (2005). ***An introduction to 'Pataphysics***.
- Cruickshank, Douglas (nd). ***Why Anti-Matter Matters***.
- Cutshall, James Anthony (1988). 'The Figure of the Writer - Alfred Jarry'. Thesis. University of Reading, p. 258.
- Damerau, Fred J (1964). 'A Technique for Computer Detection and Correction of Spelling Errors '. In: ***Communications of the ACM*** 7.3, pp. 171–176.
- Daumal, Rene (2012). ***Pataphysical Essays***. Trans. by Thomas Vosteen. Cambridge, Massachusetts: Wakefield Press.
- De Bra, Paul, Geert-jan Houben et al. (1994). 'Information Retrieval in Distributed Hypertexts'. In: ***Techniques*** (cit. on p. 37).
- De Bra, Paul and Reinier Post (1994a). 'Information retrieval in the World-Wide Web: Making client-based searching feasible'. In: ***Computer Networks and ISDN Systems*** 27.2, pp. 183–192 (cit. on p. 37).
- (1994b). 'Searching for Arbitrary Information in the WWW: the Fish Search for Mosaic'. In: ***Mosaic A journal For The Interdisciplinary Study Of Literature*** (cit. on p. 37).
- Dean, Jeffrey, Luiz Andre Barroso and Urs Hoelzle (2003). 'Web Search for a Planet: The Google Cluster Architecture'. In: ***Ieee Micro***, pp. 22–28 (cit. on p. 39).

- Deerwester, Scott et al. (1990). 'Indexing by Latent Semantic Analysis'. In: **Journal of the American Society for Information Science** 41.6, pp. 391–407 (cit. on p. 35).
- Dennis, Andrew (2016). 'Investigation of a patadata-based ontology for text based search and replacement'. University of London.
- Dictionary, Oxford English (2015). **animal, n.** URL: <http://www.oed.com/view/Entry/273779> (visited on 10/12/2015).
- Dijkstra, Edsger W. (1988). **On the Cruelty of Really Teaching Computing Science.**
- Ding, Li et al. (2004). 'Swoogle: A semantic web search and metadata engine'. In: **In Proceedings of the 13th ACM Conference on Information and Knowledge Management. ACM** (cit. on p. 38).
- Drucker, Johanna (2009). **SpecLab: Digital Aesthetics and Projects in Speculative Computing.** University of Chicago Press (cit. on pp. 21, 22).
- Drucker, Johanna and B Nowvskie (2007). 'Speculative Computing: Aesthetic Provocations in Humanities Computing'. In: **A Companion to Digital Humanities.** Ed. by Susan Schreibman, John Unsworth and Ray Siemens. Oxford: Blackwell Publishing. Chap. 29 (cit. on pp. 22, 23).
- Du, Zhi-Qiang et al. (2007). 'The Research of the Semantic Search Engine Based on the Ontology'. In: **2007 International Conference on Wireless Communications, Networking and Mobile Computing**, pp. 5398–5401 (cit. on p. 38).
- Dubbelboer, Marieke (2009). 'UBUSING' CULTURE'. Thesis. Rijksuniversiteit Groningen, p. 233.
- Eden, Amnon H. (2007). 'Three Paradigms of Computer Science'. In: **Minds and Machines** 17.2, pp. 135–167 (cit. on pp. 18, 19).
- Edmonds, E. and L. Candy (2010). 'Relating Theory, Practice and Evaluation in Practitioner Research'. In: **Leonardo** 43.5, pp. 470–476.
- Efron, Bradley and Ronald Thisted (1976). 'Estimating the number of unseen species: How many words did Shakespeare know?' In: **Biometrika** 63.3, pp. 435–447.
- Elton, Matthew (1995). 'Artificial Creativity: Enculturing Computers'. In: **Leonardo** 28.3, pp. 207–213 (cit. on pp. 16, 57).
- Flickr (2016a). **flickr.photo.search.** URL: <https://www.flickr.com/services/api/flickr.photos.search.html> (visited on 07/08/2016).
- (2016b). **Getting Started.** URL: <https://www.flickr.com/services/developer/api/> (visited on 07/08/2016).
- Foucault, Michel (1966). 'The Order of Things - Preface'. In: **The Order of Things.** France: Editions Gallimard. Chap. Preface, pp. xv–xxiv (cit. on p. 56).

- Garcia-Molina, Hector, Jan Pedersen and Zoltan Gyongyi (2004). 'Combating Web Spam with TrustRank'. In: **In VLDB**. Morgan Kaufmann, pp. 576–587 (cit. on p. 38).
- Gelernter, David (1994). **The Muse in the Machine**. London: Fourth Estate Limited (cit. on p. 14).
- Getty (2016a). **API Overview**. URL: <http://developers.gettyimages.com/api/docs/v3/api-overview.html> (visited on 07/08/2016).
- (2016b). **Search For Creative Images**. URL: <http://developers.gettyimages.com/api/docs/v3/search/images/creative/get/> (visited on 07/08/2016).
- Glover, E.J. et al. (2001). 'Improving category specific Web search by learning query modifications'. In: **Proceedings 2001 Symposium on Applications and the Internet**, pp. 23–32 (cit. on p. 38).
- Google (2016a). **Crawling & Indexing**. URL: <https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html> (visited on 04/08/2016).
- (2016b). **Search: list**. URL: <https://developers.google.com/youtube/v3/docs/search/list> (visited on 07/08/2016).
- (2012). **Google Ranking** (cit. on p. 28).
- Haveliwala, Taher H (2003). 'Topic-Sensitive PageRank: A Context Sensitive Ranking Algorithm for Web Search'. In: **Knowledge Creation Diffusion Utilization** 15.4, pp. 784–796 (cit. on p. 38).
- Heilman, Kenneth M, Stephen E Nadeau and David O Beversdorf (2003). 'Creative innovation: possible brain mechanisms.' In: **Neurocase** 9.5, pp. 369–79.
- Heisenberg, Werner (1942). **Ordnung der Wirklichkeit**. Trans. by M.B. Rumscheidt and N. Lukens.
- Hendler, Jim and Andrew Hugill (2011). 'The Syzygy Surfer : Creative Technology for the World Wide Web'. In: **ACM WebSci 11** (cit. on p. 53).
- (2013). 'The syzygy surfer: (Ab)using the semantic web to inspire creativity'. In: **International journal of Creative Computing** 1.1, pp. 20–34.
- Hersovici, M et al. (1998). 'The shark-search algorithm. An application: tailored Web site mapping'. In: **Computer Networks and ISDN Systems** 30.1-7, pp. 317–326 (cit. on p. 37).
- Hofstadter, Douglas (1981). 'A Conversation with Einstein's Brain'. In: **The Mind's I**. Ed. by Douglas Hofstadter and Daniel Dennett. Basic Books. Chap. 26, pp. 430–460.
- Holz, Hilary J et al. (2006). 'Research Methods in Computing : What are they , and how should we teach them ?' In: **ITiCSE Innovation and technology in computer science education**, pp. 96–114.
- Hotho, Andreas et al. (2006). 'Information retrieval in folksonomies: Search and ranking'. In: **The Semantic Web: Research and Applications, volume 4011 of LNAI**. Springer, pp. 411–426 (cit. on p. 38).

- Hugill, Andrew (2012). **'Pataphysics: A Useless Guide**. Cambridge, Massachusetts: MIT Press (cit. on p. 56).
- (2013). 'Introduction: transdisciplinary learning for digital creative practice'. In: **Digital Creativity** 24.3, pp. 165–167 (cit. on p. 18).
- Hugill, Andrew and Hongji Yang (2013). 'The creative turn: new challenges for computing'. In: **International Journal of Creative Computing** 1.1, pp. 4–19 (cit. on pp. 19–21).
- Hugill, Andrew, Hongji Yang et al. (2013). 'The pataphysics of creativity: developing a tool for creative search'. In: **Digital Creativity** 24.3, pp. 237–251.
- Indurkha, Bipin (1997). 'Computers and creativity'. Unpublished manuscript. Based on the keynote speech 'On Modeling Mechanisms of Creativity' delivered at Mind II: Computational Models of Creative Cognition (cit. on p. 12).
- Jarry, Alfred (1996). **Exploits and Opinions of Dr Faustroll, Pataphysician**. Cambridge, MA: Exact Change.
- (2006). **Collected Works II - Three Early Novels**. Ed. by Alastair Brotchie and Paul Edwards. London: Atlas Press (cit. on pp. 53, 56).
- Jeh, Glen and Jennifer Widom (2002). 'SimRank: A Measure of Structural Context Similarity'. In: **In KDD**, pp. 538–543 (cit. on p. 38).
- Jordanous, Anna (2015). 'Four PPPerspectives on Computational Creativity'. In: **International Conference on Computational Creativity**.
- Jordanous, Anna Katerina (2011). 'Evaluating Evaluation : Assessing Progress in Computational Creativity Research'. In: **Proceedings of the Second International Conference on Computational Creativity**.
- (2012). 'Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application'. PhD thesis. University of Sussex.
- Jordanous, Anna Katerina and Bill Keller (2012). 'Weaving creativity into the Semantic Web: a language-processing approach'. In: **Proceedings of the 3rd International Conference on Computational Creativity**, pp. 216–220 (cit. on p. 5).
- Jorn, Asger (1961). 'Pataphysics - A Religion In The Making'. In: **Internationale Situationniste** 6.
- Jurafsky, Daniel and James H Martin (2009). **Speech and Language Processing**. London: Pearson Education (cit. on p. 41).
- Kamps, Jaap, Rianne Kaptein and Marijn Koolen (2010). **Using Anchor Text , Spam Filtering and Wikipedia for Web Search and Entity Ranking**. Tech. rep. ? (Cit. on p. 38).
- Kaufman, James C. and Ronald A. Beghetto (2009). 'Beyond big and little: The four c model of creativity'. In: **Review of General Psychology** 13.1, pp. 1–12 (cit. on p. 7).

- Kim, Youjeong and S. Shyam Sundar (2012). ‘Anthropomorphism of computers: Is it mindful or mindless?’ In: **Computers in Human Behavior** 28.1, pp. 241–250.
- Kleinberg, Jon M (1999). ‘Authoritative sources in a hyperlinked environment’. In: **journal of the ACM** 46.5, pp. 604–632 (cit. on p. 37).
- Kleinberg, Jon M et al. (1999). ‘The Web as a graph : measurements, models and methods’. In: **Computer** (cit. on p. 37).
- Koestler, Arthur (1964). **The Act of Creation**. London: Hutchinson and Co (cit. on pp. 6, 9).
- Kurzweil, Ray (2013). **How to Create a Mind**. London: Duckworth Overlook.
- Levenshtein, Vladimir I (1966). ‘Binary codes capable of correcting deletions, insertions, and reversals’. In: **Soviet Physics Doklady** 10.8, pp. 707–710.
- Luo, Fang-fang, Guo-long Chen and Wen-zhong Guo (2005). ‘An Improved ‘Fish-search’ Algorithm for Information Retrieval’. In: **2005 International Conference on Natural Language Processing and Knowledge Engineering**, pp. 523–528 (cit. on p. 37).
- Macdonald, Craig (2009). ‘The Voting Model for People Search’. In: **Philosophy** (cit. on p. 35).
- Maeda, John (2001). **Design by Numbers**. MIT Press.
- Manning, Christopher, Prabhakar Raghavan and Hinrich Schuetze (2009). **Introduction to Information Retrieval**. Cambridge UP (cit. on pp. 34, 41).
- Marchionini, Gary (2006). ‘From finding to understanding’. In: **Communications of the ACM** 49.4, pp. 41–46 (cit. on pp. 31, 32).
- Marchionini, Gary and Ben Shneiderman (1988). ‘Finding facts vs. browsing knowledge in hypertext systems’. In: **Computer** 21.1, pp. 70–80 (cit. on pp. 28, 31).
- Marcus, Mitchell P, Beatrice Santorini and Mary Ann Marcinkiewicz (1993). ‘Building a Large Annotated Corpus of English: The Penn Treebank’. In: **Computational Linguistics** 19.2 (cit. on p. 47).
- Mathews, Harry and Alastair Brotchie (2005). **Oulipo Compendium**. London: Atlas Press.
- Mayer, Richard E (1999). ‘Fifty Years of Creativity Research’. In: **Handbook of Creativity**. Ed. by Robert J Sternberg. New York: Cambridge University Press. Chap. 22, pp. 449–460 (cit. on pp. 4, 5).
- McBride, Neil (2012). ‘A Robot Ethics: The EPSRC Principles and the Ethical Gap’. In: **AISB / IACAP World Congress 2012 Framework for Responsible Research and Innovation in AI**. July, pp. 10–15.
- (2013). **Robot Ethics: The Boundaries of Machine Ethics**. Leicester.
- Microsoft (2016a). **Bing Search API**. URL: <http://datamarket.azure.com/dataset/bing/search#schema> (visited on 07/08/2016).

- (2016b). **Image Search API Reference**. URL: <https://msdn.microsoft.com/en-us/library/dn760791.aspx> (visited on 07/08/2016).
 - (2016c). **Microsoft Translator - Text Translation**. URL: <https://datamarket.azure.com/dataset/bing/microsofttranslator> (visited on 07/08/2016).
 - (2012). **Bing Fact Sheet** (cit. on p. 28).
- Miller, George A. (1995). 'WordNet: a lexical database for English'. In: **Communications of the ACM** 38.11, pp. 39–41.
- Minsky, Marvin (1980). 'K-Lines : A Theory of Memory'. In: **Cognitive Science** 33.4, pp. 117–133 (cit. on pp. 14, 15).
- (1988). **The Society of Mind**. Simon and Schuster, p. 336 (cit. on pp. 14, 15).
- Miyamoto, Sadaaki (1988). **Information Retrieval based on Fuzzy Associations** (cit. on p. 35).
- (2010). **Fuzzy Sets in Information Retrieval and Cluster Analysis (Theory and Decision Library D)**. Springer, p. 276 (cit. on p. 35).
- Miyamoto, Sadaaki and K Nakayama (1986). 'Fuzzy Information Retrieval Based on a Fuzzy Pseudoththesaurus'. In: **IEEE Transactions on Systems, Man and Cybernetics** 16.2, pp. 278–282 (cit. on p. 35).
- Motte, Warren (2007). **Oulipo, A primer of potential literature**. London: Dalkey Archive Press.
- Neeley, J. Paul (2015). **Introducing the NEW Yossarian**. email communication.
- Newell, A, J. G. Shaw and H. A. Simon (1963). **The Process Of Creative Thinking**. New York: Atherton.
- Nick, Z.Z. and P. Themis (2001). 'Web Search Using a Genetic Algorithm'. In: **IEEE Internet Computing** 5.2, pp. 18–26 (cit. on p. 37).
- Nicolescu, Basarab (2010). 'Methodology of Transdisciplinarity - Levels of Reality, Logic of the Included'. In: **Transdisciplinary journal of Engineering and Science** 1.1, pp. 19–38.
- Partridge, Derek and Jon Rowe (1994). **Computers and Creativity**. Oxford: Intellect (cit. on pp. 6, 13).
- Pease, Alison and Simon Colton (2011). 'On impact and evaluation in Computational Creativity : A discussion of the Turing Test and an alternative proposal'. In: **Proceedings of the AISB**.
- Pease, Alison, Simon Colton et al. (2013). 'A Discussion on Serendipity in Creative Systems'. In: **Proceedings of the 4th International Conference on Computational Creativity**. Vol. 1000. Sydney, Australia: University of Sydney, pp. 64–71.
- Pease, Alison, Daniel Winterstein and Simon Colton (2001). 'Evaluating Machine Creativity'. In: **Proceedings of ICCBR Workshop on Approaches to Creativity**, pp. 129–137.
- Peters, Tim (2004). **PEP 20 – The Zen of Python**.

- Piffer, Davide (2012). 'Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research'. In: **Thinking Skills and Creativity** 7.3, pp. 258–264.
- Poincare, Henri (2001). **The Value of Science**. Ed. by Stephen Jay Gould. New York: Modern Library (cit. on pp. 6, 20).
- Polya, George (1957). **How To Solve It**. 2nd. Princeton, New Jersey: Princeton University Press (cit. on pp. 6, 20, 31).
- Queneau, Raymond (1961). **One Hundred Thousand Billion Poems**. Gallimard.
- Raczinski, Fania (2016). **Emails**. personal communication. feedback for his bachelor project.
- Raczinski, Fania and Dave Everitt (2016). 'Creative Zombie Apocalypse: A Critique of Computer Creativity Evaluation'. In: **International Symposium of Creative Computing**.
- Raczinski, Fania, Hongji Yang and Andrew Hugill (2013). 'Creative Search Using Pataphysics'. In: **Proceedings of the 9th International Conference on Creativity and Cognition**. Sydney, Australia: ACM New York, NY, USA, pp. 274–280.
- Ramesh, V., Robert L. Glass and Iris Vessey (2004). 'Research in computer science: an empirical study'. In: **journaltitle of Systems and Software** 70.1-2, pp. 165–176.
- Rhodes, Mel (1961). 'An analysis of creativity'. In: **The Phi Delta Kappan** 42.7, pp. 305–310 (cit. on p. 6).
- Ritchie, Graeme (2001). 'Assessing creativity'. In: **AISB '01 Symposium on Artificial Intelligence and Creativity in Arts and Science**. Proceedings of the AISB'01 Symposium on Artificial Intelligence, Creativity in Arts and Science, pp. 3–11.
- (2007). 'Some Empirical Criteria for Attributing Creativity to a Computer Program'. In: **Minds and Machines** 17.1, pp. 67–99.
 - (2012). 'A closer look at creativity as search'. In: **International Conference on Computational Creativity**, pp. 41–48.
- Schmidhuber, Juergen (2006a). 'Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts'. In: **Connection Science** 18.2, pp. 173–187.
- (2006b). **New millennium AI and the Convergence of history**.
- Schuetze, Hinrich (1998). 'Automatic Word Sense Discrimination'. In: **Computational Linguistics** (cit. on p. 35).
- Schuetze, Hinrich and Jan Pedersen (1995). **Information Retrieval Based on Word Senses** (cit. on p. 35).
- Schulman, Ari (2009). 'Why Minds Are Not Like Computers'. In: **The New Atlantis** 23, pp. 46–68.

- Searle, John (1980). 'Minds, Brains, and Programs'. In: **Behavioral and Brain Sciences** 3.3, pp. 417–457.
- Shattuck, Roger (1959). **The Banquet Years**. London: Faber.
- Shu, Bo and Subhash Kak (1999). 'A neural network-based intelligent meta-search engine'. In: **Information Sciences** 120 (cit. on p. 38).
- Singh, Push (2005). 'EM-ONE: An Architecture for Reflective Commonsense Thinking'. PhD thesis. Massachusetts Institute of Technology.
- Srinivasan, P (2001). 'Vocabulary mining for information retrieval: rough sets and fuzzy sets'. In: **Information Processing and Management** 37.1, pp. 15–38 (cit. on p. 35).
- Stahl, Bernd Carsten, Marina Jirotko and Grace Eden (2013). 'Responsible Research and Innovation in Information and Communication Technology: Identifying and Engaging with the Ethical Implications of ICTs'. In: **Responsible Innovation**. Ed. by Richard Owen. John Wiley and Sons. Chap. 11, pp. 199–218.
- Sternberg, Robert J (1999). **Handbook of creativity**. Cambridge University Press, p. 490 (cit. on pp. 7, 8).
- (2006). 'The Nature of Creativity'. In: **Creativity Research journal** 18.1, pp. 87–98.
- Sutcliffe, Alistair and Mark Ennis (1998). 'Towards a cognitive theory of information retrieval'. In: **Interacting with Computers** 10, pp. 321–351 (cit. on p. 31).
- Taye, Mohammad Mustafa (2009). 'Ontology Alignment Mechanisms for Improving Web-based Searching'. PhD thesis. De Montfort University (cit. on p. 38).
- Thomas, Sue et al. (2007). 'Transliteracy: Crossing divides'. In: **First Monday** 12.12 (cit. on p. 23).
- Turing, Alan (1950). 'Computing Machinery and Intelligence'. In: **Mind** 59, pp. 433–460.
- Varshney, Lav R et al. (2013). 'Cognition as a Part of Computational Creativity'. In: **12th International IEEE Conference on Cognitive Informatics and Cognitive Computing**. New York City, USA, pp. 36–43.
- Ventura, Dan (2008). 'A Reductio Ad Absurdum Experiment in Sufficiency for Evaluating (Computational) Creative Systems'. In: **5th International Joint Workshop on Computational Creativity**. Madrid, Spain.
- Vian, Boris (2006). **'Pataphysics? What's That?'** Trans. by Stanley Chapman. London: Atlas Press.
- Vries, Erica de (1993). 'Browsing vs Searching'. In: **OCTO report 93/02** (cit. on pp. 31, 32).
- Walker, Richard (2012). **The Human Brain Project**. Tech. rep. HBP-PS Consortium.
- Wallas, Graham (1926). **The Art of Thought**. Jonathan Cape (cit. on pp. 6, 20).

- Walsh, Dave (2001). **Absinthe, Bicycles and Merdre**.
- Wickson, F., A.L. Carew and A.W. Russell (2006). 'Transdisciplinary research: characteristics, quandaries and quality'. In: **Futures** 38.9, pp. 1046–1059 (cit. on p. [56](#)).
- Widyanthro, D.H. and J. Yen (2001). 'A fuzzy ontology-based abstract search engine and its user studies'. In: **10th IEEE International Conference on Fuzzy Systems** 2, pp. 1291–1294 (cit. on pp. [35](#), [38](#)).
- Wiggins, Geraint A (2006). 'A preliminary framework for description, analysis and comparison of creative systems'. In: **Knowledge Based Systems** 19.7, pp. 449–458.
- Yang, Hongji (2013). 'Editorial'. In: **International journal of Creative Computing** 1.1, pp. 1–3 (cit. on p. [19](#)).
- Yossarian (2015). **Yossarian**.

KTHXBYE

[Go to TOC](#)