The Title

By

Firstname Middlename Surname (email@aims.ac.rw)

June 2017

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

DECLARATION

This work was carried out at AIMS Rwanda in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS Rwanda or any other University.

Scan your signature

Student: Firstname Middlename Surname

Scan your signature

Supervisor: Firstname Middlename Surname

ACKNOWLEDGEMENTS

This is optional and should be at most half a page. Thanks Ma, Thanks Pa. One paragraph in normal language is the most respectful.

Do not use too much bold, any figures, or sign at the bottom.

DEDICATION

This is optional.

Abstract

A short, abstracted description of your essay goes here. It should be about 100 words long. But write it last.

An abstract is not a summary of your essay: it's an abstraction of that. It tells the readers why they should be interested in your essay but summarises all they need to know if they read no further.

The writing style used in an abstract is like the style used in the rest of your essay: concise, clear and direct. In the rest of the essay, however, you will introduce and use technical terms. In the abstract you should avoid them in order to make the result comprehensible to all.

You may like to repeat the abstract in your mother tongue.

Contents

1. Introduction

In today's world where the most popular and convenient way of storing information is the electronic storage. Electronic storage provides an effective way to process this form of data storage with less human effort. As the stored information increases the difficulty it turns we face in trying access and extract what we need.For the purpose of easily understanding the contents of the data or know what it talks about, topic models serves as a tool for handling this task. Topic models were discovered by researchers in machine learning (ML). It is a statistical tool with a collection of algorithms that reveals the key components to understanding a document.Huge magnitude of unlabelled text can be analysed with a topic model. Topic model algorithms provides the environment that allows users to explore the text in details and summarize it irrespective their size. Topic model is very useful in identifying the patterns of words in a document and in the event that more than one document is involved in the ML, documents with similar patterns can be related. Topic models are unsupervised method of ML, through various algorithms is able to produce cluster of words that represent somewhat a summary of a document. They are applied in search engines to recommend to users what they are interested . A typical summary for a collection of documents is for the analysis of web search, producing results for users in further search Turpin et al (2007). This research focuses on summarizing reports from the international federation of red cross and crescent societies (IFRC). The summary provides a representative topic for each document or in other words best cluster of words that summarize the document.Mathematically, it can be perceived as a function that takes a large text and converts to small one, in a way that thematic structure of the large or original document is preserved.This can be represented as :

$$f : L \longrightarrow S, \quad \text{Such that} \quad |L| << |S|$$

Where L = Large text or document and S = Summarized document or small document. Intuitively the size of S is smaller than L. Manually going through the reports and trying to understand what each is talking about can be time consuming and challenging. The IFRC is a Non governmental organization that provides humanitarian assistance to victims who suffers a disaster event. Through this aid the IFRC generates data of the occurrences of disaster worldwide. Large volumes of complex information are locked in the reports and it hard extracting them going by the manual approach.. This research will employ the Latent Drischlet Algorithm (LDA) the Latent Semantic Algorithm (LSA). Both models extract the contextual meaning from a given large text. This research is divided into five chapters, the second chapter elaborates some key concepts of topic modelling, IFRC and similar work. The third chapter discuss explicitly LDA, LSA and the tools that were also used to arrive at the final results. The fourth chapter covers results and discussions. Chapter five presents conclusion and recommendation.

2. Literature Review

2.1 Topic Model

A topic model is a statistical tool that produce a short description of an original document. Topic models can be applied on a single document or a collection of documents. Blei (2012) described topic models as algorithms that discovers the main themes existing in a large text or document and otherwise the combination of two or more documents. He further reveal that the development of probabilistic topic modelling by ML researchers as a set of algorithms that is geared towards revealing and describing large archives of documents with thematic information. Topic models analyses words in the large text document to discover the themes that pervades them, the connection that exist between the words and their with time. In topic modelling the stress of having to label the documents prior to annotations is saved as it is been done in supervised learning. from the analysis of the original document the topics are obtained. Given an very large volume of electronic archives that is impossible for human annotations, topic modelling can help to summarize and organize it.

2.2 Topic Model methods

Topic modelling techniques have been developed to automatically summarize document or large text. These techniques are: latent semantic indexing/allocation (LSI/LSA), the latent dirichlet allocation (LDA) and the probabilistic latent semantic analysis (PLSA) . This research will narrow between LDA and LSA.

2.3 Vector Space Model (VSM)

In Natural Language Processing (NLP) specially in semantics similarity documents can be represented as a vector of words in in a vector space model (Salton et al, 1975). The frequency of the word in the document determines its importance. Given two documents with words desk and shirt. Say desk appears 6 and 4 times in document 1 and document 2 respectively, and the word shirt appears 3 times in document 1 and 5 times in document 2. Geometrically this can be represented as shown in (2.1).

Table 2.1: Document of words

	desk	shirt
Doc_1	6	3
Doc_2	4	5

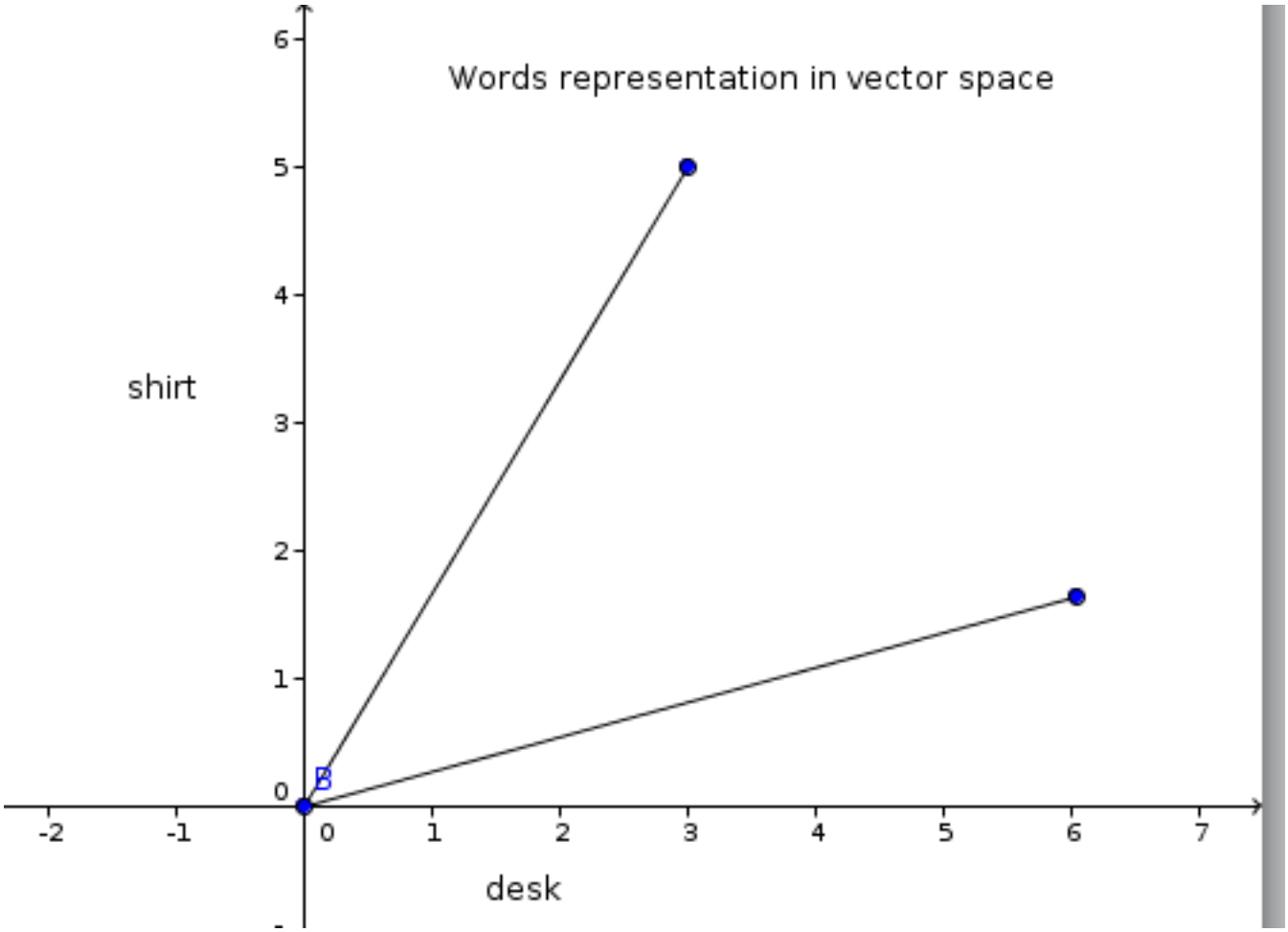


Figure 2.1: Words represented in vector space

Computing the distance between them tells the extent to which their similarity. The dot product shows how close the two vectors are to each other. The dot product of two vectors is given as

$$X \cdot Y = |X||Y|cos(\theta)$$

. Although the VSM is a simple model to use, one main limitation is that, it cannot handle polysemy and synonyms issues. Polysemy is a term that describes words with multiple meaning and synonyms are words that have similar meaning. Considering a query search in Google, by the SVM, the documents relating to the query will not be revealed in the search results. However a research conducted by Erk and Pado pointed out the the reason behind the reason leading to the limitations in SVM. The paper indicated that the existing model does not take syntatic structure into account. Their research resulted in a novel called "structured vector space (SVS)". This work incorporates the selectional preferences for words argument position. With this it is possible to integrate syntax computation of word meaning in context.

2.4 Latent Semantic Analysis

Also known as the latent semantic indexing (LSI) is a topic model method that transforms documents of high dimension to low dimension of words. One useful role played by LSI in topic modelling is its ability to deal with polysemy and synonyms (Griffiths, Steyvors, 2006).

2.5 How The LSA Works

Preliminarily it constructs a matrix $M \in \mathbb{R}^{k*n}$, from the documents $d_1, d_2, ..., d_k$ of words $w_1, w_2, ...w_n$. The rows represents the different words and the columns can be viewed as different documents. For example from (**??**), m_{ij} shows the position and the frequency of the word w_j in document d_i. .To achieve reduction in the dimension of the matrix M the truncated Singular Value Decomposition is applied, given as $M \approx A_t \sum B_t^T$. A_t and B^T are orthogonal matrices, whilst $\sum$ is a diagonal matrix. Reducing the dimension leads to reduction of noise (Deerwester et al, 1990). To achieve document similarity computing the dot product of the row vectors yields desired results.

Table 2.2: Corpus of documents

	d_1	d_2	d_3
food	0	0	2
food	2	5	0
cash	0	1	0
automobile	1	0	4

2.6 The Idea Latent Dirithchet Allocation

Blei(2012) referred to LDA as the simplest topic model . The ideas underlying this model is every document has several topics existing in it.He defined topic to be a distribution over a

fixed a topic.Each topic is made up of words that are very related to the topic. Considering an article with a title "Seeking Life's Bare (Genetic) Necessities," for which data analysis was used to determine the number of genes an organism needs to survive. By hand words pertaining to three different vocabularies were highlighted with different colours. Words such as computer, prediction linked to the topic data analysis highlighted blue, life and evolve about evolutionary biology highlighted pink and words like gene, DNA describing the topic genetics is highlighted yellow. Stop words that occur frequently in the article are removed. The LDA as a statistical tool uses this idea with the assumption that topics are generated prior to words assignment. All words in each vocabulary has a probability value and depending on the topic each word finds itself would be high or low. For example the word "gene" will have a low probability value if it is in the domain of the vocabulary "data analysis" compared to when it belongs to the topic "genetics".He described the process of assigning words in the document to each vocabulary as:

1. From the documents a random distribution over topics is chosen.
2. for each word in the documents:

2a. Randomly choose a topic from the distribution over topics in step 1.

2b. Randomly choose a word from the corresponding distribution over the vocabulary.

The LDA model reflects the idea of multiple topics exhibited by documents. Figure gives a picture of the whole intuition of this generative probabilistic process:

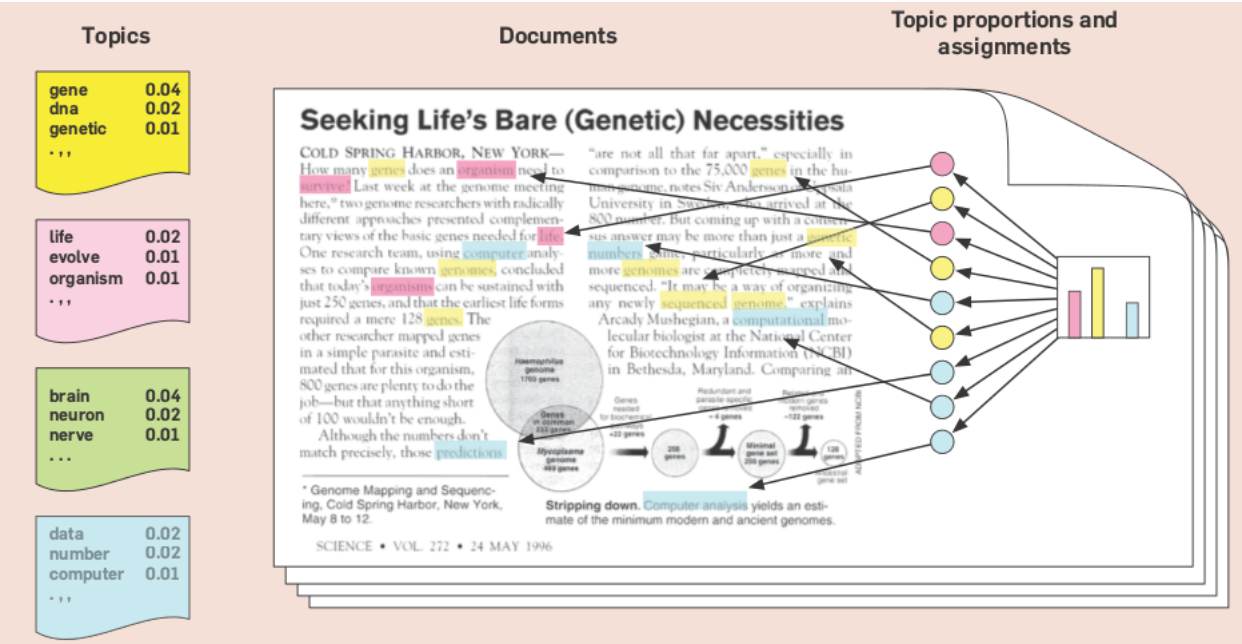
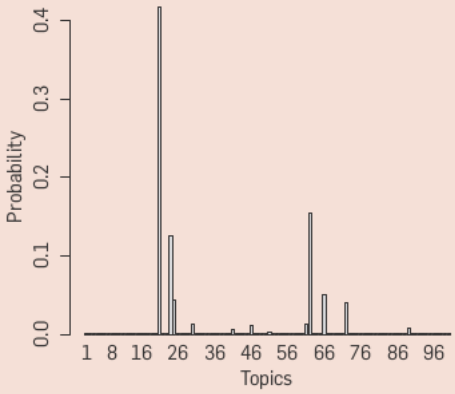


Figure 2.2: Generative intuition of LDA model

In short description of Figure (2.2) the idea underlying LDA is that, first of all some number of topics that are distribution over words is assumed (far left). In generating for each document, firstly choose a distribution over topics (far right ie. histogram), then the circles of different colours are topic assignment for which words drawn from the document corresponds to.

Figure (2.3) shows real inference with LDA, using 17000 articles from the journal of science. Genetics, Evolution, Disease and Computers represents the topics from one article and the words below each are top 15 most frequent words. The graph on the left shows the probability values for each topic. The probability values for this article for a given set of topics may be different from another article. in effect, even though some documents or articles may share the same topics, each article exhibits the topics in different proportions.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 2.3: Infered topics from one article of the 17000 articles froom the journal of science.

2.7 Graphical Model of LDA

Figure (2.4)provides a graphical representation, showing the both the observed and latent variables involved in the generative process. The only observed part is the shaded circle $W_{d,n}$, α and η are parameters from the Diritchlet distribution. What the notations stands for:

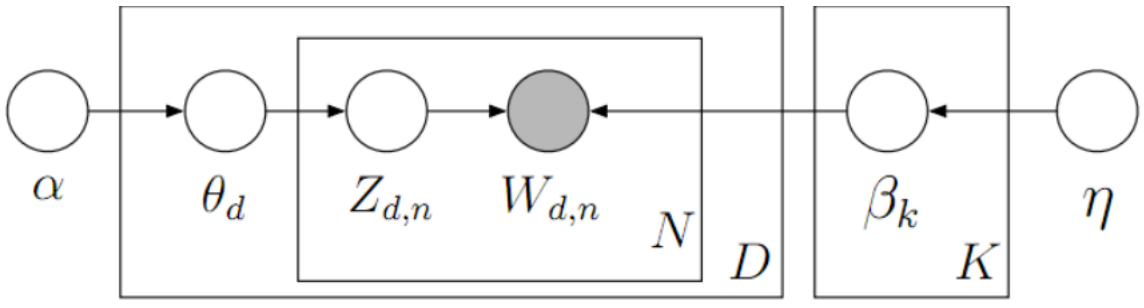


Figure 2.4: Graphical representation of LDA model

- D:the number of documents

- N: number of words in each document
- K: number of topics
- θ_d: the topic proportion for each document d.
- $Z_{d,n}$: is the topic assignment for word n in document d.
- $W_{d,n}$: is the observed worn n in document d,
- β_K:topics

From (2.4) the joint distribution or the total probability of both latent and observed variables is given by:

$$P(\beta, \theta, Z, W, \alpha, \beta) = \prod_{k=1}^{K} P(\beta_i) \prod_{d=1}^{D} P(\theta_d) \prod_{n=1}^{N} P(Z_{d,n}|\theta_d) P(W_{d,n}|\beta_k, Z_{d,n}) \tag{2.7.1}$$

2.8 Diritchlet Distribution

It is from the exponential family of continuous multivariate probability with the parameter α of positive real. It is denoted by $D(\alpha)$ Let $S = [S_1, S_2, ..., S_d]$ as probability mass function, implies $S_i \leqslant 0$ for $i = 0, 1, 2, , ..., d$ and $\sum_{i=1}^{d} S_i$. Also suppose $\alpha = [\alpha_1, \alpha_2, ..., \alpha_d]$ with $\alpha_i > 0$ for each i, and let $\alpha_0 = \sum_{i=1}^{d} \alpha_i$. Then S is said to have a Dirithchlet distribution with parameter α, which is denoted by $S \backsim$Dir(α), if it has $f(s, \alpha) = 0$, if$s$ is not a pmf and if s is a pmf then,

$$f(s, \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} s^{\alpha_i - 1} \tag{2.8.1}$$

where $\Gamma()$ is the Gamma distribution.

2.9 Advantages of LDA

LDA can be used as a module in other complex models to achieve undertake complicated tasks. The LDA model is used in other applications, not only in text.In recent work we have used pairs of LDA modules to model relationships between images and their corresponding descriptive captions (Blei and Jordan,2002). But also include problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics.

2.10 Disadvantages of LDA

The bag of words assumption (the order words in a document does not matter) of LDA makes it unrealistic, however it is reasonable if only our task is to uncover the course thematic structure of the texts (Blei,2012).

2.11 Extentions of LDA

Wallach (2006) developed a model that does not ignore the assumption that the order of words does not matter (bag of words). His work uses the combination of the n gram statistic and latent topic variable by extension of the uni-gram topic model to include properties of a hierarchical Dirichlet language model. Griffiths et al(2005) combined the idea of syntactic and semantics to produce a generative model. This model is capable of simultaneously finding syntactic classes and semantic topics despite having no knowledge of syntax or semantics beyond statistical dependency

3. Methodology

3.1 Unsupervised Learning (SL)

This is a machine learning task of inferring a function to describe hidden structure from unlabelled data. This type of ML does not require any prior manual categorization of observations in the data. The distinction between supervised learning and unsupervised learning (UL) is that in supervised learning there is evaluation of accuracy of the algorithm used used, because data fed to the learner is unlabelled . Also one advantage of UL over SL is that time and cost is saved in labelling as required in SL.

3.2 Natural Language Processing (NLP)

It is a multidisciplinary area that deals with the automatic processing of human language. This automation allows communication between humans and computers. The computer accept input in the form of text or speech and then produces structured representations showing the meaning of those strings as their output.

3.3 Data

The source of the data for this research is from the website of IFRC. Practically the data was obtained by algorithms implemented in the R studio, automatically downloaded the over one thousand pdf reports from the website. Each report is named a a name of a country depicting that the reports describes disaster that occurred in a particular country. Each report has an appeal id, several documents might refer to the same appeal id.

3.4 Word Embeddings

Word embeddings is a dense representation of words in a low dimensional vector space. Bigo et al(2003) introduced the concept of word embedding and then train them in neural language jointly with model parameters. Mikolov et al (2013)came out with the popular word embedding model known as the Word2vec. Pemigton et al (2014) released Glove. The Glove and the Wor2vec are both aimed at producing word embeddings that ecode the general semantic relationship.

3.5 Term Frequency Inverse Document Frequency (TFID)

This measures the extent to which words are important in a document. In topic modelling we to find a group of words that describes a vocabulary. For example topic modelling a document that talks about a university, words such as classrooms, library, lectures, Courses, Grades would tend to be the most important words that describe the topic. It is worth noting that important words are not necessarily the most frequent words, possible to be judged by our intuitive notions.

The TFIDF transforms a vector of integer values into a vector of real values, maintaining the dimension of the original vector.After transformation features which are not frequent in the corpus will have their values increased. That does not mean that rare words all rare words, some may not be significant at all in the description of the topic. For instance dealing with our "university" document, a word lik e"congregation" may rare but then it is significant towards describing the vocabulary. On the other hand a word such as "consequently" may appear very frequent which in this case does not really say anything about the topic. The most frequent words are most words such as "the" or "and," which helps to construct a more sentence, thereby making it readable and understandable.These words do not carry any importance to help topic model a document.They are stop words and they are removed before the modelling irrespective of their number. Given a collection of document with each document d containing words, where each word in the document is denoted i. The frequency of occurrence of a word i in document d is denoted f_{id}. The term frequency TF_{id} computed as

$$TF_{id} = \frac{f_{id}}{max_t f_{tj}}$$

. Which means that the frequency of the word i in document d is fij normalized by dividing it by the term with the highest frequency in the same document of occurrence with stop words exclusive. Intuitively the word which occurs would have a TF of 1, and other words get fractions as their term frequency for this document.

4. The Second Squared Chapter

An average essay may contain five chapters, but I didn't plan my work properly and then ran out of time. I spent too much time positioning my figures and worrying about my preferred typographic style, rather than just using what was provided. I wasted days bolding section headings and using double slash line endings, and had to remove them all again. I spent sleepless nights configuring manually numbered lists to use the LaTeX environments because I didn't use them from the start or understand how to search and replace easily with texmaker.
Everyone has to take some shortcuts at some point to meet deadlines. Time did not allow to test model B as well. So I'll skip right ahead and put that under my Future Work section.

4.1 This is a section

Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.
Some essays may have 3, 5 or 6 chapters. This is just an example. More importantly, do you have at most 35 pages? Luck has nothing to do with it. Use the techniques suggested for writing your essay.
Now you're demonstrating pure talent and newly acquired skills. Perhaps some persistence. Definitely some inspiration. What was that about perspiration? Some team work helps, so every now and then why not browse your friends' essays and provide some constructive feedback?

5. Testing

References

Alan Adolphson, Steven Sperber, and Marvin Tretkoff, editors. *p-adic Methods in Number Theory and Algebraic Geometry*. Number 133 in Contemporary Mathematics. American Mathematical Society, Providence, RI, 1992.

Alan Beardon. From problem solving to research, 2006. Unpublished manuscript.

Matthew Davey. *Error-correction using Low-Density Parity-Check Codes*. Phd, University of Cambridge, 1999.

Leslie Lamport. *LaTeX: A Document Preparation System*. Addison-Wesley, 1986.

D. J. C. MacKay and R. M. Neal. Good codes based on very sparse matrices. Available from www.inference.phy.cam.ac.uk, 1995.

David MacKay. Statistical testing of high precision digitisers. Technical Report 3971, Royal Signals and Radar Establishment, Malvern, Worcester. WR14 3PS, 1986a.

David MacKay. A free energy minimization framework for inference problems in modulo 2 arithmetic. In B. Preneel, editor, *Fast Software Encryption (Proceedings of 1994 K.U. Leuven Workshop on Cryptographic Algorithms)*, number 1008 in Lecture Notes in Computer Science Series, pages 179–195. Springer, 1995b.

Claude Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

Claude Shannon. The best detection of pulses. In N. J. A. Sloane and A. D. Wyner, editors, *Collected Papers of Claude Shannon*, pages 148–150. IEEE Press, New York, 1993.

Web12. Commercial mobile robot simulation software. Webots, www.cyberbotics.com, Accessed April 2013.

Wik12. Black scholes. Wikipedia, the Free Encyclopedia, http://en.wikipedia.org/wiki/Black%E2%80%93Scholes, Accessed April 2012.