NOTTINGHAM TRENT UNIVERSITY

College of Science and Technology
School of Science and Technology

Analysing Narratives: Automatic Descriptive Feature Extraction Through Latent Entity modelling

Adam Slack
N0499528

COMP30151: Project Planning Document

Introduction

The web is a seemingly ever-expanding resource [1], providing access to news, games, research, films, social media, books, and many forms of entertainment. Given the amount of information available to an individual, it becomes necessary to filter and reduce what is available, producing a set of data that is both relevant and in alignment with our personal interests. Without such a reduction in volume a reader would spend an increasing amount of time evaluating the suitability of documents before dedicating any time to reading them.

Analysis of narratives provides an opportunity to utilise or develop novel methods for the field of Natural Language Processing (NLP) that also fall within the wider field of Information Retrieval (IR) [2]. A driving focus of the study will be the use or modification of statistical techniques essential to the identification, relation, and resolution of inter-document entities. The identification of latent abstract entities common to numerous documents, may allow the automatic generation of summaries or comparisons of distinct narratives. Techniques relating to the extraction of latent information and topic modelling, such as Latent Dirichlet Allocation (LDA) [3] or Non-Negative Matrix Factorisation (NMF) [4] may prove useful when attempting to model distributions of entities and descriptive features.

This research aims to determine the viability of a system for the identification and automatic extraction of features present within narratives; The purpose of which is to provide insight into a set of documents without the need to manually extract information. It shall consider attributes deemed potentially useful for the extraction of abstract entities. This could include features specific to individual texts as well as those that frame a document in relation to the larger corpus. The practical applications of such features will also be considered as well as an evaluation of the efficacy of each narrative attribute when summarising a document. Such information could also provide useful insights into an individual's personal preferences, additionally It could help identify methods that can be utilised by authors to write books that are more appealing to their target audiences.

Topic modelling is a subset of Machine Learning (ML) and NLP. Primarily the focus of which is to identify and extract information about abstract topics and concepts within texts. It is also an area of

research that closely relates to the aim of this project. Related literature could potentially be adapted to provide insight into entities present within documents. Research in topic modelling also extends to temporal information, with consideration to the changing prevalence of topics over time [5]. It is possible that similar techniques could be applied to observe the changing pervasiveness of abstract entities across documents.

Automatic summarisation of documents also relates to this investigation. Often the task of producing summaries focuses on the extraction of keywords or salient sentences as well as other features of a document [6,7]. Techniques for producing summaries can apply to both single and multi-document problem spaces [8]. A challenge frequently faced when summarising document is evaluating the outputs of specific abstraction and extraction based methods, It is likely that the methods resulting from this study will encounter the same challenges. Methods for evaluating related techniques do exist [9,10], and this investigation will consider the suitability of these methods when evaluating any novel techniques that are proposed going forward.

Aims and Objectives

The overall aim of this study, is to develop methods for the extraction of information relating to abstract entities present across a corpus. This can be broken down into 3 aims and objectives that will direct the progression of the study.

AO1: *To Identify features of narratives that are effective for the task of document summarization.*
By identifying abstract entities within narratives, it would be possible to describe a corpus in terms of those entities, thus providing a summary. Modelling entity distributions across narratives may require input from various features of a document first. As a result, one objective of this study is to determine exactly which features of narratives are required and the extent of their use.

AO2: *To determine the efficacy of selected existing analytical methods when used individually and in conjunction with each other.*
With topic modelling already being an established field of study within NLP, there exists a suite of analytical methods that provide a decent place to start an investigation of this kind. To avoid

reiterating what numerous academics have previously undertaken it is necessary to consider the efficacy of many of the existing methods with regards to the identification a corpus' latent abstract entities. The evaluation of methods can be carried out through direct comparisons and benchmarking. It is possible that improved performance can be attained through various combinations of statistical and grammatical based methods, so hybrid models will also be considered.

AO3: *To define an extensible pipeline process for the automated analysis of narratives.*

An underlying objective of the investigation will be to produce a pipeline from which data (in the form of narratives) can be provided to a system, and extracted information will be provided as an output. The satisfaction of this method can be measured through the efficiency of such a pipeline as well as through the amount of human interaction that is required by any resulting system. The need for such a system stems from the rate at which new data is generated, using machines to assist in the production of information from data will make the whole process more manageable.

Project Scope and Deliverables

The scope of the project is deliberately flexible. It contains a number of opportunities for refinement, either reducing or increasing the scope as appropriate. As outlined in the Aims and Objectives the project is oriented around identifying specific features of narratives. There are tasks that need to be carried out prior to identifying features as well as after they have been extracted. As a result the project can be described as a series of deliverables..

D1: *Obtain Narratives - identify suitable types and sources of narratives*

In order to begin analysing narratives, it is necessary to determine what types of narratives should be obtained as well as a source for them. The repository of narratives needs to be large enough in scale so that suitably complex analysis can be investigated.

D2: *investigate methods of analysis*

NLP is a large field that is utilised in many other areas of computer science. As a result of this there is plenty of literature available. By reading and evaluating previous related investigations progress with this investigation can be made without repeating mistakes made by others, whilst also identifying potentially effective and novel methods. Given the amount of existing related research, a limit on the

amount of time and resources dedicated to this deliverable must be placed. This ensures that progress on the project can be made.

D3: *Analysis Implementation - testing and evaluating efficacy of methods*

Once methods of analysis have been identified, they need to be investigated. In order to fully determine whether they are of any use and also the extent at which they can be utilised. It is also necessary to compare and contrast different techniques. Benchmarks can be formed by using standard methods to establish a baseline level of performance.

D4: *Review Point One and Two*

Review points are positioned throughout the duration of the project to ensure that progress is being made at an acceptable rate and that any arisen issues can be circumnavigated. Review point one is scheduled for the end of teaching week 4, at this stage this project planning document is to be submitted for review. Review point 2 is to take place during teaching week 15, where additional documents will be completed.

D4: *Project Report*

As the project is being undertaken for the completion of a Bachelors of Science in Computer Science, a formal document reporting the study is to be written. This document will detail all parts of the investigation, including a review of current literature relating to the project, an evaluation of any implementation, and a discussion of any results from the project. A deadline for the report has been set for Wednesday of teaching week 24.

D5: *Project Demonstration*

Scheduled for the end of teaching week 25, a demonstration of what has been completed during the project is expected. This will provide an opportunity to present any findings from the study whilst also satisfying requirements placed by the NTU for the attainment of a BSc (Hons) Computer Science. It is expected that no further work will be required for the project by this point. The only remaining tasks will be preparation for the demonstration.

Project Scope

The flexible nature of the project doesn't prevent certain tasks from being out of scope from the

outset. Producing a system that assimilates a corpus in its entirety would be completely out of scope. Likewise, in depth analysis of numerous features of a corpus is also out of scope, there is however a distinction that needs to be made between features of a corpus and features of an individual narrative. It isn't necessarily out of scope that a select amount of features present in narratives be investigated. This is because they may be essential to achieving the ultimate aim of analysing an individual aspect of the corpus.

The development of a fully autonomous system for the processing of documents is on the fringe of being out of scope of the project. The task of accumulating documents and integrating them into the system would likely prove to be as big a task as the investigation of common abstract entities. It wouldn't be out of scope however to define the behaviour, or to implement a basic prototype of such a system. Such a system would likely require a human to build a corpus prior to analysis. Whilst development may take time to carry out, it would assist in evaluation of any devised methods, ultimately easing other tasks.

Sources of Information

Existing literature is an essential resource. It offers insight into what researchers have previously worked on. The search engine, Google [11], provides search methods for finding prior and related research. The levels of accessibility that Google provides to locations where journal and conference papers are available, means that to not use such a tool would be a hindrance to the project.

Similarly Wikipedia [12] is also a valuable source of information. In such an active area of research, wikipedia is frequently updated with new information. Whilst the reliability of wikipedia is commonly questioned, and the depth at which topics are discussed is not necessarily as detailed as required [12,13] It does however succeed at being a source for finding related work on a topic. Specifically wikipedia provides collections of sources for what is discussed earlier on the page. It is this list of references that make wikipedia a valuable tool.

Subject experts will also be useful sources of information. A number of Academics at Nottingham Trent University (NTU) have extensive experience with a variety of NLP, ML and Computational

Intelligence topics. Through discussions about the field it is hoped that new ideas can be generated and guidance on specific topics can be given.

Required Resources

The Gutenberg project [14] is a valuable resource for project involving natural language. Featuring over 54,000 freely available narratives, the site can be scraped with relative ease. Such a resource will prove useful to this project for numerous reasons. Primarily it makes enough data available that meaningful analysis can be performed on inter-document descriptive features, whilst making other ML methods an option should early investigations prove useless. Additionally should the resource be used as part of an automated pipeline, there are enough instances that meaningful consideration can be given to the efficiency of such a pipeline.

From obtaining large corpuses of text, to performing complex analysis, some of the tasks required by this project may take hours, or even days to complete. For that reason access to a machine that can be left running 24 hours a day is required. This will mean that processes can be executed continuously overnight and even over a couple of days. An example scenario where such a resource would be of use is seen in the obtaining of books from Project Gutenberg, It is estimated that it would take approximately 48 hours to download and unzip all available texts. NTU have kindly offered an appropriately specced Virtual Machine (VM) on a University owned server for academic purposes. This VM will be able to provide much of the throughput required by the project.

Project Risks

There are a number of risks to the project that would be detrimental to its success. Anything from failing to obtain any or enough narratives for analysis, to not having access to computational resources that allow analysis to be performed. This section will outline some of the more likely and impactful risks to this project.

R1: *Failure to Obtain any or enough Narratives*

As the project relies on having data to analyse, one of the obvious risks is that suitable data cannot be obtained. The most likely reason for this is legal issues, given that work in the UK is subject to

copyright laws, it is possible that common public collections of narrative work is not legally available in the UK. The impact this would have on the project should this occur would be detrimental. As discussed later, only work that can be accessed in a legal manner can be used in textual analysis. Should UK copyright law prohibit the use of works featured in Project Gutenberg, then it would mean purchasing access to narratives or finding alternative means of obtaining them. It is possible however that the impact can be mitigated by the fact that british libraries can provide a means of legally accessing work, even in eBook form [15,16].

Whilst access to libraries can mitigate some of the impact of having no narratives to analyse, it is possible that obtaining enough narratives to perform in-depth analysis would still be difficult. Many methods rely on having dense networks being formed between elements, such collaborative methods suffer from what is referred to as the sparsity problem [17]. Should there be a lack of narratives available, then it is likely that this study also falls foul to the sparsity problem and other features of narratives will need to be investigated.

R2: *Failure to successfully analyse any narratives*

The severity of this problem essentially means that the project would be rendered useless. There are many different methods of analysis that can be investigated, however the overall aim of the project is to investigate the extraction of common latent abstract entities and the composition of narratives featuring them. Should this prove to be too challenging a task then there are other aspects that would have to take precedence. In order to mitigate the risk of failing the primary aim of the project, the comparison of analysis methods and implementation of an automated pipeline would become the new focus of the project. This would still provide enough substance that a project could be structured around them.

The challenge of mitigating this risk is identifying a suitable time to stop. It might not be clear that the task is unachievable until it is too late. To combat this, regular reviews of the current state of the project will be carried out. This will take the form of personal progress reviews, as well as weekly meetings with a project supervisor.

R3: *Can't carry out analysis for long enough*

Working on the assumption that a large corpus can be built, extended periods of time may need to be allocated for a system to assimilate such a corpus. Statistical methods that require iterating over the contents of a corpus will take increasing amounts of time to complete for every additional narrative. To limit the effect of time constraints associated with this project, the size of the corpus will be adapted according to the methods used as well as increased incrementally. Controlling this risk is essentially an optimisation problem, what amount of books provides the most amount of insight, whilst also taking a reasonable amount of time to analyse. Whilst mathematically modelling this as a function and finding optimal input parameters is an option, it may be easier to incrementally increase the size of a corpus, limiting its size when only marginal gains are made with results from analysis, or when the amount of time required to carry out a process begins to prevent progress.

Evaluation of Professional, Social, Ethical and Legal issues.

Many narratives exist in the form of books, and given that this study will focus on the analysis of narratives present in books, it is prudent that consideration of the legal issues surrounding the analysis of published works be considered. Many books are subject to copyright and local laws, many countries have different rules and regulations regarding the use of published works and intellectual property. The World Intellectual Property Organisation (WIPO) is a United Nations (UN) agency concerning itself with the protection of property rights across the world, some of the tasks involved with this include standardising law across UN member states. [18]

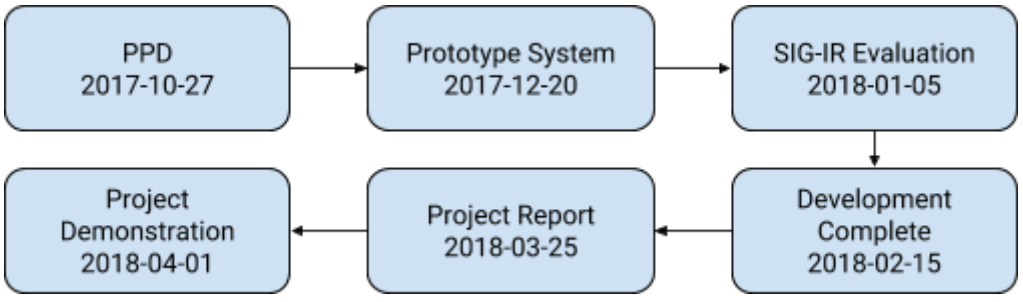
The UK Gov't summarises that copyright protects work from being: copied, distributed, rented or loaned, performed or displayed publically, and adapted. [19] Section 29 of Copyright, Designs and Patents Act 1988 outlines the conditions of copyright for research and private study [20], section 29(1) specifically states that use of protected work is considered fair dealing if protected work is used "*for the purposes of research for a non-commercial purpose* [...] *provided that it is accompanied by a sufficient acknowledgement*." Text and Data Mining is also outlined as an exception to UK copyright laws. [21] The UK Gov't summaries the legislation as meaning textual analysis of protected work is allowed if there is lawful access to the work in question. [22]

There are limited direct ethical issues with the proposed project, however there are still points for

discussion. Informed consent is frequently raised as a point for discussion in research projects involving people. Whilst there will be no direct involvement of individuals, there will extensive use of work produced by others. This raises the question of whether consent to use their work should be requested. Analysis of narratives could be used to provide insights into an author, and thus could be considered secondary data. Arguably the use of such data should be held to the same ethical standards as clinical data. Additionally should abstract entities be identified across narratives, there would then exist information that could be used to compare distinct narratives, if two works contain the same abstract entities then it could raise the question of whether one work is derivative of the other. This raises moral issues both socially and professionally, given that the integrity of an author could then be brought into question. Whilst the legality of such analysis is lawful, socially and professionally there are still questions. However given that it is common practice to review books and other products, the ethical implications can be lessened as authors open themselves to criticism with each publication.

Timeline

This project deadline for all tasks of this project is the end of teaching week 25 (2018-04-01) however practical development work on the project should be completed much earlier than that (2018-02-01) leaving only documentation and experimentation to report on.



References

1. Web Server Survey | Netcraft Available at: https://news.netcraft.com/archives/category/web-server-survey/ [Accessed October 16, 2017].

2. Smeaton, A.F. (1999). Using NLP or NLP Resources for Information Retrieval Tasks. In Natural Language Information Retrieval Text, Speech and Language Technology. (Springer, Dordrecht), pp. 99–111.

3. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. J. Mach. Learn. Res. *3*, 993–1022.

4. Lee, D.D., and Seung, H.S. (2001). Algorithms for Non-negative Matrix Factorization. In Advances in Neural Information Processing Systems 13, T. K. Leen, T. G. Dietterich, and V. Tresp, eds. (MIT Press), pp. 556–562.

5. Newman, D.J., and Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. J. Am. Soc. Inf. Sci. *57*, 753–767.

6. Erkan, G., Radev - Journal of Artificial Intelligence Research, D.R., and 2004 (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. jair.org. Available at: http://www.jair.org/papers/paper1523.html.

7. Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '99. (New York, NY, USA: ACM), pp. 121–128.

8. Goldstein, J., and Kantrowitzt, M. Multi-Document Summarization By Sentence Extraction. Available at: http://delivery.acm.org/10.1145/1120000/1117580/p40-goldstein.pdf?ip=152.71.207.142&id=1117580&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=997908537&CFTOKEN=73239202&__acm__=1508768857_e68037103a20e32986729d89a20842fc.

9. Wallach, H.M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation Methods for Topic Models. In Proceedings of the 26th Annual International Conference on Machine Learning ICML '09. (New York, NY, USA: ACM), pp. 1105–1112.

10. Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, S. S. Marie-Francine Moens, ed. (Barcelona, Spain: Association for Computational Linguistics), pp. 74–81.

11. Google Available at: https://www.google.co.uk/?gfe_rd=cr&dcr=0&ei=06LpWfKTEKfS8AeZyJKICA [Accessed October 20, 2017].

12. Wikipedia contributors (2017). Main Page. Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=798174323 [Accessed October 20, 2017].

13. Azer, S.A. (2014). Evaluation of gastroenterology and hepatology articles on Wikipedia: Are they suitable as learning resources for medical students? Eur. J. Gastroenterol. Hepatol. *26*, 155.

14. Project Gutenberg Project Gutenberg. Available at: http://www.gutenberg.org/ [Accessed October 19, 2017].

15. Nottingham City Council Borrow E-books and E-magazines. Nottingham City Council. Available at: http://www.nottinghamcity.gov.uk/libraries/borrow-e-books-and-e-magazines/ [Accessed

October 20, 2017].

16. e-books and e-audio Available at: http://libraries.essex.gov.uk/e-books-e-audio-e-magazines-and-book-groups/e-books-and-e-audio/ [Accessed October 20, 2017].

17. Grčar, M., Mladenič, D., Fortuna, B., and Grobelnik, M. (2006). Data Sparsity Issues in the Collaborative Filtering Framework. In Lecture Notes in Computer Science, pp. 58–76.

18. Inside WIPO Available at: http://www.wipo.int/about-wipo/en/ [Accessed October 22, 2017].

19. Intellectual property: Copyright - GOV.UK Available at: https://www.gov.uk/topic/intellectual-property/copyright [Accessed October 20, 2017].

20. Participation, E. (1988). Copyright, Designs and Patents Act 1988. Available at: https://www.legislation.gov.uk/ukpga/1988/48/section/29 [Accessed October 20, 2017].

21. Exceptions to copyright Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

22. Exceptions to copyright - GOV.UK Available at: https://www.gov.uk/guidance/exceptions-to-copyright [Accessed October 20, 2017].