UNIVERSITY OF
LINCOLN

Project Report

Comparing the Results of Topic Models Against the Academic Consensus on the Primary Motifs in a Dataset of Poetry.

A dissertation submitted in partial fulfillment of the requirements for the degree of
BSc (Hons) Computer Science

Tomos Davies (DAV17639930)
School of Computer Science
University of Lincoln
April 2020

Table of Contents

Acknowledgements

I'd like to thank the university's Computer Science Society and its Discord server for the academic and moral support that's been a constant aid to my motivation. Thank you to Dan, Billy, and Will for the support, compassion, and solidarity through the ups and downs of this project.
Special thanks to Astrid for reminding me to take breaks and stay positive.

Many thanks to my project supervisor, Yvonne James, for all the support.

Thank you to the poets whose work I have attempted to get computers to read.

"*Science and literature alike are readers of the world. And, sooner or later, both lead us to the unreadable, the boundary at which the unintelligible begins*." - Karl Ove Knaussgard

Abstract

Topic models, which detect latent themes in a corpus of documents to group co-occurring keywords together in thematically comprehensible ways, were generated using the Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) algorithms with three datasets of poetry from different time periods. A close reading of the results as well as a study to measure interpretability were used to measure which algorithm was the most successful at uncovering specific themes in each dataset established using relevant literary studies. Comparison between the two algorithms' performances served to indicate which method was the most successful in modelling this highly figurative language. Our findings indicated that LDA generated the most thematically comprehensible topics, owing to improved performance in identifying context and polysemy in the vocabulary used throughout the corpora, as well as having more parameters available to tune and optimise performance.

Introduction

'Understanding how topic modelling algorithms handle figurative language means allowing for a … beautiful failure — not a failure of language, but a necessary inclination toward form that involves a diminishing of language's possible meanings.' (Rhody, 2012).

Topic modelling works on the principle that, in a given textual dataset, 'each document can be described by a distribution of topics and each topic can be described by a distribution of words' (Ganegedara, 2018). By utilising this principle, one can measure the similarities and differences between two or more documents based on the way in which a topic model categorises their semantic patterns. A website may use a topic model on two different articles to see if one should be recommended to be read after another, or a large corpus could be analysed to measure gradual changes in tone and vocabulary as significant historical events unfold.

Most high profile applications of topic modelling (e.g. on newspapers and to measure popular trends in scientific journals) tend to share a common feature in their textual datasets - the use of denotative and literal language. Should an abstract from a scientific journal speak of an 'atom', a topic modelling algorithm would associate it with other words in the corpus one may associate with the semantic meaning of 'atom' ('molecule', 'element' etc). In this project, topic modelling will be used on corpora of poetry to measure the extent to which a model can give us insight into latent semantic patterns in figurative and connotative language, wherein a word like 'atom' could instead work as part of an analogy, stripping away its literal meaning. This involves a process of abstracting this figurative language in a way that is captured in the quote above by Lisa M. Rhody - a process whose potential for success is the problem tackled by this project.

This is relevant and insightful as a Computer Science project. Applying topic modelling to poetry is not an entirely novel use of the algorithms, as detailed in the literature review, but its application to the problem domain of these specific poetic movements has so far been overlooked and thus justifies proper investigation into its potential. Two topic modelling algorithms will be implemented with the datasets - Latent Dirichlet Allocation and Latent Semantic Analysis. Comparing the results of these two against each other and analysing the different ways in which these algorithms have modelled the topics in the poetry will give insight towards both the potential of topic modelling with figurative language in general and towards measuring which algorithm is the most effective for doing so. Based on topic theme categorisation, an estimate of readability using a survey, and some qualitative calculations, we will aim to discover the best performing algorithm for this goal.

Background and Literature Review

Discussion regarding the academic background of this project can be divided into previous experiments of LDA vs LSA, a general review of topic modelling with poetry, and a discussion on the motifs featured in our datasets that we will use to contextualise our results.

While both of these algorithms are being used to generate topic models, their results may differ in some ways. Bergamaschi et al. (2014) performed a study comparing the results of the two algorithms in a project where topic modelling was used to create a film recommendation system. Using similarities between established keywords featured in the topic models created, films were recommended to users and the extent of their successes recorded in a survey. Participants were asked whether the best recommendations were made using LDA or LSA, and the study concluded that LSA had outperformed LDA fairly drastically, its performance being estimated as almost twice as successful as that of LDA. It was noted, however, that LSA ran into issues in its models with polysemy, wherein a word may have different meanings depending on the context. LDA was able to account for polysemy in its topics where LSA could not. This could indicate that LDA is better able to account for thematic context in the keywords - should this be the case, LDA could potentially outperform LSA in capturing motifs in our figurative dataset wherein context can define the way a word is incorporated into a theme.

A comparison of the two algorithms was also performed by Cvitanic et al. (2016), where they were used to measure similarities between documents in a dataset of patents from the United States Patent and Trademark Office. It concluded that LSA had been more effective at identifying similar documents, and that the primary reason LDA had underperformed was due to its performance relying entirely on the amount of topics specified (whereas LSA's results were more consistently coherent), said amount having no method of being optimised beyond trial-and-error. They noted that 'a method to determine the best number of topics for the LDA algorithm is much needed' (ibid.). These projects had a different aim to ours in that they aimed to accurately measure similarities between documents in their application of topic modelling, whereas we will be measuring the algorithms' effectiveness in identifying pre-established themes in our dataset. Despite this, their results tell us we should be very careful in specifying the amount of topics we generate, and that it may be this variable that has the most impact on the quality of our conclusions. In addition, despite LSA being chosen as the best-performing algorithm in both of these comparative studies, the datasets being used are far more objective than our dataset of poetry, meaning our results have a high chance of being incongruent with these conclusions.

The primary sources for information that will contribute to this project's foundational viability as an application of topic modelling with a dataset of poetry are '*Topic Modeling and Figurative Language: Revising Ekphrasis*' by Lisa M. Rhody (2012) and '*On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry'* by Borja Navarro-Colorado (2018). Using two very different poetic datasets, these studies both go into great detail on the different ways one can interpret the results of topic modelling with figurative language, how the datasets should be pre-processed, and how the results can be analysed. The distinctions between this project's dataset and theirs will be kept in mind during the exploration of their findings and conclusions.

Poetry is an atypical corpus to be used with topic modelling. Rhody (2012) goes as far as to describe topic modelling as being 'designed to be used with texts that employ as little figurative language as possible'. Despite this, the results of both projects were quite illuminating, but had to be interpreted in distinctly different ways to more typical applications of the Latent Dirichlet Allocation (LDA) algorithm. Rhody's (ibid.) goal in using topic modelling with poetry was to see if it was possible to discover specific tropes, identified using relevant literary studies, in Ekphrastic poetry (poems about the visual arts) using a

dataset of 4500 poems. Navarro-Colorado's (2018) goal was to uncover and analyse what kind of topics would be unearthed from a dataset of 5000 sonnets from the cultural Spanish 'Golden Era', covering a much broader range of themes, with no prior predictions for the results. This project's aim falls somewhere in the middle - while we are exploring specific poetic movements/genres in order to see if established tropes and themes can be successfully identified through topic modelling, similar to Rhody, these movements are broad enough that they are likely to approach as wide of a range of themes and language as Navarro-Colorado's dataset. Therefore, when it comes to pre-processing the poems, our focus should be placed on Navarro-Colorado's discoveries on the best methods for doing so; for analysis of the results, we should turn to Rhody's article.

Navarro-Colorado (2018) experiments with pre-processing his dataset of Spanish sonnets in two different ways. One exclusively incorporates *stopword removal*, where specific words from a prepared list, such as connectives and pronouns (which are not substantive and generally useless in natural language processing) are filtered out and ignored as the model trains so that each keyword is inherently meaningful. The other makes use of stopword removal in addition to a process called *lemmatisation*, which reverts inflected words into their *lemma*, or the 'base form' one might find of a word in a dictionary - for example, the lemma of the inflected words (or *lexemes*) 'runs' and 'running' would be 'run'. This has the potential to be beneficial for topic modelling, as lemmatising every inflection of the word 'love' ('loves', 'loved', 'loving' etc.) would allow for less semantically identical keywords in a topic concerning the theme of 'love' and feature different, equally meaningful ones in their place. Ultimately, Navarro-Colorado concludes that for his application of topic modelling with poetry, lemmatisation is not appropriate and in fact leads to less coherent results; he notes this may be due to linguistic features in languages which involve 'rich verbal inflection' that inherently carry 'temporal, modal and aspectual' features of the subjects. For example, the Spanish word 'fue', meaning 'was', disappears from a topic which concerns the downfall of the Spanish Empire after being lemmatised into 'is', removing a semantically significant word which reflects the sense of looking-back explored in the rest of the topic. We will assume that a similar loss in figurative meaning could occur with our datasets of English poetry, and will also skip any kind of lemmatisation.

There is no consensus among these researchers on the optimal number of topics for a poetic topic model. Navarro-Colorado's finalised model generates 100 topics (after experimenting with 10, 25 and 50); Rhody's generates 60. It would seem that having a high amount of topics that stay thematically coherent depends on having a large dataset of several thousand poems.

Rhody (2012) categorises the topics in her results in four different ways.

- **OCR and Dialect Features** - This includes digit recognition errors and topics that consist only of words exclusive to a dialect or language. In Rhody's project, these included topics consisting of Spanish words, which came from poems in the dataset that had not yet been translated. As mentioned later, some poems in our Harlem Renaissance dataset include dialect poetry whose distinct features are inseparable from the themes and motifs at the core of the movements, and any topics we uncover which would fall into this category should still be analysed. Dialect poetry in our Romantic dataset such as the Scots poems of Robert Burns may also appear in topics such as these.
- **Large 'Chunk' Topics** - Some poems include much more use of repetition than others, and inclusion of these in the corpus can sometimes pull the keyword distribution in the topics towards these poems. For example, Topic 12 in Rhody's results includes keywords that consist exclusively of words that are repeated frequently in just one of the poems in her dataset. In our dataset this is likely to occur with longer poems such as Lord Byron's 'Don Juan', which includes the title character's name almost 400 times (Byron, 1837). In cases where these topics appear,

research could be briefly conducted to find out from which poem they could have appeared and if there is a relation between the keywords beyond a linguistic one.

- **Semantically Evident Topics** - Topics whose results appear as expected, consisting of keywords that are clearly semantically related, e.g. 'night, stars, moon, sky, dusk', etc. Rhody notes that these topics should not always be interpreted at face value, but since these will reflect a theme in the most coherent manner, they may be the most important ones to analyse and present in studies.
- **Semantically Opaque Topics** - Topics with 'little to no comprehensibility' in terms of a clear thematic relationship between keywords. We should not throw these topics away entirely - a close reading of the poems whose contents are primarily included in the keywords could be carried out in order to see if the connection between these words is thematic and/or figurative rather than the potentially literal connections of the semantically evident topics.

The result categorisation system will be useful for this project and will be used when analysing results. There is a chance (especially with the semantically opaque topics) that a relationship between the keywords will be incomprehensible or nonsensical for human reading. Quantifying how many of these topics appear in our final model is a task that should not be carried out exclusively through self-reporting, as there are several other illuminating methods we should consider.

One goal in the evaluation of a topic model is to quantify the topics' *coherence*, which will be discovered formulaically in the same way for both models using the Python library Gensim's CoherenceModel function. This library uses a formula proposed by Röder et al. in '*Exploring the space of topic coherence measures*' (2015), using a pipeline of segmentation, probability estimation, confirmation measurements and aggregation.

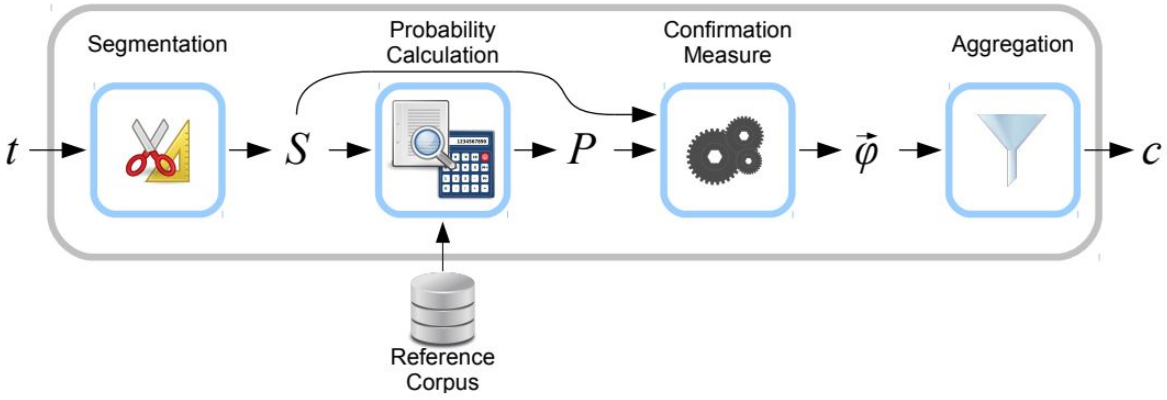


Fig. 3 - Flow diagram of the stages in the Gensim CoherenceModel pipeline (ibid.)

This is a justified way of quantitatively evaluating different classes of topic models. Srinivasa-Desikan (2018) writes that this allows for the comparison of the performance of two different classes of topic models, including LDA and LSA - he remarks that Gensim's CoherenceModel method gives a 'quantitative way to measure the optimal number of topics for a corpus, as well as a way to compare between completely different classes of models'. We must consider the unusual application of this project, however, and a high likelihood of quantitative statistics such as these having unpredictable results. Nonetheless, we will compute this statistic to see if Srinivasa-Desikan's assumption stays applicable.

Another method will involve assistance from human participants; rather than using a 'word intrusion' test where subjects are asked to pinpoint which unrelated word has been inserted at random into each topic in

order to establish a general idea of the models' success, we will instead use a 'topic intrusion' method. Both Rhody and Navarro-Colorado mention the possibility of using 'word intrusion' evaluation, but Rhody advises against it on the basis of poetic topic modelling often producing the semantically opaque topics mentioned earlier, because poems 'purposefully access and repurpose language in unexpected ways' (2012). These may be 'coherent' in a sense that is not detected in human reading, and the 'intruder word' system could therefore be misleading in its representation of the model's success, being biased towards the more human-readable semantically evident topics. The alternative 'topic intrusion' method adds a topic created from an entirely unrelated (potentially randomised) dataset and places them among the real topics, and the participant is asked to identify it. Should any of the real topics be classified as 'fake', it would indicate it as appearing to be thematically unrelated to the rest of the topics. This 'topic intrusion' method of qualitative data gathering is used by Guzman et al. in a project using topic modelling to group semantically related tweets to measure 'if the association between topics and tweets is accurate according to human judgement' (2017). Rhody (2012) also mentions it as a valid way of evaluating the comprehensibility of poetic topic models specifically, further justifying this method over smaller-scale word intrusions.

The more successful these participants are at recognising the 'fake' topics, the more thematically consistent the results. Since we are aiming to see if the primary themes and motifs of the poetry can be identified, high thematic consistency could be considered indicative of a successful model.

Next, to contextualise the results, we will discuss the specifics of our dataset and natural language processing with poetry. This project will make use of three datasets - one for Romantic poetry, one for Metaphysical poetry, and another for 20th century Harlem Renaissance poetry. In order to assess the results of our topic models, we must first understand what kind of motifs and vocabulary we should hope to see, a task for which we should turn to academic writing on our three poetic datasets.

The themes of 19th century Romantic poetry primarily involve an emphasis on emotional, dramatic language addressing nature, love, and grand depictions of sublime events, whether they are beautiful or grotesque. Some poems turn everyday events or sights into observations on life, the spirit and the divine, binding mundane and divine topics together - for example, Tennyson's 'Flower in the Crannied Wall', a poem about a religious experience attained through the discovery of a flower in the cracks of wall (Tennyson, 1863). Religious themes pervade Romanticism; Harold Bloom firmly categorises the movement as 'religious poetry', noting that 'there is no more important point to be made about Romantic poetry' than its background of 'Protestant dissent' (Bloom, 1971, p.17).

As a result of the prevalence of these themes, the topics we will unearth from our dataset of Romantic poetry are likely to involve nature, life, death, religion, love, the spirit, supernaturalism and superlative or dramatic adjectives. The general vocabulary used in Romantic poetry could be considered fairly antiquated - with use of 'thee','thou' etc being quite common - though not to the same extent as our Metaphysical dataset. Word spellings and grammatical syntax are more or less identical to language in use today - the English of the 19th century is similarly considered 'Late Modern English' (Mastin, 2011a). Nonetheless, we will include any antiquated variations on connectives and pronouns we may not need to analyse in our list of stopwords, which the topic modelling algorithm will ignore during analysis.

The primary motifs of Metaphysical poetry (17th century) as a whole are more difficult to categorise than those of Romantic poetry, with T.S. Eliot remarking that 'any definition of Metaphysical poetry will be only a partial success' (Eliot, 2014, p.60). Speaking very generally, it is often thematically involved with religion, love, and the relationship between man and the divine by 'making the Word Flesh' (ibid., p.61), typically including astronomical imagery, elaborate vocabulary, satirical elements, and

allusions to nature. It includes such poets as John Donne and Richard Crashaw, studies of the latter being remarked by Eliot as necessitating a prior study of religion in the first half of the 17th century due to its pervasion of his work (ibid., p.63), making religion a likely topic to appear in our models in some form. The oldest dataset explored in this project, the language is particularly antiquated and primarily categorised as 'Early Modern English', whose features include the letter 'e' at the end of words which did not go on to retain it and different vowel sounds to Modern English (Mastin, 2011b). The movement in general is rife with archaic spellings (e.g. *'Let mans Soule be a Spheare'* from Donne's 'Good Friday, 1613, Riding Westward', *'Come and let us live my Deare, / Let us love and never feare, / What the sowrest Fathers say'* from 'Out of Catullus' by Richard Crashaw).

From our analysis of the general motifs of the movement, it can be concluded that our topics are likely to explore religion, the soul, the supernatural, nature, love, mortality, and humanity's place in the universe. The antiquated vocabulary and spelling in this dataset may cause issues during the modelling process - for example, the algorithms would not recognise that 'Soule' and 'Soul' are the same word and would categorise them as separate terms - ideally, however, they would still be classified in the same topic. Such antiquated variations on the spellings of common words may be too numerous to account for with stopwords, a fact that may need to be taken into account during the preprocessing of the poems should the topics feature too many duplicates.

Our third dataset explores works from the era in the early 20th century known as the Harlem Renaissance, an artistic movement of African-American cultural expression which included music, film, and poetry, roughly corresponding to the 'Great Northward Migration' of African-Americans from the southern USA after the implementation of Jim Crow segregation and an increase in racist violence. The primary themes of the movement include race, African history, urban life, the suffering induced by systematic racial oppression and the lasting effects of slavery, as well as uplifting African-Americans by encouraging ethnic pride and positive black identity (Wintz, 2015).

The topics identified by these models are likely to be a mix of positive, encouraging vocabulary (which may reference city life, music and love) and language tackling racial oppression that affected African-Americans in the past and during the time of the movement itself. The vocabulary in this dataset includes slang such as "ain't", "till", and "folks". Some poems, such as James Weldon Johnson's 'An Explanation', are written in a style known as *dialect poetry*, in which alternate spellings and grammatical structures are used to portray speech in a dialect or accent ("*Look heah! 'Splain to me de reason…*"). As opposed to the antiquated alternate spellings featured in the Romantic and Metaphysical datasets, the slang and dialectal language are inseparable from the themes in the poets themselves - their use encourages such linguistic features as part of the general publication of black identity, and attempts should not be made to filter them using stopwords as they represent 'a self-conscious rejection of dominant literary models and of dominant cultural models' (Encyclopedia.com, 2020). The poet Paul Dunbar is said to have lamented to James Weldon Johnson: 'I've got to write dialect poetry; it's the only way I can get them to listen to me' (Baker Jr., 2013, p.38). Some poems in the same dataset do include antiquated words such as 'thy' and 'thee', e.g. James Weldon Johnson's '*Lift Every Voice and Sing*' - '*Shadowed beneath Thy hand, / May we forever stand. / True to our God, / True to our native land.'* We will therefore apply the same stopword filter from the other datasets, as most standard prebuilt stopword collections will not include them.

Methodology

Project Management and Software Development

Our project management methodology and strategies should follow from a successful identification of our aims and objectives, and the specific demands that would qualify the project as a success. These metrics will be based around the extent to which the primary themes of the poetic datasets can be identified in the models using several data gathering techniques. If a model consistently identifies prevalent themes and shows trends in the survey that indicate it as being more thematically consistent than the other model class, it will be considered the most successful. As it is ultimately concerned with a largely subjective area of natural language processing, intuitively interpretable results will indicate success. Project management should ensure enough time is dedicated to take the steps necessary to train, optimise and properly measure the success of the models.

Since our aims and objectives were concretely established, and with a risk assessment taken into close consideration (risk assessment available in appendices), the project's time scale can be managed with a Gantt chart.

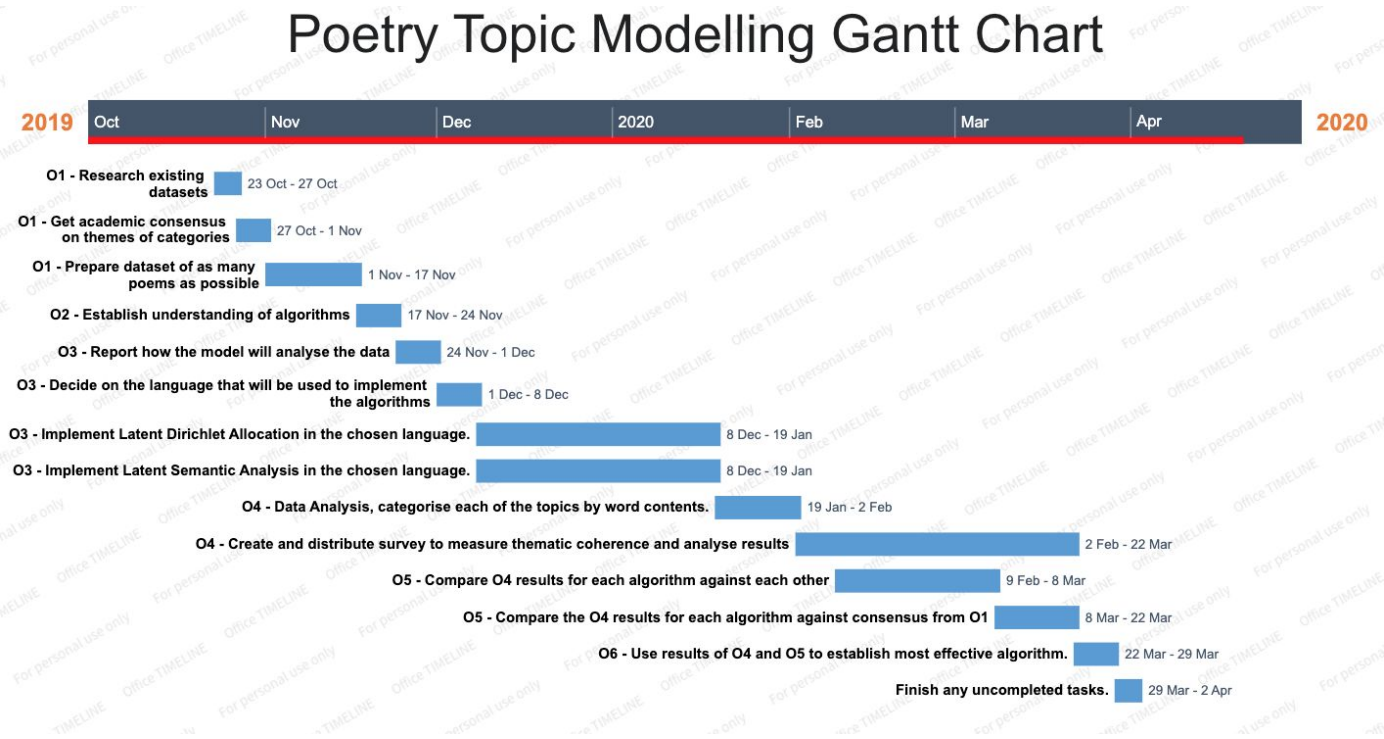


Fig. 4 - Gantt Chart detailing tasks and timescales for project management.

Adopting a plan-based Waterfall methodology would be appropriate for this project's management. Since aims, objectives, evaluation techniques and risks can be identified early on, there is a low chance of mid-development risks emerging (e.g. scope creep) which Agile or Spiral methodologies are designed to prepare for and adapt to (Sliger, 2010). There is little uncertainty in the actual software development aspect of this project, which will involve gathering a dataset, performing appropriate preprocessing and then utilising the prepared corpus in Gensim modelling libraries. These libraries have excellent

documentation and adequate support available through StackOverflow and other such forums, further decreasing the risk of this project's development features being so niche as to make debugging difficult or impossible. The Gantt chart would aid in identifying the development stages used in the Waterfall methodology.

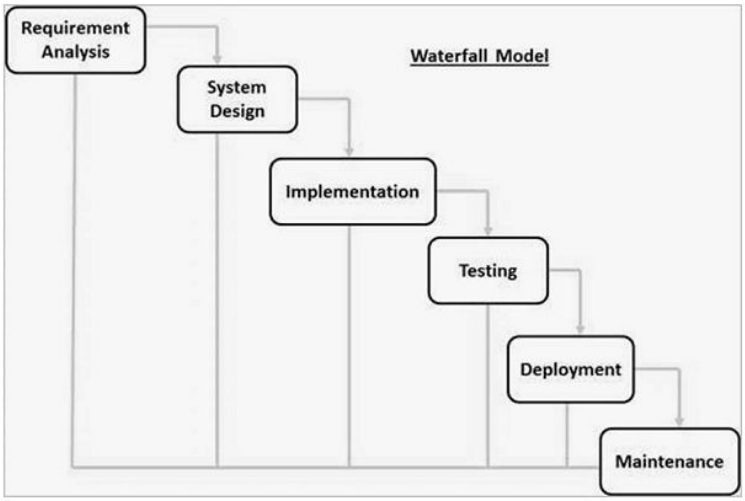


Fig. 5 - Example diagram for the stages of the Waterfall methodology (Tutorialspoint.com, 2020)

Having established that our project can be managed using Waterfall, with the aid of Gantt charts, our actual software development methodology to be utilised during O3 and any further tuning to the artefact will be an Agile-based Scrum-style system with iterative development stages. These stages will be controlled using branches in our Git repository, aiming to have a stage for requirement elicitation and dataset gathering, modelling functional prototypes, then improving further and further on those prototypes with hyperparameter tuning and visualisations for closer result analysis until our model reaches its final state.

Since actual development was planned to take around a month, this was enough for two relatively short sprints. At the end of the first sprint, a deliverable should be prepared in the form of prototypes for each model created using successfully elicited base requirements and datasets. At the end of the second sprint, requirements for what parameters should be tested and tuned as well as what kind of visualisations are planned should be elicited and used to improve the artefact into its final deliverable state, ready for intuitive analysis and comparisons using the research methods. Yan (2019) writes that an Agile methodology for development stages in data science projects generally incorporates sprints of 1-2 weeks in length, as we are doing, and whose benefits for data science involve the planning for each stage, clearly defining tasks and timelines, and planning for demonstrable deliverables at the end of sprints.

Toolsets and Machine Environments

For this project, we will need an effective artefact hosting tool and a programming language in which to write the artefact with all the necessary natural language processing libraries available for model implementation.

Tool	Git Repository using GitHub	Google Drive Cloud Storage	Local Backups
Advantages	• Very widely used for software hosting , large community to offer support for any issues. After development is finished, more likely to be viewed here by other developers than on other hosting platforms. • Simple-to-use GUI (GitHub Desktop) as well as powerful command line features • Unlimited private repositories (as of January 2019) • Effective organisation tools using branching • Version control using commits, pulls and reverts for short and long-term backups and development history	• Easy to use - drag and drop folders and files on the website or client • 15GB free storage and a 5TB hard file size limit, good for hosting large files • Secure backups on Google servers	• Easiest to use, copy and paste project folder onto a USB or a separate folder on the machine • Convenient, no download or pull times
Disadvantages	• Takes time to learn to use effectively with its many features, steep learning curve with the CLI • Doesn't handle large files very well (100MB max per file)	• No version control beyond viewing changes to files, no reverting to or viewing old files • No organisation systems beyond directories	• Extremely insecure - data corruption, accidental deletions, losing USB etc could all compromise the project • Less flexible, no access on other machines

Table 1. - Software hosting tool comparison as a matrix diagram

As is the standard for any software development project, we will be hosting our artefact in a Git repository using GitHub as a hosting service. Since experience with Git has already been attained and there will be no files involved in this project that would go over the 100MB size limit, the disadvantages do not apply, and so we can make great use of the advantages that using Git with GitHub has over the other options. We will effectively utilise its version control tools and branch organisation features to clearly track development stages - we will have a branch for collating the dataset, for prototyping, for parameter tuning and testing, for producing visualisations etc. Should the current version of the artefact incur any catastrophic failures, we can revert to previous versions fairly simply.

Our choices for programming languages came down to three of the most popularly used languages in Data Science projects (Hayes, 2019), which all had libraries aided in implementing the algorithms:

Tool	Python	R	MATLAB
Advantages	● NLP libraries freely available that can be utilised to preprocess, train and visualise the models using NLTK, Gensim, pyLDAvis, Matplotlib and more ● Powerful OOP features for good software development practices ● Code syntax designed to be readable	● Powerful libraries for visualisation and NLP model training including ggplot2, textmineR and LDAvis ● Widely used by statisticians	● Text Analytics Toolbox makes preprocessing and model training simple using functions like ldaModel(), normalizeWords() and removeStopWords() ● Effective built-in visualisation features
Disadvantages	● Running the code on other machines requires re-downloading up to date versions of all the libraries being used, making distribution somewhat harder ● Interpreted language that generally runs slower than most other languages	● Requires use of the R interpreter to run files through specific IDEs such as RStudio, difficult to use with text editors ● Far less syntactic sugar, less readable	● Proprietary software, base version available to students for free but the Text Analytics Toolbox which would be required for this project is an expensive paywalled addition

Table 2. - Programming language and library comparison as a matrix diagram.

Taking this evaluation into consideration we will go ahead with using Python to implement our artefact. Its libraries are freely available, straightforward to use, excellently documented and the code can be written in a lightweight text editor such as VSCode with Python extensions installed to aid with debugging and accessing library documentation. Development would likely take longer and be a more complex process using R, and using MATLAB would be impossible due to the toolkit's paywall.

With the choice of programming language evaluated and justified, we will now evaluate which Python library we wish to use to implement the algorithms.

Tool	Gensim	Scikit-Learn
Advantages	● Implementations for LDA using '*LdaModel*' and the '*LdaMallet*' wrapper and LSA using *'LsiModel'* (Řehůřek, 2019). ● *'LdaMallet'* wrapper uses Gibbs sampling to greatly improve model performance over typical Bayes sampling (ibid.) ● Quantitative comparison metric for both model classes using CoherenceModel pipeline (Röder, 2015) ● Model parameters can control random state seeds, chunk sizes, passes and iterations, and alpha and beta values for LDA	● Implementation for LDA with *decompositions.LatentDirichletAllocation* which uses variational Bayes sampling (Scikit-learn.org, 2020) ● LSA can be implemented using the *decomposition.TruncatedSVD* tool on a tf-idf matrix ● Quantitative performance metric for LDA using Perplexity values ● Parameters for both models include the number of components (topics) and random seeds. For LDA includes alpha and beta values, how often to calculate perplexity for parameter tuning, and the learning rate
Disadvantages	● Less parameters to experiment with to improve LDA performance than Scikit-Learn	● No quantitative methods of measuring LSA performance making comparison between the models' performance more difficult ● No Gibbs sampling to improve LDA performance

Table 3. - Comparison between Python libraries that could be used to implement the algorithms in a matrix diagram.

Since model comparison is a key part of this project, we will be using Gensim to implement our topic models. In addition, despite having slightly less parameters to play with while tuning, the increased performance using Gibbs sampling available in Gensim's LdaMallet wrapper (Řehůřek, 2019a) will prove advantageous.

Research Methods

Qualitative and quantitative data measurements will both be produced as a way of evaluating this project's results. Measuring the coherence values (as defined in the literature review) with Gensim's *CoherenceModel* pipeline for both models gives us a quantitative and directly comparable metric for performance evaluation. Qualitative analysis using the results of a questionnaire with human participants as well as self-reported close study of the results and our corpus will provide insight both into how the algorithms produced their results and a subjective account of the extent to which these results are

intuitively human-interpretable. The results of the questionnaire will also aid in categorising our results into one of the four distinct topic types specified in the literature review.

As this is an atypical application of topic modelling, it is likely that statistical measurements such as coherence may give unusual or unreliable results. Any disparity between the most mathematically coherent algorithm and the most human-approved algorithm as measured from the questionnaire's results will be worth noting - it may even occur that the most technically coherent algorithm is the least thematically consistent. Since this is at its essence a natural language processing project, the qualitative data we measure and self-report will be ultimately more important than quantitative coherence calculations. Both could be illuminating, however, and as such both will be measured. Measuring the average coherence value for each topic will also aid in tuning the hyperparameter controlling the number of topics generated by the models. An optimal value would greatly aid both models' performance, but especially LDA.

Our qualitative data, the questionnaire results, will be nominal. The participants will select which topic seems 'out of place' in order to measure the thematic consistency among the topics, which is more important to measure than any strict semantic relation between keywords. Our quantitative data, the coherence results, will be ratio data as it has a true zero starting point (results being between 0 and 1) and is a number with measurable differences between results.

Establishing an aim for an effective number of participants for this survey took into consideration the ongoing developments of COVID-19 making in-person recruitment impossible but still ensuring enough opinions were collated to identify trends in the survey's results. Creswell (2016) suggests that for research into qualitative measurements that establish a phenomenon of interest, a good aim is 20-30 participants. 20 participants should therefore be our minimum.

Topic interpretation will involve several steps. After analysing both the coherence measurements and the qualitative results of the questionnaire, we will produce a table which will showcase every topic with a justified categorisation, a topic index, LDA prior or topic weight if relevant or similar for LSA, the keywords, and a 'title' we will attempt to manually assign to it (love, mortality, religion etc). We will also visualise the results of both algorithms with the Python library pyLDAvis.

Design, Development and Evaluation

Requirements and Data Collection, Dataset Analysis

Since we worked with poetry as our dataset, our first task was to find or collate a corpus of poems from the Romantic, Metaphysical and Harlem Renaissance movements. In order to get a ballpark figure for a sufficient amount of poems to train a poetic topic model, consideration was given to the two primary examples of similar poetic topic modelling projects by Rhody (2012) and Navarro-Colorado (2018) explored in the literature review: Rhody uses 4500 poems and Navarro-Colorado uses 5000. We may not need such a high amount of poems, as many of these poems are short (Navarro-Colorado's dataset consisted exclusively of sonnets, with a length of 14 lines each) and having a fair amount of poems of a significant length (as can be found in plentiful amounts in Romantic poetry especially) could lead to good results with a smaller dataset. As long as we have enough poems to accurately reflect to at least some degree the main themes of each poetic movement, an aim for several hundred poems from each movement is feasible.

In order to minimise the amount of data scraping and data wrangling that would need to be done when collecting the dataset, we first looked to see if there were any publicly available datasets that were already preprocessed into a CSV format, a standard for data science projects. A dataset of poems from poetryfoundation.org, one of the largest foundations for hosting and showcasing poetry, was found on Kaggle (Bramhecha, 2019) and includes around 14,000 of the 45,000 poems featured on the website. Since there was no way of filtering by poetic movement due to the arrangement of the CSV, a list was collected instead of which poets from the movements we would want to search for in order to extract poems falling into the movements associated with said poets. This list was created for two of our datasets using the names of the poets officially classified into the 'Romantic' or 'Harlem Renaissance' movements on the poetryfoundation.org website (Poetry Foundation, 2020). Metaphysical poetry, as discussed in the literature review, is a much more difficult movement into which to classify specific poems and poets. Ultimately, our list of Metaphysical poets was created using sources ranging from Samuel Johnson's original essay coining the term (Johnson, 1868) to Colin Burrow's choice of Metaphysical poets for Penguin Classics' Metaphysical poetry anthology (2013), eventually leading to a list of 10 poets.

Movement	Poems	Total Word Count	Unique Words
Romantic	392	140,837	16,264
Metaphysical	126	33,321	5,767
Harlem Renaissance	85	18,201	5,192

Table 4. - Details for the size of the datasets.

Romantic poetry was the movement with the highest quantity of poems included in the base dataset, and features by far the most poems and unique words. Despite lacking in an equivalently high amount of

poems and word counts, the other two datasets with their high amounts of unique words should still result in models with fairly good results.

Design

Firstly, we will discuss how the algorithms actually work, their inputs and outputs, and how we can interpret the results.

The original paper which introduced Latent Dirichlet Allocation details the underlying mathematical processes and effective applications of the algorithm. LDA takes a corpus of documents as its input - each word is defined as a unit-basis vector, with each document containing a sequence of words and the corpus being a collection of documents (Blei et. al, 2003). It breaks down each document into a distribution of topics, and outputs the topics as distributions of words across the documents. Every unique word in the entire corpus is assigned a probability that it belongs to each topic, so we will only display the 20 words with the highest probabilities to see which ones the algorithm calculates as being the most likely to co-occur thematically. The output will therefore be a list of topics whose keywords have a probabilistic 'weight' associated with them indicative of its strength in that topic, and the probability it belongs in that topic.

Latent Semantic Analysis takes as its input a term-document matrix or a TF-IDF matrix. Using the singular value decomposition on this matrix, a list of topics will be produced, similarly with a weight associated with each keyword to indicate its strength (frequency of appearance in the topic). This weight is not, however, a probability, and may even be a negative value as it instead indicates the cosine similarity value taken from a query that this term appears in this topic, and therefore the contribution of that word to the topic when taken as an absolute value (Řehůřek, 2019b).

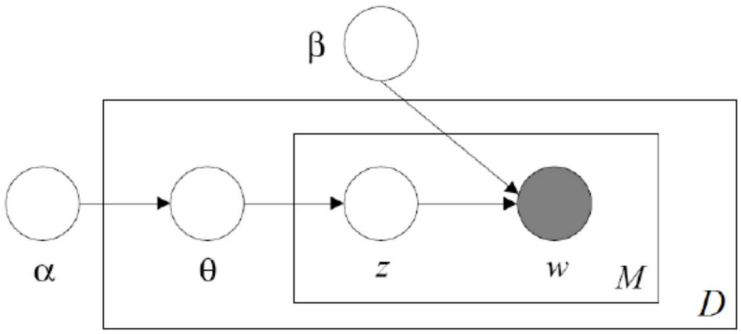


Fig. 6 - Latent Dirichlet Allocation visualised in plate notation (Blei et. al, 2003)

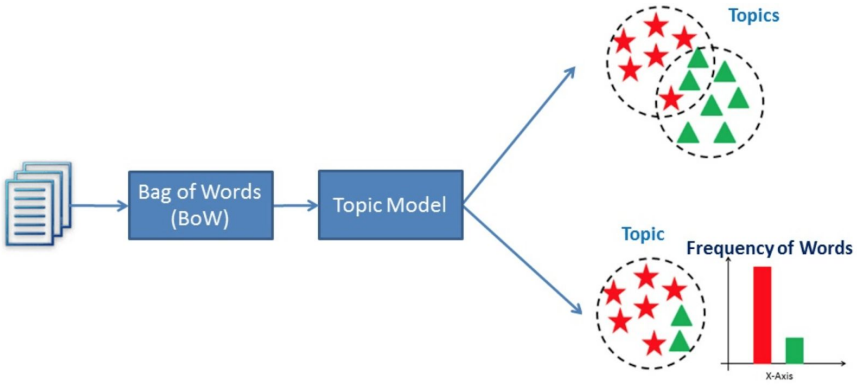


Fig. 7 - Latent Semantic Analysis concept diagram (Navlani, 2018)

The Latent Dirichlet Allocation algorithm begins by randomly assigning each word in each document to one of *K* topics. Then, for each document *D*, it assumes that all topic assignments except for the one currently being explored is correct. Two proportions are then calculated - the proportion of words in document *D* that are currently assigned to topic *T*: P(topic *T* | document *D*) and the proportion of assignments to topic *T* over all documents that come from this word *W*: p(word *W* | topic *T*) (Sullivan, 2017). Multiplying these two proportions gives a new probability to assign the word *W* to a new topic. This is repeated until assignments converge at the point at which the algorithm no longer changes the topics in which each word is placed. Topics and keywords are also affected by priors in the Dirichlet distribution across each topic, where a topic with a higher Dirichlet prior is presumed to be the most 'keyword-dense' in the topic distribution. The top keywords in a dense topic are likely to appear the most across the corpus. The algorithm takes as input two hyperparameters α and β; α affects document-topic distribution, with a higher amount meaning more topics are distributed per document, and β specifies word-topic distribution, with a higher amount meaning some topics will have more strongly weighted keywords than others..

Latent Semantic Analysis can be considered an improvement on calculating the TF-IDF (Term Frequency - Inverse Document Frequency) statistic on a document. It generally takes a TF-IDF matrix as input, but also works with a document-term matrix (a matrix that describes the frequency at which each unique word occurs across the corpus). The purpose of both the TF-IDF and document-term matrices is to measure how important a word is in a dataset by measuring the frequency of its usage across the entire corpus. The Latent Semantic Analysis algorithm improves on these by using a singular value decomposition (SVD) of this matrix to compress it to *k* rows where *k* is the amount of topics we are generating.

The SVD works by splitting X into 3 different matrices. The first is *S*, the term matrix, containing the eigenvectors in the document-term matrix where the rows are documents and the columns are topics. The second is Σ, a diagonal matrix of the square roots of X's eigenvalues in descending order, where each element is the amount of variation captured from each topic. The third is the transpose of *V*, the document matrix, containing the eigenvectors in the term-topic matrix where the rows are terms and the columns are topics. The cosine similarity method with a new document query is then used to generate a matrix (the final topics) of co-occurring keywords. LSA, like LDA, is a *bag-of-words* algorithm - that is to say, the order of the words does not matter, and the entire process is performed mathematically.

The structure of the artefact will be as several Python files. One will collate the datasets, one will implement LDA, one will implement LSA, and there will be a few other files for testing and tuning hyperparameters. Development stages were split into branches in the Git repository:

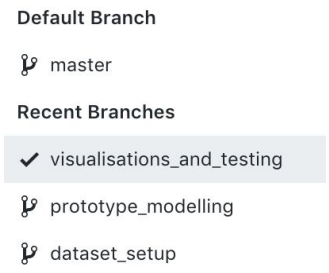


Fig. 8 - Git branches from this project's repository. Each branch covered a different stage of iterative development.

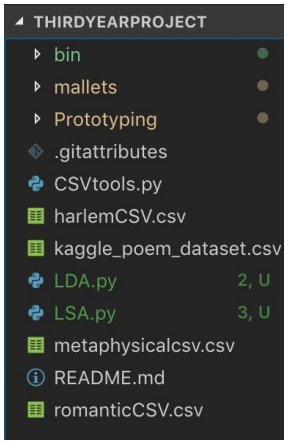


Fig. 9 - File structure, where LDA.py and LSA.py implement the respective algorithms with preprocessing, CSVtools.py creates subsets of the larger Kaggle poetry dataset for each movement, and each of the other CSVs contain the specified movements generated using CSVtools.py. The directories include older versions of the implementations as well as files for use with the Gensim LDAMallet wrapper.

Coding - Preprocessing and Model Training

Preprocessing the corpora involved largely identical steps for both algorithms. The first step involved removing all the punctuation from the poems and converting them to lowercase. This ensured more consistency across the vocabulary in each corpus - ‘Love’ and ‘love’ would not be considered separate words, for example.

Following the preprocessing steps for poetic topic modelling identified in the literature review, stopwords were then identified and removed from the poems. The list of stopwords, which ideally would include any connectives, pronouns or other words which of themselves are not substantive, included all the words in the Python NLTK (Natural Language Toolkit) library’s ‘stopwords’ module as well as some additional ones identified by generating a term frequency table for the corpus. Extended stopwords included, for example, several archaic words used in the Romantic and Metaphysical poetry datasets (‘thee’, ‘thou’, ‘ye’ etc). Before creating a dictionary of numerical tokens to the associated words, we also use Gensim’s Phraser module to build bigram models in our keywords, which uses Markov modelling to form phrases of two words which, should they occur together often enough, are connected with underscores as one term to account for short phrases.

At this stage of preprocessing for topic modelling, lemmatisation or stemming is often performed on the data to remove inflections (e.g. changing all verbs to the same tense) and create more consistency across the keywords in a corpus. However, we did not do this for our poetic topic modelling, as suggested by Navarro-Colorado (2018) in our reviewed literature. We therefore proceeded in creating the dictionary, which gives a numerical identifier to each unique word for passing into the models (then used to convert back into a word for topic output), and creating term-document frequency matrices with Gensim’s doc2bow method, both of which are taken as input by the Gensim functions which train our models.

For our Latent Dirichlet Allocation implementation, we will use Gensim’s wrapper for MALLET’s (Machine Learning For Language Toolkit) LDA functionality, improving the results of the

models using Gibbs sampling, which samples conditional distributions and aids with convergence (McCallum, 2002). For our Latent Semantic Analysis implementation, we will use Gensim's default LSIModel function (LSA and LSI are terms often used interchangably depending on the purpose for which the algorithm is used).

LDA:

```
ldamallet = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=13, id2word=id2word,
                                             random_seed=100, iterations=3000)
```

LSA:

```
# generate LSA model
lsamodel = LsiModel(doc_term_matrix, num_topics=number_of_topics, id2word = dictionary)  # train model
```

Figs. 10&11 - The lines of Python code which build the models using the final parameters.

The parameters taken by these models which have not been expanded upon already will be detailed in the next section, where experiments with tuning parameters appropriately will be recorded and discussed.

Testing and Tuning Hyperparameters

The MALLET LDA implementation we are using takes as input two hyperparameters α and β, which we will leave as the default values of α = 5/k (where k is the number of topics) and β = 0.01.

α affects document-topic distribution, with a higher amount meaning more topics are distributed per document, and β specifies word-topic distribution, with a higher amount meaning some topics will have more strongly weighted keywords than others. There is an optional parameter for the MALLET wrapper's LDA function - '*optimize_interval*' - which aims to optimise the training of the model by adjusting the value of β after a certain amount of training iterations in such a way that results in some topics having very high weights to the keywords and the rest of the weights being fairly low, dehomogenising the weight distribution (Schöch, 2016). Experimentation with *optimize_interval* with our Romantic dataset gives us, as expected, some topics which are very coherent and some which are more opaque, while a model without *optimize_interval* is more consistent. The following topic from this model is a good example:

```
(10,
 '0.025*"lady" + 0.016*"christabel" + 0.010*"geraldine" + 0.008*"sir_leoline" '
 '+ 0.007*"maid" + 0.007*"ladys" + 0.006*"lord" + 0.006*"porphyro" + '
 '0.006*"st_agnes" + 0.005*"oak" + 0.005*"knight" + 0.005*"side" + '
 '0.005*"madeline" + 0.005*"hall" + 0.004*"eyes" + 0.004*"honour" + '
 '0.004*"dame" + 0.004*"shield" + 0.004*"back" + 0.004*"knees"'),
```

Fig. 12 - Topic 10 in a model with 13 topics and using *optimize_interval* with a value of 20 (β changes every 20 iterations)

On the surface, this appears to be a coherent topic about chivalrous knights and ladies. However, it does not entirely reflect themes in the wider scope of the dataset, instead mainly giving high weights to words in the long poem *Christabel* by Samuel Taylor Coleridge, which repeats the keywords many times and features the characters of Christabel, Geraldine and Sir Leoline. By contrast, the same topic from a model without *optimize_interval*:

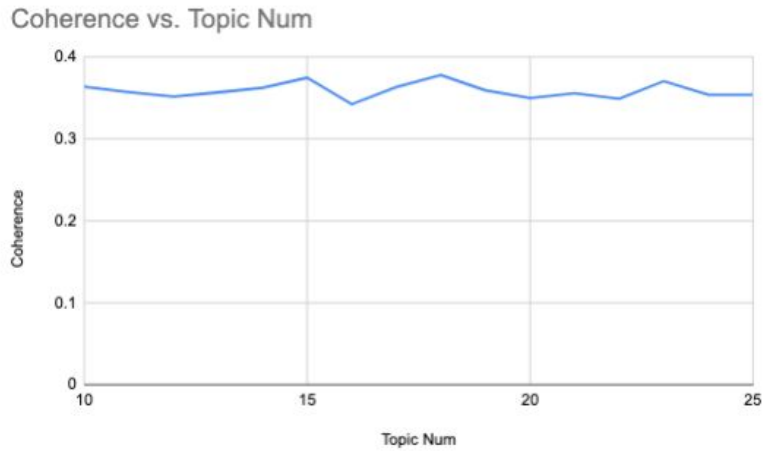
```
(10,
 '0.015*"face" + 0.013*"sweet" + 0.013*"eyes" + 0.013*"lady" + 0.011*"maid" + '
 '0.011*"fair" + 0.011*"hath" + 0.009*"full" + 0.009*"bright" + 0.008*"side" '
 '+ 0.008*"wide" + 0.007*"made" + 0.007*"rose" + 0.007*"words" + '
 '0.007*"night" + 0.006*"christabel" + 0.006*"told" + 0.006*"ah" + '
 '0.006*"pray" + 0.006*"child"'),
```

Fig. 13 - Topic 10 in a model with 13 topics without *optimize_interval*

While still including some of the same keywords, this topic (while having lower weights to its top keywords) seems to capture more of a sense of how women are generally described in Romantic poetry rather than simply including the names of characters repeatedly featured together in concentration.

While *optimize_interval* would be useful if our goal was to discover if the document contains very specific themes which may only feature in small parts of our dataset, as in the example above, we are instead looking to identify wider themes present across the whole dataset. As a result of this, we will not be using it in our final model, and will keep α and β as their default values.

Another parameter to tune for LDA is the number of topics. As mentioned in the literature review, the number of topics greatly affects LDA's performance while it does not affect LSA (beyond not displaying more topics past the set limit). Deciding on the best number of topics for LDA depends on a close look at the generated topics and experimentation with CoherenceModel values for models with different amounts of topics. Calculating the coherence of models with 10 to 25 topics led to the following results:



Topic Num	Coherence
10	0.3642722561
11	0.3576774566
12	0.352172884
13	0.3572959885
14	0.3629940087
15	0.3752479907
16	0.3428679987
17	0.363839282
18	0.37869212
19	0.3596233728
20	0.3504727953
21	0.3562536744
22	0.3495788368
23	0.3709555803
24	0.3545136279
25	0.3545136279

Fig. 14 - Measuring CoherenceModel values over a range of different topic counts in an LDA model

The highest coherence is with 18 topics at 0.379, but variance is extremely small. The difference between the highest and lowest (16, with 0.343) coherence values in this range of topic numbers is only 0.035. Since the numbers have very little trend or difference in the range of the values, we can assume that the CoherenceModel value, while still potentially being useful for comparing each algorithm later, is not a substantive statistic we should rely upon for choosing our amount of topics for a poetic topic model. Instead, we should keep the random seed so that results remain consistent, and self-report a semantically evident and sufficiently high value to represent as many themes as can be identified without much overlap, an issue that worsens with a high amount of topics. Self-reporting indicated that topics began to

repeat at a value of *k=14,* for example - topics 8 and 11 in a Romantic test model speak of nature in almost identical ways:

```
(8,
 '0.036*"sweet" + 0.018*"green" + 0.016*"hear" + 0.016*"spring" + '
 '0.014*"flowers" + 0.014*"meet" + 0.013*"sing" + 0.009*"flower" + '
 '0.008*"bring" + 0.008*"true" + 0.008*"wild" + 0.008*"birds" + '
 '0.008*"winter" + 0.008*"bird" + 0.007*"nest" + 0.007*"fancy" + 0.007*"eye" '
 '+ 0.007*"happy" + 0.006*"clear" + 0.006*"bees"'),
(11,
 '0.011*"joy" + 0.010*"joys" + 0.009*"oer" + 0.008*"vain" + 0.008*"rude" + '
 '0.007*"song" + 0.007*"leaves" + 0.007*"sunny" + 0.007*"sound" + '
 '0.006*"natures" + 0.006*"mossy" + 0.006*"muse" + 0.006*"early" + '
 '0.005*"wild" + 0.005*"love" + 0.005*"heart" + 0.005*"summer" + '
 '0.005*"pleasant" + 0.005*"beauty" + 0.005*"ease"'),
```

Fig. 15 - Topics 8 and 11 from an LDA model with k= 14, showing some duplicate concepts involving nature and seasons.

As a result, and to ensure topics stay compact and readable enough to ensure people finish the study performed with them later, we will use a value of *k*=13 for our amount of topics in every model.

Other parameters to LDA have less of an effect on the results, but should be taken into consideration. Rather than running until a statistically established convergence is reached (which, in LDA, would occur when the change in model perplexity after certain amounts of iterations goes below a certain threshold (Blei et al., 2003)), we will also instead manually define the amount of iterations in case the perplexity statistic behaves in a similarly unusual way to the coherence. As seen in the following figures, there is some difference between topics from a model using the default amount of iterations (1000) and using a higher amount such as 3000 - some keywords and weights change, and it could be argued that the latter topic is more appropriately categorised as ‘sky’ rather than something like ‘waters’.

```
(12,
 '0.013*"sea" + 0.012*"beneath" + 0.011*"life" + 0.011*"spirit" + '
 '0.009*"stream" + 0.009*"calm" + 0.009*"sky" + 0.009*"thought" + '
 '0.008*"moon" + 0.008*"mighty" + 0.008*"dark" + 0.008*"sound" + '
 '0.007*"world" + 0.007*"darkness" + 0.007*"woods" + 0.007*"thoughts" + '
 '0.007*"power" + 0.007*"voice" + 0.007*"waters" + 0.006*"wind"')]
```

Fig. 16 - Topic 12 from the Romantic poetry model after 1000 iterations.

```
(12,
 '0.015*"dark" + 0.013*"world" + 0.010*"sky" + 0.010*"clouds" + '
 '0.009*"stream" + 0.009*"dream" + 0.009*"deep" + 0.009*"beneath" + '
 '0.008*"sea" + 0.008*"mighty" + 0.008*"moon" + 0.007*"stars" + 0.007*"oer" + '
 '0.007*"sound" + 0.007*"wild" + 0.007*"calm" + 0.007*"life" + 0.007*"voice" '
 '+ 0.006*"spirit" + 0.006*"thought"')]
```

Fig. 17 - Topic 12 from the Romantic poetry model after 3000 iterations.

3000 iterations were used to train our final model to ensure that convergence is approached as closely as possible.

Latent Semantic Analysis does not rely on as many hyperparameters, and our model only takes the document-term matrix, number of topics and preprocessed tokenized dictionary as its specified parameters. Its other parameters include whether or not to use parallel execution and whether to use a one-pass or stochastic multi-pass algorithm for training (Řehůřek, 2019b). We will use the default values of serial execution and one-pass algorithm, as tests indicate next to no change in the topics in changing these. In order for more direct comparison with the results of LDA, the amount of topics for LSA will be the same as with LDA, whose performance improves much more with an optimal amount of topics.

Model Results and Visualisations

Categorisation into 'semantically evident', 'semantically opaque', 'large chunk' and 'dialect feature' topics has been performed on the results, as well as attempts to title the topics with one or two-word thematic labels. The full results of the model as well as the self-reported categorisations can be found in an appendix - here general findings will be discussed that indicate the models' level of performance.

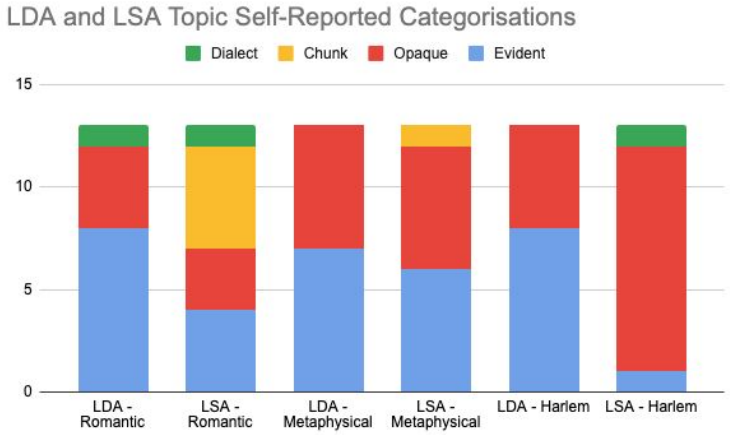


Fig. 18 - Chart showing the proportions of each topic category in each model.

Evident/opaque categorisation was performed based on whether or not there were at least 5 clearly related keywords from the top 20 most heavily weighted keywords in each topic. A relation between keywords could mean that they have connotations relating to the same theme (e.g. topic 5 from the Romantic LDA model, manually labelled 'Afterlife' - '**life, die, live, death, shalt, heavn, blood, bright...**'. Another example is in topic 5 from the Harlem Renaissance LSA model labelled 'Culture' - '**us, africa, soul, people, came, artist, song, color...**). In addition, it could also mean that several keywords are essentially synonyms (e.g. topic 1 from the Metaphysical LSA model labelled 'Mortality' - **'love, world, man, dead, loves, sweet, souls, die, mankind...**').

Similarly, Dialect topics were categorised as such if they contained 5 or more words from a dialect (e.g. topic 6 from the Romantic LDA model labelled 'Scots'- '**meet, wi, ill, thro, body, john, neer, till, bonie, sae, tam im, tho, poor, auld, frae**', and topic 7 from the Harlem LSA model - **'lak, seems, de, went, away, sence, folks, people, aint**').

Chunk topics were more difficult to label. Generally, in these datasets, the kinds of words that would be repeated to the extent where they visibly formed chunk topics would be the names of characters. The LSA models seemed the most susceptible to generating these kinds of topics - the Romantic model contained several examples, such as topics 1 and 2 which both drew heavily from *'Christabel'* by Samuel Taylor Coleridge, and topic 12 from the Metaphysical model which repeated terms from Andrew Marvell's '*A Dialogue Between Thyrsis and Dorinda*' ('**dorinda, thyrsis, souls, elysium, away, way…**').

Every topic can tell us something about how the algorithms are modelling the themes and distributionally related keywords in the corpora, but the ones which will reflect in the most intuitive manner the kind of themes which we may identify as the primary motifs from the academic literature will be the semantically evident topics (Rhody, 2012). LDA consistently has a higher amount of evident topics compared to LSA performed on the same corpus, especially with the Harlem Renaissance dataset, making thematic identification easier.

LSA often outputs duplicate topics. This is especially notable in the final 4 topics in its Harlem Renaissance model, which are all opaque and contain noticeable crossovers in the keywords, meaning it is quite difficult to perform meaningful analysis. The keywords 'hunger', 'dry', 'throat' and 'oats' in these topics are also all coming from the same poem, '*Harvest Song*' by Jean Toomer, almost making them chunk topics. LDA does not suffer from the same issue, and likely owing to choosing a good amount of topics for the size of our dataset, has no duplicate topics in its models, either thematically or in its keywords.

Owing to its higher amount of semantically evident topics, the LDA models appear to be reflecting general themes in the clearest manner for evaluation with respect to the writings on the poetry. As evidenced in the Testing section, where experiments with the *optimize_interval* parameter were performed, it would also appear that sticking with an unchanging value for β has led to there being no identifiable 'large chunk' topics in any of the LDA models, again suggesting that themes present across a large part of the corpus are being reflected rather than being influenced by repetition. The results will be visualised to see how the models reflect themes across the entire dataset. One such visualisation tool for our Python implementation of LDA is the *pyLDAvis* library.

pyLDAvis takes an LDA model (into which we can convert our MALLET LDA model using Gensim's *malletmodel2ldamodel* wrapper) as input in addition to the document-term matrix and dictionary, and outputs the visualisation in an interactable HTML file. Each topic is represented as a circle in a bi-dimensional space with four quadrants. The areas of the circles are set to be proportional to the proportions and relevance of the topics across the entire corpus. The distance between each circle is calculated using Jensen-Shannon divergence between topics, which calculates a quantitative metric for distances (representing differences) between probability distributions (Mabey, 2018). When a topic's circle is highlighted, a histogram of red bars is shown to represent the frequency of the terms' appearance in this topic. A histogram of blue bars is also shown to represent the frequency of the term's appearance across the entire corpus. Additional circles, which can be viewed by clicking on each keyword in a topic, have areas that are proportional to the frequency of the term's appearance in said topic (reflecting the red-bar histogram).

First, we will explore the visualisation of the LDA model for the Romantic poetry dataset.

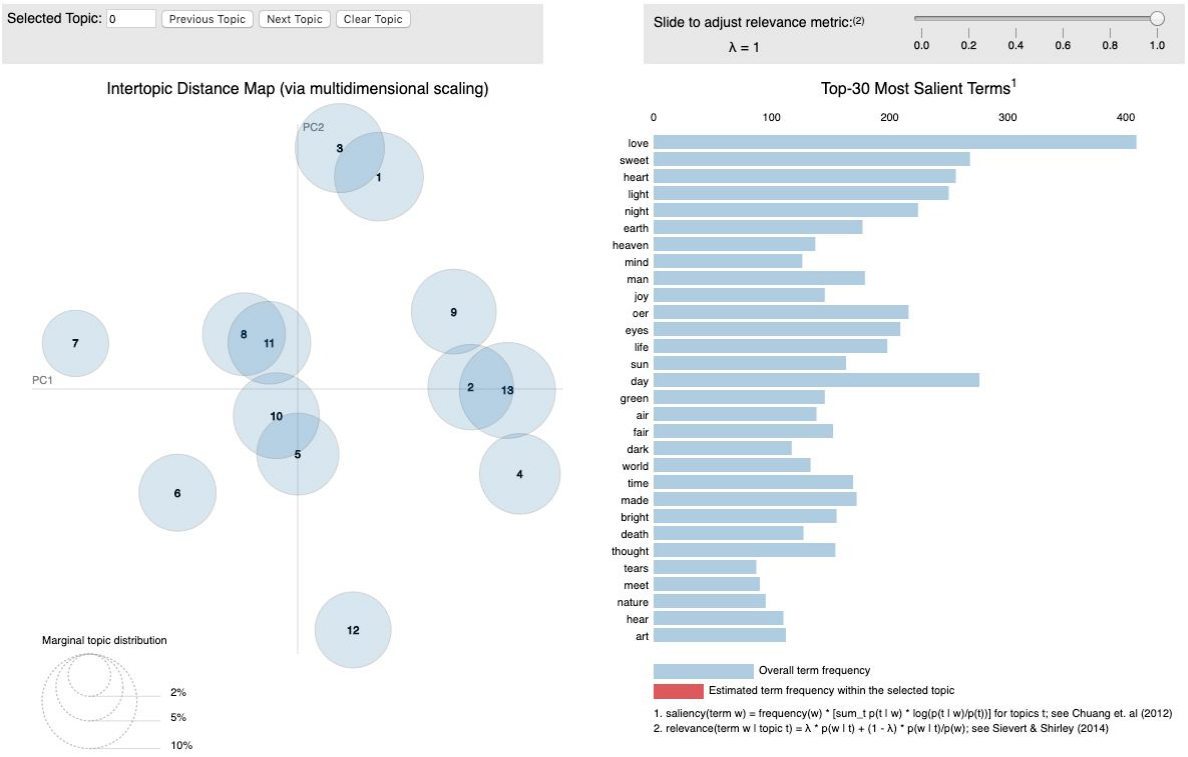


Fig. 19 - pyLDAvis visualisation of the Romantic model.

The 30 most salient terms independent from any particular topic (calculated using the formula at the bottom right) may reflect the themes most relevant to the entire dataset. Two distinct clusters can be seen around the centre of the bidimensional plane and on the right edge of the positive x-axis. This would indicate a high degree of similarity (i.e. a low Jensen-Shannon divergence value) between the topics in the respective clusters. Topics 7 and 12 are notable outliers whose distribution and distance from other topics would imply a thematic uniqueness.

The cluster around the centre of the plane seems to contain topics that speak of and describe people and associated adjectives. The cluster on the right features emotional themes.

Topic in Middle Cluster	Top Keywords
5	oer bear age man sad pride pain poor youth
8	man time years twas head heard child place find
10	soft fair voice form air eyes pale brow silent
11	eyes lady fair maid sweet face full hath rose

Topic in Right Cluster	Top Keywords
2	love heart life tears art thought past smile human
4	oer joys joy till rude early sound wild sunny
9	mind joy nature things thoughts sense friend time day
13	dark world sky clouds stream dream deep beneath sea

Tables 5&6. - The keywords in the clusters in the Romantic visualisation.

One of the most notable outliers, topic 7, is certainly a topic that would be classified as ‘dialect features’, and the inclusion of its keywords in just a small subset of the corpus (Robert Burns’ poems, which were mostly written in Scots) reflects its distance away from the other topics. Topic 12, the other outlier, is a semantically evident topic speaking about joy in nature:

Outlier Topic	Top Keywords
7	meet wi ill thro body john neer bonie sae till tam
12	sweet green hear spring flowers happy sing song fancy

Table 7. - The outlier topics in the Romantic visualistion.

Their distance from the other topics would imply both outliers are semantically or linguistically unrelated to the others. One possible explanation for topic 12’s distance from the others is that it is the only topic that speaks exclusively of nature and joy rather than including many other keywords that overlap into many other topics.

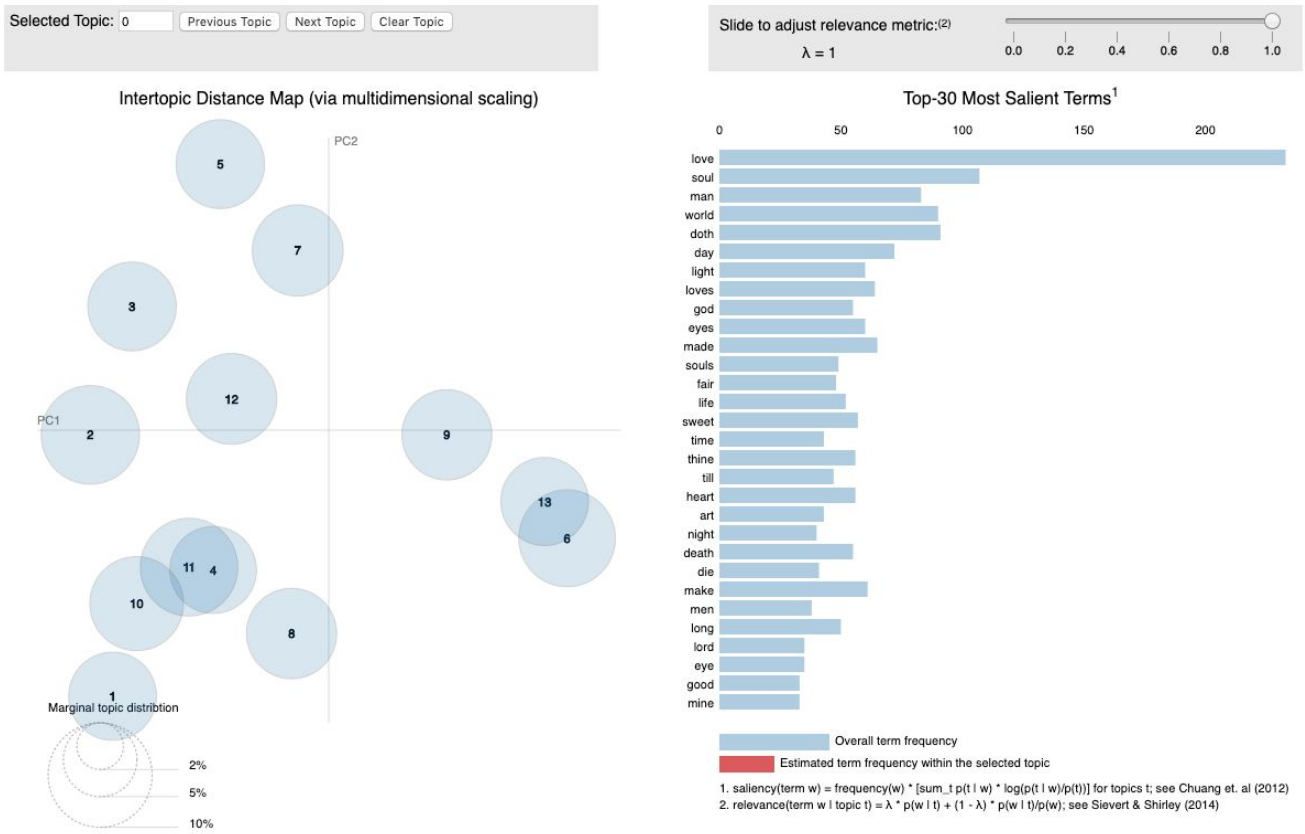


Fig. 20 - Metaphysical dataset pyLDAvis visualisation.

The biggest cluster in our Metaphysical visualisation, and thereby the group of topics most likely to have a relationship between the themes they explore, is at the bottom left, in the third quadrant. The relation between these topics is not as semantically evident as with the clustered topics in the Romantic visualisation. For example, topics 4 and 11, which have a distinct overlap in the diagram:

Topic	Top Keywords
4	place thoughts sun grass sing wound pain lead dew
11	eyes sweet fair thine eye lie earth bed heaven

Table 8. - Metaphysical cluster.

One possible connection is ‘morning’, since topic 4 appears to talk about a natural environment and the morning sun and topic 11 appears to mention a romantic setting in a bedroom.

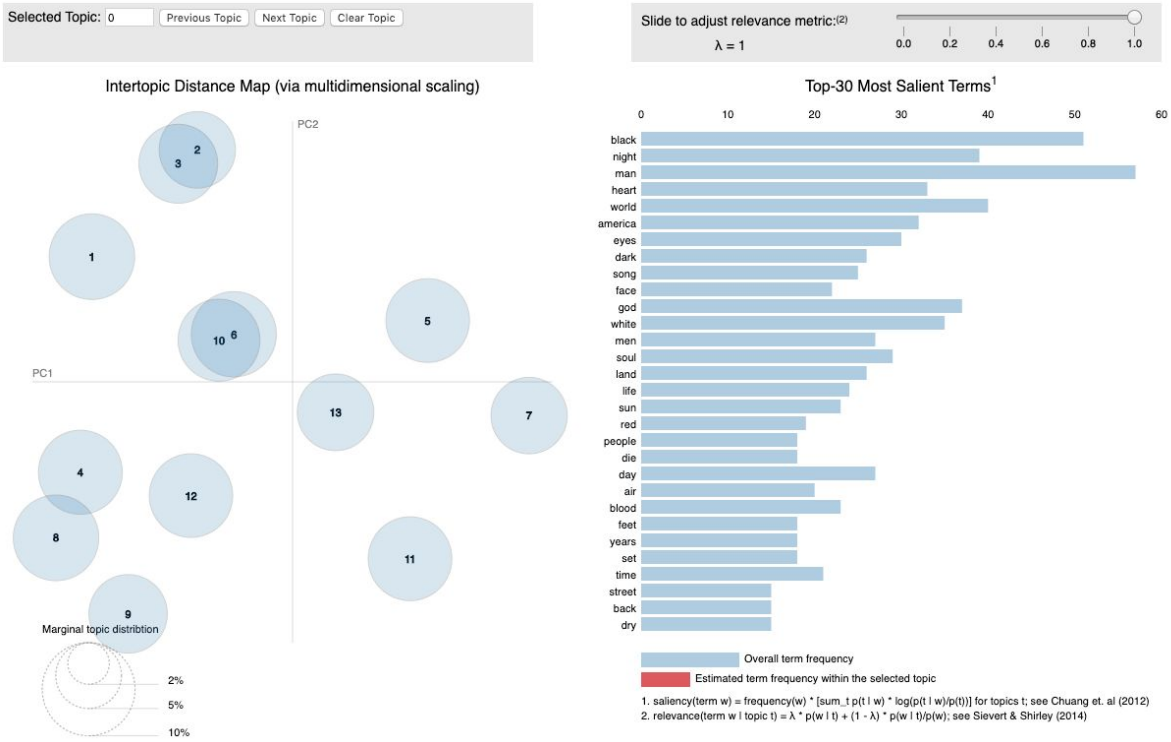


Fig. 21 - Harlem Renaissance dataset pyLDAvis visualisation.

There are no distinct clusters with overlap between several circles in the Harlem Renaissance diagram, though there are a few overlapping pairs. For example topics two and three, which roughly seem to talk about ‘night’ and ‘day’:

Topic	Top Keywords
2	night world people street sky blues rain seems_lak john
3	heart day make mind turned beauty bar grow strong

Table 9. - Harlem cluster

Attempts were made to fit the LSA models into the same kind of visualisation with pyLDAvis, but were unsuccessful. A few discussions were found regarding how this could be done, but with no concrete answers and fatal errors remaining (Matisetorm, 2016). This is likely due to statistics such as Jensen-Shannon divergence being unable to work with models that are not probability distributions. As a result, the only method used in this project to observe the kind of themes explored by LSA are by categorising the results and using CoherenceModel results for an attempt at performance quantification.

Coherence results calculated for each final model are as follows:

	Romantic	Metaphysical	Harlem
LDA Coherence	0.367	0.3204	0.4008
LSA Coherence	0.371	0.345	0.4354

Table 10. - Coherence values for each model using Gensim's CoherenceModel pipeline

Operation

The 3 main Python files were run from the command line. The LDA and LSA files ask the user to input which movement of poetry they wish to explore with a topic model.

```
(base) Toms-MacBook-Pro:ThirdYearProject admin$ python LDA.py
Please specify the poetry movement (romantic, metaphyiscal, harlem)
```

Fig. 22 - User input for the algorithms at runtime.

The visualisations are produced and output as HTML files in the same directory as the folder of the Python script.

Qualitative Data Gathering For a Research Project

Participant Recruitment

Our 'topic intrusion' method of qualitative data gathering was performed with a survey/questionnaire hosted on Google Forms, completed anonymously. Participants were recruited online by having the link to the form sent to them, any in-person recruitment having been rendered impossible by the ongoing developments of COVID-19.

This was an appropriate method for collecting this data. No personal information whatsoever was requested from the participants since information such as subjects studied, experience with reading poetry etc. was not required or measured in this study. The only requirement from the participants was that they were willing and able to read the topics. A brief description of the movements is provided along with the topics presented in order to provide some context to what kind of themes are likely to be explored. Care was taken to ensure the information given did not overtly hint towards the 'fake' topic. Google Forms was a secure way of storing the results electronically, and the results can only be viewed by Google accounts with editing permission.

Since no IP addresses, email addresses or personal identification were collected, and it is made clear that the study only requests anonymous opinions on the models, there was no need for any kind of system for a participant to retract their participation in accordance with recital 26 of GDPR (Intersoft, 2018). Before the main questions, an overview of the study was given informing the participant of what they were going to be asked, and for what purposes. A brief disclaimer was given regarding the fact that any profanity would be censored to ensure there was no chance to cause offence, and the participant was asked to check a box confirming their consent in participating in the study anonymously. The full consent form and questions can be viewed in the appendices.

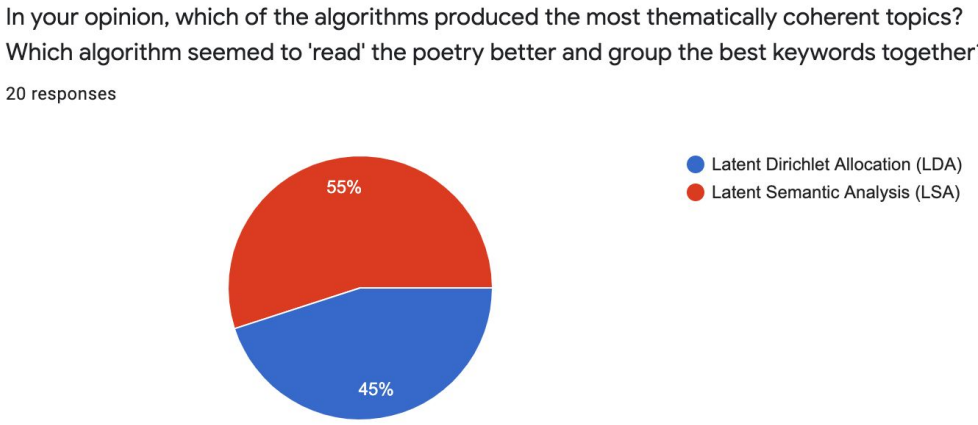
Study Design and Hypotheses

For each model (6 in total, 2 for each corpus of poetry), a set of topics will be presented from which to choose a 'fake' topic that was not generated from this dataset. The fake topic was chosen from a random index of the 50 topics generated by Rhody in *Revising Ekphrasis* (2012). At the end of the study, the participant was asked which algorithm they believe generated the best topics, the overall difficulty of identifying the fake topics for every model, what kind of features highlighted the topics they chose as being fake, and whether they had any additional comments. The decision to draw the fake topics from the results of *Revising Ekphrasis* (ibid.) was based on most applications of topic modelling being on more objective data such as scientific journals, the results of which may have been too obviously separate from our topics. Drawing from a pool of topics that were also generated from poetry means that some of the semantically opaque elements that are expected to appear in any poetic topic model continue to appear even in the fake topics, making a successful process of identifying the 'fake' generally more meaningful and ideally less obvious. As a result of the indices of the fake topics being chosen randomly, there is a chance that some of the fake topics may be more 'obvious' than others. Discussion in the reflection will touch on whether it is believed that was the case, and what effect this could have on the results.

Our hypothesis for the results of the Romantic models, based on the fact these were the ones trained on the largest amounts of poems, is that theoretically these would have the highest success rates of the three corpora when it came to identifications of the fake topic. Between LDA and LSA, based on the self-reported topic classifications from each model, we can hypothesise that LDA will perform the best since it has the most semantically evident topics, thereby making a thematically inconsistent fake topic more easily identifiable. It can be assumed *a priori* that the algorithm with the highest percentage of successful fake topic identifications should also be the algorithm that participants identify as being the most successful at the end of the study, but should participants have a low level of confidence in their responses, this is not guaranteed.

Study Results and Replicability

At the time of reporting these results, 20 responses were recorded, reaching the minimum that was expected to give meaningful results as established in the Research Methods section. The first part of analysing our results will involve comparing the best performing algorithm as reported by the user at the end of the survey against the models whose fake topics were the most successfully identified.



Model	% of Correct Fake Topic Selections
LDA - Romantic	20%
LSA - Romantic	25%
LDA - Metaphysical	65%
LSA - Metaphysical	85%
LDA - Harlem Renaissance	80%
LSA - Harlem Renaissance	35%

Fig. 23 & Table 11. - The best-performing algorithm as chosen by participants compared to their overall performance

LDA average fake selection accuracy: 55%
LSA average fake selection accuracy: 48%

From looking at these figures, we can see a disparity between the algorithm that was self-reported as the most successful by the participants and the algorithm with the highest average fake selection accuracy, going against our hypothesis. There a few reasons why this could have happened:

- Participants were not confident in their answers, either in choosing the fake topics or in choosing the best performing algorithm (this would also explain the almost unanimous opinion by participants that it was difficult to find the fake topic).
- Despite picking the wrong fake topic, participants were more confident with LSA that this was the correct choice.

Whatever the reason, despite LSA being chosen by a slightly higher number of participants (2) as being the best-performing algorithm, there is no overwhelming consensus. As a result, the rest of the results may be more insightful.
As hypothesised, the Romantic models, despite being trained on the most poems (and the LDA model having the most semantically evident topics of them all), had the smallest amount of correct fake topic identifications, largely due to the linguistic uniqueness of the Scots topics leading people to believe it was fake:

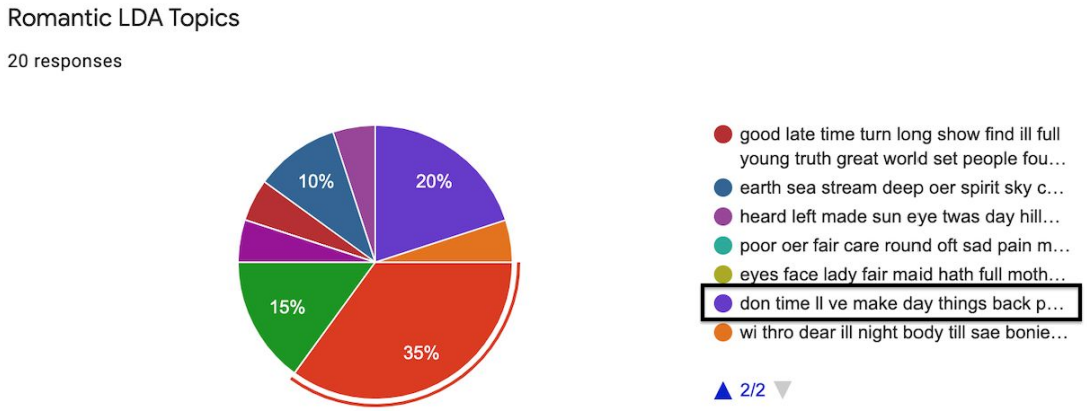


Fig. 24 - Pie chart showing that 35% of participants chose the Scots dialect topic as being the fake in the LDA model, while 20% identified the true fake, highlighted in the list and shown in the chart in dark blue.

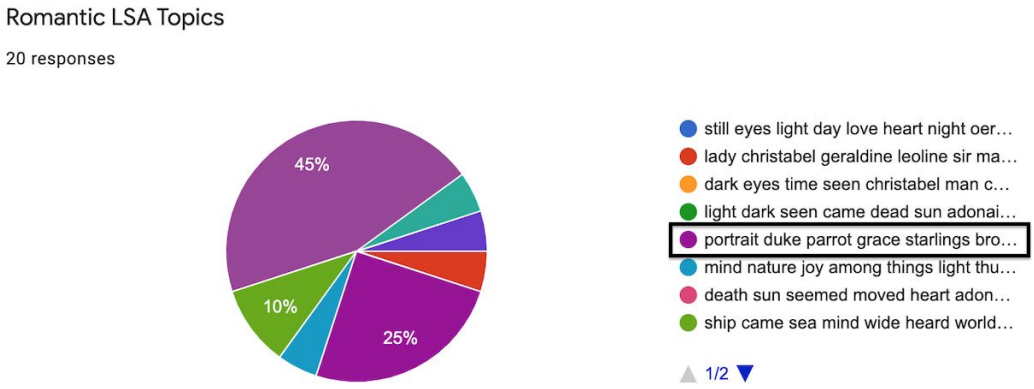


Fig. 25 - Pie chart showing that 45% of participants chose the Scots dialect topic as being the fake in the LSA model, while 25% identified the true fake, highlighted in the list and shown in the chart in a dark purple (bottom right).

This does not mean that the evaluation of the Romantic models was a failure, rather it shows us that noticeable linguistic differences in dialect feature topics catch the eye and seem more obviously thematically inconsistent to participants than a fake topic that, upon close inspection, differs thematically. This differs slightly from the opinion expressed near the end of the survey where only 5% of participants mentioned they prioritised linguistic differences over semantic ones, though since 35% of participants mentioned they took both semantic and linguistic differences into account, the Scots topics could be being referred to.

Participants had more success in identifying the fake topic in the Metaphysical results.

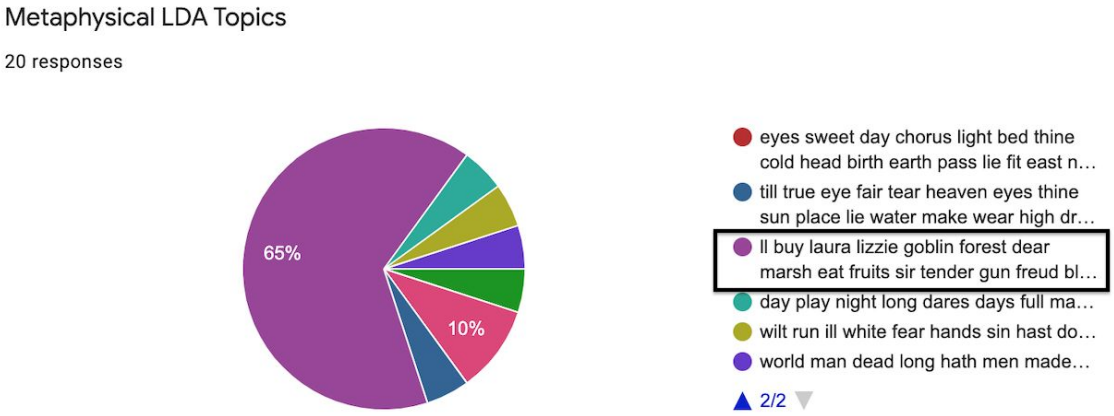


Fig 26. - Pie chart showing 65% of participants successfully identified the fake topic in the Metaphysical LDA model, whose keywords are highlighted.

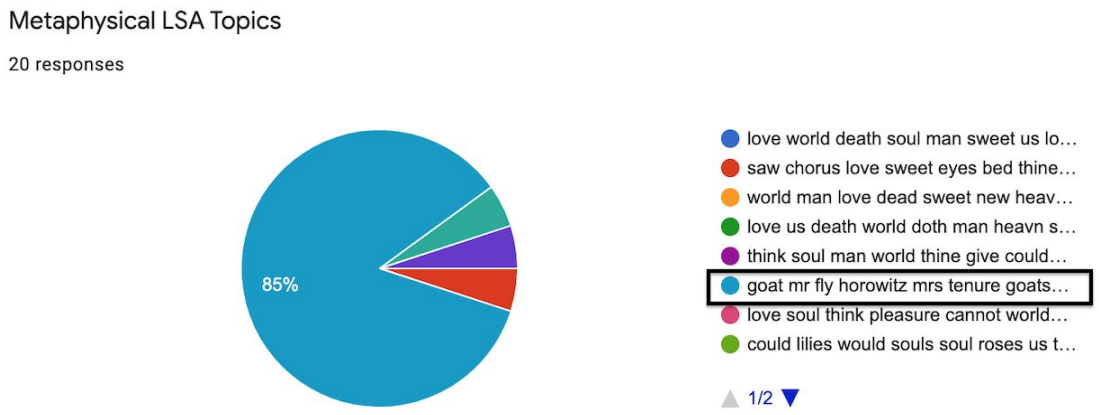


Fig 27. - Pie chart showing 85% of participants successfully identified the fake topic in the Metaphysical LSA model, whose keywords are highlighted.

It can be seen here that topics seemed consistent enough in their language that a large majority of participants were able to identify the fake in both models. This indicates a success, especially for LSA, in presenting themes and maintaining a good level of linguistic and thematic consistency.
Participants struggled with identifying the fake topic in the Harlem Renaissance LSA results, but had more success with LDA:

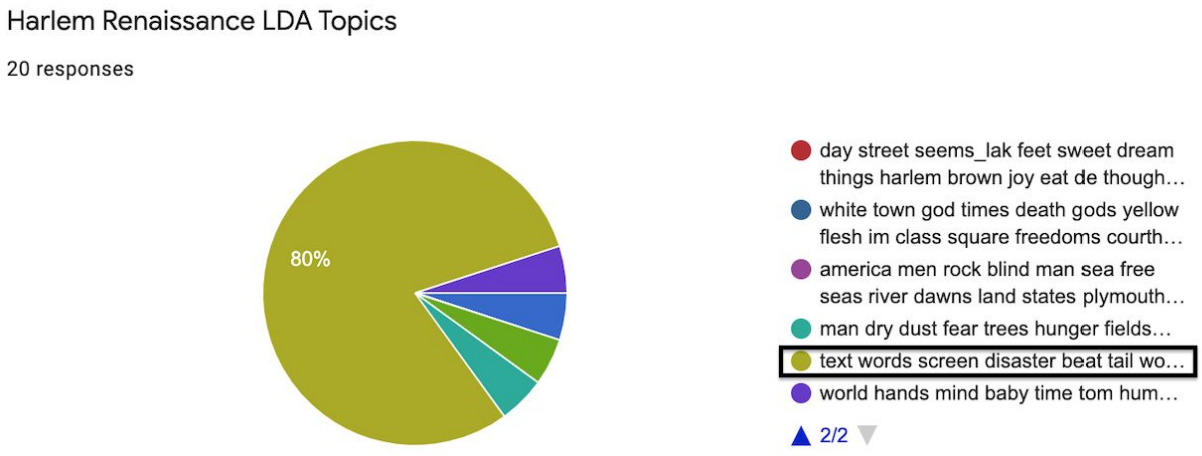


Fig. 28 - Pie chart showing that 80% of participants identified the fake topic (highlighted) in the Harlem LDA model

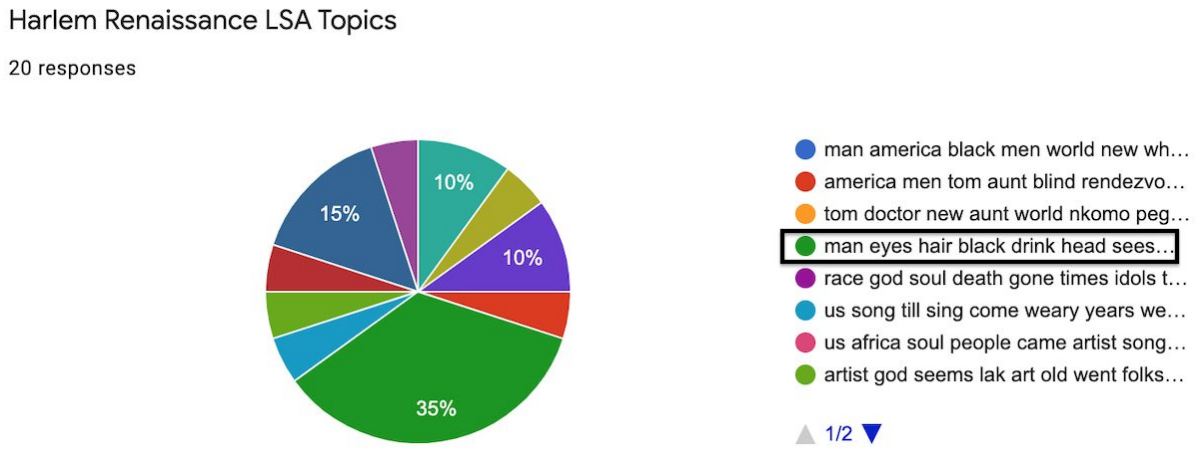


Fig. 29 - Pie chart showing that only 35% of participants identified the fake topic (highlighted) in the Harlem LSA Model

LDA greatly outperformed LSA for the Harlem dataset, but as will be expanded in the reflection, this may be due to LSA being randomly allocated a fake topic that stood out less obviously than the one for LDA, which seemed to talk about modern gadgetry.

Overall, the results would indicate that LDA produced the most thematically consistent topics, but that the fake topics were presented in such a way that participants were not confident in which algorithm performed the best.

Overall Result Analysis

The self-reported topic categorisations as well as the results of the survey would indicate overall that LDA produced the most thematically comprehensible and consistent topics. The CoherenceModel values seemed incongruent with the human-interpreted categorisations and survey results, with the Harlem models which had the least amount of semantically evident topics being computed as the most coherent. The model with the highest CoherenceValue score was the third worst-performing model in the survey and the model categorised with the least amount of evident topics. This would indicate that for comparing model performance as well as measuring the optimal amount of topics, the CoherenceModel is misleading for this particular application. The self-reported categorisations and survey results should therefore be prioritised in the data which they reflect, which is that LDA performed the best.

Leading on from this, we will address the goal identified in the introduction - to identify themes established through academic literature in the poetic datasets using our topic models. In the Romantic dataset, LDA identified themes such as Nature, Joy in Nature, Love, the Afterlife, and Family, successfully reflecting some of the themes identified in the literature review. The visualisations also

proved that clusters of topics were thematically related, reflecting themes on an even broader level. Similarly for the Metaphysical model, themes like Mortality, Love, Death, the Afterlife and Nature were able to be identified in more or less evident topics. The Harlem LDA model reflected such themes as America, Summer, Singing and Colour representing the wide aspects of positive African-American identity being explored.

Not every topic in each of these models corresponds to a pre-established theme in an evident way, but overall there was a good level of success in identifying primary themes in an intuitively readable way using LDA.

Project Conclusion

True to the nature of this project as an unusual application of algorithms typically designed for objective documents, the conclusions that can be made are drawn from mostly subjective sources, with the only quantitative metric (CoherenceModel) having been mostly shown to be misleading or inconsistent. As mentioned in the previous section, the human-interpreted metrics point towards LDA as being the most consistent and interpretable algorithm for identifying themes in these poetic corpora. Both LDA and LSA successfully identify the primary themes in these datasets of poetry to some extent, though a critical reflection of the results as well as what we can gather from each of the comparative metrics leads to the conclusion that LDA does so in a more semantically evident and human-readable manner. Should a participant unfamiliar with these poetic datasets be asked what kind of themes they think seem to primarily feature in the poetry, viewing the results of the LDA topic models would give them a successful idea of the most common and general motifs, assuming a relative accuracy to the self-reported labels made for each topic and to the thematic clusters identified in the visualisations.

In addition to consideration of thematic interpretability from the topics alone, general observations regarding topic modelling with poetry can be drawn from the overall process. Even in the most readable topics, there are always keywords whose reasons for co-occurring with the rest are unclear. The topics should be viewed and interpreted with a close reading of both the dataset and the ways in which the models have pulled themes from the dataset in order to identify trends in the themes, rather than novel interpretations. Rhody (2012) notes that topic modelling with poetry can lead to new insights about poetic datasets not because it works perfectly, but because poetry causes it to ‘fail in ways that are potentially productive for literary scholars’.

The kind of discoveries made from our models that could be illuminating to literary scholars include the noticeable influence of dialect poetry in Romantic and Harlem Renaissance models, as well as the identification of a number of very readable themes pulled from the Metaphysical dataset, a movement somewhat known for its murky definability and lack of obvious motifs. That the Romantic and Metaphysical models have a great overlap in the themes explored could also indicate a shift in the cultural expressions utilised in English poetry after the 1800s, and its use (which remains today) more explicitly for political and social expression.

Reflective Analysis

The main way the performance of our models could have been improved would have been through utilising a dataset with more poems for each movement. As it stands, the only dataset with an amount of poems that could be deemed adequate for this project was the Romantic one, and this is reflected in the results (bar the survey results, which were thrown off by the Scots topics). In my opinion, the LDA model for the Romantic dataset is a genuinely impressive reflection of its themes. Topic 8 even captures a very specific motif in an evident and elegant way - that of finding joy in nature, a theme prevalent in what is perhaps the most famous Romantic poem of all time, Wordsworth's *Daffodils*. LDA's results are impressive overall, whereas LSA's were mostly difficult to interpret. Improving the models would require a process of data scraping some poems in the public domain and adding them to the poems scraped from the Kaggle dump. This was an intimidating and potentially unrealistic task, as it is likely that a different data scraping algorithm would have had to be written for each anthology being explored, which the timescale of the project may not have made possible. This is the main change I would have made, however, and the aspect in which the project is weakest is in the small datasets.

It was a continual frustration throughout the project that finding any kind of external measurements for the models' success and readability was almost impossible. The only directly comparable quantitative metric between the two algorithms was found to be misleading and of no substance, and there are several flaws with the distributed questionnaire. While this method of 'topic intrusion' itself was supported by several pieces of academic literature, its implementation was questionable. How can one choose a fake topic for each model that maintains a consistent level of 'otherness' without factoring bias from the principal investigator? My choice to use a random number generator to choose a topic from another project as the fake removed any potential for bias, but at the cost of a wildly inconsistent level of 'otherness' in each fake. The fake topic for the Harlem LDA model, for example, seemed to speak explicitly of modern technology in a discernible way, and the participants noticed this. Other methods for qualitative data gathering with participants could have been more illuminating had the project been repeated; for example, asking participants which topics they thought were evident and opaque to measure comprehensibility. This would have made justification with academic literature more difficult, however, as the categorisation system used is fairly niche, only being defined in Rhody's project (2012).

Analysis of LSA was not as detailed as that of LDA due to it taking less complex parameters and having no libraries with which to perform intuitive visualisations. Research may have had an element of bias towards LDA's performance as a result of it having more resources with which to delve into analysis.

Finally, there are several directions in which the results of this research could be taken to produce more actionable applications. For example, a 'thematic search' function could be developed wherein keyword overlap in topic models could display poems detected as being similar to that of a user's choice so they can enjoy similar works. Data scraping programs could be written so there were many more documents in each topic, and further research could build on this as a result. Overall, I enjoyed working on this project. It was an interesting multidisciplinary endeavour that I only wish was more easily quantifiable.

References

Baker Jr, H.A., 2013. *Modernism and the Harlem renaissance*. University of Chicago Press.

Bergamaschi, S. and Po, L., 2014, April. Comparing LDA and LSA topic models for content-based movie recommendation systems. In International conference on web information systems and technologies (pp. 247-263). Springer, Cham.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Bloom, H., 1971. *The visionary company: a reading of English romantic poetry* (Vol. 117). Cornell University Press.

Bramhecha, D., 2019. Poetry Foundation Poems. [online] Kaggle.com. Available at: https://www.kaggle.com/tgdivy/poetry-foundation-poems [Accessed 2 April 2020].

Burrow, C. ed., 2013. Metaphysical poetry. Penguin UK.

Byron, G.G.B.B. and Dalling, H.L.B.B., 1837. The Complete Works of Lord Byron. A. and W. Galignani.

Coleridge, S.T (1914). *On Poesy or Art. 1909-14. English Essays: Sidney to Macaulay. The Harvard Classics*. [online] Bartleby.com. Available at: https://www.bartleby.com/27/17.html [Accessed 28 Jan. 2020].

Creswell, J.W. and Poth, C.N., 2016. Qualitative inquiry and research design: Choosing among five approaches. Sage publications.

Cvitanic, T., Lee, B., Song, H.I., Fu, K. and Rosen, D., 2016, January. *LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents.* In *International Conference on Case-Based Reasoning*.

Eliot, T.S., 2014. *The varieties of metaphysical poetry*. HMH.

Encyclopedia.com. (2020). *Dialect Poetry | Encyclopedia.com*. [online] Available at: https://www.encyclopedia.com/history/encyclopedias-almanacs-transcripts-and-maps/dialect-poetry [Accessed 12 Feb. 2020].

Ganegedara, T. (2018). *Intuitive Guide to Latent Dirichlet Allocation*. [online] Medium. Available at: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158 [Accessed 1 Dec. 2019].

Guzman, E., Ibrahim, M. and Glinz, M., 2017, September. *A little bird told me: Mining tweets for requirements and software evolution.* In 2017 IEEE 25th International Requirements Engineering Conference (RE) (pp. 11-20). IEEE.

Hayes, B., 2019. Programming Languages Most Used And Recommended By Data Scientists |. [online] Businessoverbroadway.com. Available at: https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/ [Accessed 2 April 2020].

Hofmann, T., 1999, July. Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann Publishers Inc..

Intersoft, 2018. *General Data Protection Regulation (GDPR): Recital 26 – Final Text Neatly Arranged*. [online] Available at: https://gdpr-info.eu/recitals/no-26/ [Accessed 9 Apr. 2020].

Johnson, S., 1868. Lives of the most eminent English poets: with critical observations on their works. AT Crocker.

La Rosa, M., Fiannaca, A., Rizzo, R. and Urso, A., 2015. Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC bioinformatics*, *16*(6), p.S2.

Mabey, B., 2018. Pyldavis Documentation - Vignette For R Package. [online] Cran.r-project.org. Available at: https://cran.r-project.org/web/packages/LDAvis/vignettes/details.pdf [Accessed 2 April 2020].

Matisetorm, 2016. Gensim Google Group - Trying To Understand The Feasibility Of LSA. [online] Available at: https://groups.google.com/forum/#!topic/gensim/ylfyzuiw0TY [Accessed 4 April 2020].

McCallum, A., 2002. MALLET: A Machine Learning For Language Toolkit. [online] Mallet.cs.umass.edu. Available at: http://mallet.cs.umass.edu [Accessed 3 April 2020].

MALLET. (2018). Topic model diagnostics. [online] Available at: http://mallet.cs.umass.edu/diagnostics.php [Accessed 24 Feb. 2020].

Mastin, L. (2011a). *The History of English - Early Modern English (c. 1500 - c. 1800)*. [online] Thehistoryofenglish.com. Available at: https://www.thehistoryofenglish.com/history_early_modern.html [Accessed 11 Feb. 2020].

Mastin, L. (2011b). *The History of English - Late Modern English (c. 1800 - Present)*. [online] Thehistoryofenglish.com. Available at: https://www.thehistoryofenglish.com/history_late_modern.html [Accessed 11 Feb. 2020].

Navarro-Colorado, B., 2018. On Poetic Topic Modeling: extracting themes and motifs from a corpus of Spanish poetry. *Frontiers in Digital Humanities*, *5*, p.15.

Navlani, A., 2018. Latent Semantic Analysis Using Python. [online] DataCamp Community. Available at: https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python [Accessed 1 March 2020].

Nelson, R. (2019). *Mining the Dispatch*. [online] Dsl.richmond.edu. Available at: https://dsl.richmond.edu/dispatch/Topics [Accessed 22 Nov. 2019].

Poetry Foundation. 2020. Browse Poems | Poetry Foundation. [online] Available at: https://www.poetryfoundation.org/poems/browse [Accessed 10 April 2020].

Řehůřek, R., 2019a. Gensim: Latent Dirichlet Allocation Via Mallet. [online] Radimrehurek.com. Available at: https://radimrehurek.com/gensim/models/wrappers/ldamallet.html [Accessed 7 April 2020].

Řehůřek, R., 2019b. Gensim: Lsimodel Documentation. [online] Radimrehurek.com. Available at: https://radimrehurek.com/gensim/models/lsimodel.html [Accessed 4 April 2020].

Röder, M., Both, A. and Hinneburg, A., 2015, February. *Exploring the space of topic coherence measures*. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

Rhody, L. (2012). *Topic Modeling and Figurative Language*. [online] Journalofdigitalhumanities.org. Available at: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ [Accessed 20 Apr. 2020].

Santilli, S., Nota, L. and Pilato, G., 2017. A Comparison on the Use of LSA and LDA in Psychology Analysis on “Courage” Definitions. International Journal of Semantic Computing, 11(03), pp.373-389.

Schöch, C., 2016. Topic Modeling With MALLET: Hyperparameter Optimization. [online] The Dragonfly's Gaze. Available at: https://dragonfly.hypotheses.org/1051 [Accessed 1 April 2020].

Scikit-learn.org. 2020. API Reference — Scikit-Learn 0.22.2 Decompositions Documentation. [online] Available at: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition [Accessed 3 April 2020].

Sliger, M., 2010. *Goodbye, scope creep—hello, agile*. In PMI® Global Congress.

Srinivasa-Desikan, B., 2018. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

Sullivan, S. (2017). *LDA Algorithm Description*. [online] Available at: https://youtu.be/DWJYZq_fQ2A [Accessed 2 Mar. 2020].

Tennyson, A. (1863). *Flower in the Crannied Wall by Alfred Tennyson, 1st Baron. Edmund Clarence Stedman, ed. 1895. A Victorian Anthology, 1837-1895*. [online] Bartleby.com. Available at: https://www.bartleby.com/246/394.html [Accessed 28 Jan. 2020].

Tutorialspoint.com. 2020. SDLC - Waterfall Model - Tutorialspoint. [online] Available at: https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm [Accessed 1 April 2020].

Wintz, C. (2020). *The Harlem Renaissance: What Was It, and Why Does It Matter?* | *Humanities Texas*. [online] Humanitiestexas.org. Available at: https://www.humanitiestexas.org/news/articles/harlem-renaissance-what-was-it-and-why-does-it-matter [Accessed 12 Feb. 2020].

Appendices

Appendix 1 - Risk Assessment

Objective	Risk Details	Risk Effects	Severity	Likelihood	Precautions
O1 - Collect Dataset	There may be no currently existing datasets for the categories of poetry explored in this project	We will have to manually add poetry details or write a data scraping script on collections of poetry from each movement	Small	Medium	Research in detail what kind of existing datasets are already available on Kaggle or PoetryFoundation.org
O2 - Explain how the model will analyse the data	The mechanisms of the relevant machine learning algorithms may involve advanced mathematics which may require significant time to understand	Should they be too advanced then the descriptions of the algorithms' mechanisms may lack detail or significant understand of the processes	Medium	High	Utilise knowledge from the original papers for the algorithms, online resources that explain step-by-step, and lecture/workshop materials to build a solid understanding
O3 - Successfully implement each algorithm in Python	Since external libraries are being used and relied upon to implement the algorithms, bugs and incompatibilities may occur and be harder to natively debug in Python. It may also be difficult to fine tune parameters since this is an atypical application of the algorithms.	A flawed or buggy implementation of the algorithms may result in misleading results due to incorrectly tuned parameters or other such issues	Medium	Small	Ensure that the implementations of LDA and LSA work with our corpus by using preliminary samples before running the entire dataset, and make good use of online documentation and support networks e.g. StackOverflow and Gensim's Google group forum.
O4 - Gather the results of the algorithms and explain results	Results may be difficult to intuitively interpret, making analyses, comparisons and explanations difficult and/or impossible	Difficulty in properly analysing the results could greatly increase the bulk of the project's core objectives (performance	Large	Medium	Ensure all the steps for O2 have been followed to their fullest extent to ensure that the algorithms are understood well enough that explaining the results is a matter

		and coherence comparison with respect to identifying themes)			of interpretation rather than a ground-up understanding. Consider that semantically opaque topics must have come from somewhere in the corpus and draw conclusions based on what can be seen.
O5 - Compare the outputs given by each of the models with academic literature and survey results	The results may not reflect the academic consensus on the themes of each category of poetry, or there may not even be an academic consensus to be synthesised and explored	If no consensus can be properly defined in the first place it will be impossible to gauge the success of the results	Medium	Small	Establish a consensus early on and then test it against the model results once they have been properly prepared. Should the results not reflect any of the themes, document the process behind this happening so as to still making the conclusions of the project meaningful even if identification of themes was unsuccessful.
O6 - Decide which algorithm produced the most accurate topic model and reflection	There may not be a better performing algorithm, or some other predicament which would make the results of their comparisons difficult or abstract	A key objective of the process that would keep it within the scope of a Computer Science project would not be able to be completed	Large	Medium	Compare results of preliminary samples like in O3 so that an issue like this is spotted soon rather than later in order to cause less risk to the purpose of the project as a whole

Appendix 2 - Full Topic Results and Classifications

Romantic LDA

Topic ID	Keywords and Weights	Label	Category
0	'0.016*"sun" + 0.015*"day" + 0.014*"made" + 0.014*"round" + 0.012*"air" + ' '0.011*"fell" + 0.010*"white" + 0.010*"grew" + 0.010*"light" + 0.009*"sea" + ' '0.009*"lay" + 0.009*"fear" + 0.009*"beneath" + 0.009*"cold" + 0.008*"high" ' '+ 0.008*"ere" + 0.007*"heard" + 0.007*"died" + 0.007*"bare" + 0.007*"snow"'	Daytime	Semantically Evident
1	'0.066*"love" + 0.031*"heart" + 0.016*"life" + 0.015*"tears" + 0.013*"art" + ' '0.012*"thought" + 0.012*"past" + 0.012*"smile" + 0.011*"human" + ' '0.011*"delight" + 0.011*"doth" + 0.010*"dear" + 0.010*"live" + 0.010*"feel" ' '+ 0.010*"give" + 0.010*"hath" + 0.008*"gentle" + 0.008*"hope" + ' '0.008*"mine" + 0.008*"face"'	Love	Semantically Evident
2	'0.027*"light" + 0.025*"night" + 0.022*"earth" + 0.021*"heaven" + ' '0.016*"death" + 0.012*"bright" + 0.010*"till" + 0.010*"dead" + ' '0.009*"sleep" + 0.009*"day" + 0.009*"spirit" + 0.009*"ocean" + 0.008*"weep" ' '+ 0.008*"high" + 0.007*"star" + 0.007*"feet" + 0.007*"sun" + 0.007*"fire" + ' '0.007*"shore" + 0.007*"winds"'	Afterlife	Semantically Evident
3	'0.014*"oer" + 0.010*"joys" + 0.009*"joy" + 0.008*"till" + 0.007*"rude" + ' '0.007*"early" + 0.007*"wild" + 0.007*"sound" + 0.006*"sunny" + 0.006*"eye" ' '+ 0.006*"song" + 0.006*"summer" + 0.005*"leaves" + 0.005*"mossy" + ' '0.005*"free" + 0.005*"muse" + 0.005*"makes" + 0.005*"calm" + 0.005*"smooth" ' '+ 0.005*"health"'	Summer	Semantically Evident
4	'0.012*"oer" + 0.010*"bear" + 0.009*"age" + 0.008*"man" + 0.008*"sad" + ' '0.007*"pride" + 0.007*"pain" + 0.007*"poor" + 0.007*"youth" + 0.007*"hand" ' '+ 0.006*"hour" + 0.006*"stand" + 0.006*"proud" + 0.006*"vain" + ' '0.006*"praise" + 0.006*"fate" + 0.006*"toil" + 0.006*"long" + 0.006*"eyes" ' '+ 0.005*"chain"'	Responsibility	Semantically Opaque

5	'0.013*"good" + 0.010*"great" + 0.009*"young" + 0.008*"world" + 0.007*"ill" ' '+ 0.007*"show" + 0.006*"read" + 0.006*"door" + 0.006*"time" + 0.006*"lie" + ' '0.005*"late" + 0.005*"die" + 0.005*"people" + 0.005*"find" + 0.005*"war" + ' '0.005*"truth" + 0.005*"turn" + 0.005*"morning" + 0.004*"set" + ' '0.004*"sword"'	Adjectives	Semantically Opaque
6	'0.019*"meet" + 0.018*"wi" + 0.012*"ill" + 0.011*"thro" + 0.009*"body" + ' '0.007*"john" + 0.007*"neer" + 0.007*"till" + 0.007*"bonie" + 0.007*"sae" + ' '0.006*"tam" + 0.006*"im" + 0.006*"tho" + 0.006*"poor" + 0.005*"auld" + ' '0.005*"frae" + 0.005*"dear" + 0.005*"mary" + 0.005*"ha_ha" + ' '0.005*"honest"'	Scots	Dialect Features
7	'0.020*"man" + 0.013*"time" + 0.012*"years" + 0.011*"twas" + 0.011*"head" + ' '0.010*"heard" + 0.010*"child" + 0.009*"place" + 0.009*"find" + 0.008*"day" ' '+ 0.007*"left" + 0.007*"misery" + 0.007*"water" + 0.007*"hill" + ' '0.007*"mothers" + 0.007*"born" + 0.007*"father" + 0.006*"thought" + ' '0.006*"poor" + 0.006*"stone"'	Family	Semantically Opaque
8	'0.021*"mind" + 0.017*"joy" + 0.015*"nature" + 0.013*"things" + ' '0.013*"thoughts" + 0.010*"sense" + 0.009*"friend" + 0.009*"day" + ' '0.009*"time" + 0.009*"made" + 0.009*"soul" + 0.009*"silent" + 0.008*"long" ' '+ 0.008*"life" + 0.007*"pleasure" + 0.007*"days" + 0.007*"lake" + ' '0.007*"quiet" + 0.006*"earth" + 0.006*"stood"'	Joy In Nature	Semantically Evident
9	'0.015*"soft" + 0.011*"fair" + 0.010*"voice" + 0.010*"form" + 0.010*"air" + ' '0.010*"eyes" + 0.009*"pale" + 0.008*"brow" + 0.008*"silent" + ' '0.008*"sorrow" + 0.008*"shade" + 0.008*"cloud" + 0.007*"men" + 0.007*"deep" ' '+ 0.007*"golden" + 0.007*"wide" + 0.007*"bright" + 0.007*"faint" + ' '0.006*"sweet" + 0.006*"head"'	Description	Semantically Opaque
10	'0.017*"eyes" + 0.013*"lady" + 0.012*"fair" +	Women	Semantically

	0.012*"maid" + 0.011*"sweet" + ' '0.011*"face" + 0.011*"full" + 0.009*"hath" + 0.009*"rose" + 0.009*"night" + ' '0.008*"side" + 0.008*"eye" + 0.007*"mother" + 0.007*"ah" + 0.007*"child" + ' '0.007*"bright" + 0.007*"christabel" + 0.006*"words" + 0.006*"wild" + ' '0.005*"lay"'		Evident
11	'0.031*"sweet" + 0.018*"green" + 0.017*"hear" + 0.015*"spring" + ' '0.013*"flowers" + 0.012*"happy" + 0.011*"sing" + 0.010*"song" + ' '0.009*"flower" + 0.009*"fancy" + 0.008*"birds" + 0.008*"warm" + 0.008*"die" ' '+ 0.008*"make" + 0.007*"winter" + 0.007*"true" + 0.007*"bring" + ' '0.007*"bird" + 0.007*"leaves" + 0.007*"rest"'	Nature	Semantically Evident
12	'0.015*"dark" + 0.013*"world" + 0.010*"sky" + 0.010*"clouds" + ' '0.009*"stream" + 0.009*"dream" + 0.009*"deep" + 0.009*"beneath" + ' '0.008*"sea" + 0.008*"mighty" + 0.008*"moon" + 0.007*"stars" + 0.007*"oer" + ' '0.007*"sound" + 0.007*"wild" + 0.007*"calm" + 0.007*"life" + 0.007*"voice" ' '+ 0.006*"spirit" + 0.006*"thought"'	Nighttime	Semantically Evident

Metaphysical LDA

Topic ID	Keywords and Weights	Label	Category
0	'0.038*"time" + 0.020*"years" + 0.017*"age" + 0.014*"gold" + 0.014*"sun" + ' '0.014*"found" + 0.013*"land" + 0.013*"glass" + 0.013*"call" + 0.013*"grave" ' '+ 0.013*"late" + 0.013*"king" + 0.012*"free" + 0.012*"star" + 0.009*"awake" ' '+ 0.009*"run" + 0.008*"pass" + 0.008*"graves" + 0.008*"ease" + ' '0.008*"hear"'	Time	Semantically Opaque
1	'0.058*"man" + 0.057*"world" + 0.021*"dead" + 0.018*"doth" + 0.017*"give" + ' '0.015*"worlds" + 0.013*"heaven" + 0.012*"frame" + 0.012*"hath" + '	Mortality	Semantically Evident

	'0.012*"long" + 0.010*"mans" + 0.009*"great" + 0.009*"kind" + 0.009*"hour" + ' '0.009*"thing" + 0.009*"house" + 0.008*"decay" + 0.008*"work" + 0.008*"rich" ' '+ 0.008*"part"'		
2	'0.031*"lord" + 0.022*"dear" + 0.016*"great" + 0.016*"sit" + 0.016*"make" + ' '0.015*"tear" + 0.015*"lay" + 0.014*"straight" + 0.013*"find" + 0.013*"back" ' '+ 0.012*"thing" + 0.011*"meat" + 0.011*"compare" + 0.011*"round" + ' '0.010*"wind" + 0.010*"lines" + 0.008*"past" + 0.008*"face" + 0.008*"poem" + ' '0.008*"sphere"'	N/A	Semantically Opaque
3	'0.022*"place" + 0.021*"thoughts" + 0.019*"sun" + 0.017*"grass" + ' '0.015*"sing" + 0.012*"wound" + 0.011*"pain" + 0.009*"dew" + 0.009*"lead" + ' '0.008*"water" + 0.008*"evry" + 0.008*"heat" + 0.008*"sorrow" + ' '0.008*"grief" + 0.008*"holy" + 0.008*"year" + 0.007*"dream" + ' '0.007*"juliana" + 0.007*"brought" + 0.007*"scythe"'	Emotion in Nature	Semantically Opaque
4	'0.200*"love" + 0.051*"loves" + 0.029*"make" + 0.026*"true" + 0.019*"makes" ' '+ 0.013*"find" + 0.012*"spring" + 0.011*"book" + 0.011*"vain" + ' '0.011*"hearts" + 0.010*"prove" + 0.010*"fit" + 0.010*"hope" + ' '0.010*"whilst" + 0.009*"spheres" + 0.009*"meant" + 0.008*"verse" + ' '0.008*"stars" + 0.008*"mind" + 0.007*"fate"'	Love	Semantically Evident
5	'0.036*"life" + 0.030*"die" + 0.024*"live" + 0.024*"death" + 0.020*"shalt" + ' '0.015*"heavn" + 0.015*"blood" + 0.015*"bright" + 0.014*"full" + ' '0.014*"fire" + 0.013*"joys" + 0.013*"heart" + 0.011*"breath" + 0.011*"home" ' '+ 0.010*"white" + 0.010*"teach" + 0.010*"deaths" + 0.009*"thousand" + ' '0.009*"flame" + 0.009*"strong"'	Afterlife	Semantically Evident
6	'0.022*"mind" + 0.021*"nature" + 0.021*"flowers" + 0.016*"air" + ' '0.016*"pleasure" + 0.015*"green" + 0.014*"heavens" + 0.012*"hand" + ' '0.012*"twas" + 0.012*"roses" + 0.012*"high" + 0.011*"winds" + 0.010*"feet" '	Nature	Semantically Evident

	'+ 0.010*"shade" + 0.010*"plain" + 0.010*"fed" + 0.009*"dwell" + ' '0.009*"find" + 0.009*"flowrs" + 0.009*"trees"'		
7	'0.022*"wilt" + 0.021*"alas" + 0.021*"hast" + 0.020*"sin" + 0.020*"fear" + ' '0.016*"dost" + 0.014*"sea" + 0.013*"reach" + 0.013*"power" + 0.012*"taught" ' '+ 0.011*"lovd" + 0.010*"end" + 0.010*"flesh" + 0.009*"run" + 0.008*"bear" + ' '0.008*"truth" + 0.008*"ways" + 0.008*"honour" + 0.008*"seek" + ' '0.008*"storms"'	Wrath	Semantically Opaque
8	'0.037*"till" + 0.033*"heart" + 0.030*"made" + 0.028*"mine" + 0.028*"good" + ' '0.024*"tears" + 0.024*"hath" + 0.018*"stay" + 0.015*"oft" + 0.014*"wit" + ' '0.013*"music" + 0.012*"fears" + 0.011*"praise" + 0.011*"part" + ' '0.011*"heard" + 0.010*"set" + 0.009*"false" + 0.009*"pleasures" + ' '0.009*"spent" + 0.009*"wear"'	Emotion	Semantically Opaque
9	'0.073*"soul" + 0.050*"doth" + 0.038*"souls" + 0.030*"men" + 0.021*"grow" + ' '0.018*"things" + 0.016*"move" + 0.015*"angels" + 0.015*"hands" + ' '0.013*"thought" + 0.013*"pure" + 0.012*"body" + 0.011*"peace" + ' '0.011*"bodies" + 0.011*"sense" + 0.010*"friends" + 0.009*"glorious" + ' '0.008*"subtle" + 0.008*"voice" + 0.008*"foot"'	Souls	Semantically Evident
10	'0.042*"eyes" + 0.035*"sweet" + 0.034*"fair" + 0.034*"thine" + 0.025*"eye" + ' '0.023*"lie" + 0.019*"earth" + 0.018*"bed" + 0.017*"heaven" + 0.013*"ill" + ' '0.013*"cold" + 0.012*"chorus" + 0.011*"rise" + 0.011*"soft" + 0.011*"ere" + ' '0.011*"crown" + 0.011*"birth" + 0.011*"thyrsis" + 0.011*"sight" + ' '0.010*"sleep"'	Morning	Semantically Evident
11	'0.054*"day" + 0.049*"light" + 0.034*"night" + 0.027*"head" + 0.027*"long" + ' '0.026*"rest" + 0.019*"days" + 0.017*"poor" + 0.016*"glory" + 0.013*"lovers" ' '+ 0.011*"care" + 0.010*"shine" + 0.010*"sacred" + 0.009*"clear" + ' '0.009*"shadows" + 0.009*"dead" + 0.009*"short" + 0.008*"bright" + ' '0.008*"eternal" + 0.008*"play"'	Day and Night	Semantically Evident

12	'0.048*"god" + 0.037*"art" + 0.027*"dust" + 0.026*"made" + 0.022*"fall" + ' '0.021*"beauty" + 0.019*"death" + 0.014*"trust" + 0.013*"grace" + ' '0.011*"face" + 0.011*"leave" + 0.010*"things" + 0.010*"hell" + 0.010*"true" ' '+ 0.010*"learn" + 0.009*"mee" + 0.009*"door" + 0.009*"word" + 0.009*"put" + ' '0.009*"shame"'	Divinity	Semantically Opaque

Harlem Renaissance LDA

Topic ID	Keywords and Weights	Label	Category
0	'0.023*"world" + 0.023*"time" + 0.020*"folks" + 0.017*"hands" + 0.016*"baby" ' '+ 0.015*"tom" + 0.015*"bed" + 0.015*"lost" + 0.013*"bad" + 0.013*"big" + ' '0.013*"rock" + 0.012*"white" + 0.012*"things" + 0.009*"aint_got" + ' '0.009*"hard_times" + 0.009*"child" + 0.009*"aunt_peggy" + 0.008*"wind" + ' '0.008*"legs" + 0.008*"year"'	Family	Semantically Opaque
1	'0.060*"night" + 0.032*"world" + 0.027*"people" + 0.024*"street" + ' '0.023*"sky" + 0.019*"blues" + 0.019*"seems_lak" + 0.019*"rain" + ' '0.013*"john" + 0.013*"power" + 0.011*"born" + 0.011*"hand" + 0.011*"put" + ' '0.010*"peace" + 0.010*"de" + 0.008*"breaks" + 0.008*"dawn" + ' '0.008*"strength" + 0.008*"satchmo" + 0.008*"start"'	Night	Semantically Evident
2	0.047*"heart" + 0.024*"day" + 0.022*"make" + 0.021*"mind" + 0.012*"beauty" ' '+ 0.012*"turned" + 0.010*"grow" + 0.010*"strong" + 0.010*"bar" + ' '0.009*"leaves" + 0.009*"shed" + 0.009*"palm" + 0.009*"eat" + 0.009*"summer" ' '+ 0.009*"laugh" + 0.007*"waters" + 0.007*"writ" + 0.007*"poets" + ' '0.007*"straight" + 0.007*"birds"'	Summery Joy	Semantically Evident
3	'0.030*"soul" + 0.026*"sun" + 0.024*"air" +	Life	Semantically

	0.023*"set" + 0.023*"blood" + ' '0.016*"wood" + 0.016*"fire" + 0.011*"day" + 0.009*"dare" + 0.009*"coal" + ' '0.009*"joy" + 0.009*"tree" + 0.009*"war" + 0.009*"gold" + 0.008*"money" + ' '0.008*"flowing" + 0.008*"pour" + 0.008*"souls" + 0.008*"law" + ' '0.007*"seventh"'		Evident/Semantically Opaque
4	'0.045*"man" + 0.024*"die" + 0.014*"harlem" + 0.014*"doctor" + ' '0.011*"yellow" + 0.011*"sound" + 0.011*"young" + 0.009*"lips" + ' '0.009*"nkomo" + 0.008*"eagle" + 0.008*"im" + 0.006*"ill" + 0.006*"talk" + ' '0.006*"table" + 0.006*"walls" + 0.006*"raw" + 0.006*"bosom" + 0.006*"rid" + ' '0.006*"hair" + 0.006*"pass"'	Disease	Semantically Opaque
5	'0.039*"america" + 0.026*"men" + 0.019*"sea" + 0.017*"deep" + 0.017*"mans" + ' '0.014*"free" + 0.014*"rivers" + 0.012*"blind" + 0.011*"ive" + 0.010*"seas" ' '+ 0.010*"dawns" + 0.010*"freedom" + 0.010*"river" + 0.008*"plymouth" + ' '0.008*"states" + 0.007*"rendezvous" + 0.007*"golden" + 0.007*"valley" + ' '0.007*"worlds" + 0.007*"midnight"'	America	Semantically Evident
6	'0.024*"dry" + 0.020*"dust" + 0.019*"head" + 0.019*"fear" + 0.017*"fields" + ' '0.017*"hunger" + 0.015*"call" + 0.015*"good" + 0.013*"throat" + ' '0.013*"grain" + 0.013*"hear" + 0.012*"soft" + 0.010*"mother" + 0.010*"beat" ' '+ 0.010*"dying" + 0.010*"man" + 0.010*"corn" + 0.008*"stone" + ' '0.008*"hundred" + 0.008*"oats"'	Agriculture/Slavery	Semantically Evident
7	'0.031*"song" + 0.025*"land" + 0.023*"feet" + 0.023*"god" + 0.023*"years" + ' '0.020*"sing" + 0.019*"till" + 0.016*"weary" + 0.016*"light" + ' '0.015*"children" + 0.015*"rise" + 0.013*"high" + 0.013*"tears" + ' '0.011*"taught" + 0.011*"earth" + 0.011*"heaven" + 0.011*"full" + ' '0.011*"liberty" + 0.010*"true" + 0.010*"sang"'	Singing	Semantically Evident
8	'0.026*"life" + 0.023*"back" + 0.022*"boy" + 0.018*"thing" + 0.017*"dont" + ' '0.015*"left" + 0.014*"bright" + 0.012*"truth" + 0.012*"art" + 0.011*"true" '	N/A	Semantically Opaque

	'+ 0.011*"aint" + 0.009*"hard" + 0.009*"door" + 0.009*"round" + 0.009*"days" ' '+ 0.009*"clean" + 0.009*"lies" + 0.009*"poor" + 0.009*"trade" + ' '0.008*"half"'		
9	'0.068*"black" + 0.039*"eyes" + 0.035*"dark" + 0.030*"face" + 0.027*"white" ' '+ 0.026*"red" + 0.020*"stars" + 0.018*"dead" + 0.017*"blue" + 0.014*"pain" ' '+ 0.012*"smoke_king" + 0.012*"color" + 0.009*"open" + 0.009*"artist" + ' '0.009*"hair" + 0.008*"notes" + 0.008*"brow" + 0.008*"fall" + 0.006*"find" + ' '0.006*"bowels"'	Colour	Semantically Evident
10	'0.019*"race" + 0.018*"long" + 0.018*"man" + 0.015*"human" + 0.015*"love" + ' '0.015*"flesh" + 0.012*"bone" + 0.011*"spirit" + 0.010*"cry" + 0.010*"water" ' '+ 0.010*"eye" + 0.008*"speaks" + 0.008*"spring" + 0.008*"wrong" + ' '0.008*"hate" + 0.008*"life" + 0.007*"memories" + 0.007*"hold" + ' '0.007*"makes" + 0.007*"fierce"'	Race	Semantically Opaque
11	'0.017*"god" + 0.016*"times" + 0.015*"death" + 0.015*"town" + 0.013*"ground" ' '+ 0.012*"gods" + 0.010*"lie" + 0.009*"justice" + 0.009*"class" + ' '0.009*"gray" + 0.009*"heads" + 0.007*"fool" + 0.007*"freedoms" + ' '0.007*"alien" + 0.007*"courthouse" + 0.007*"southern" + 0.007*"men" + ' '0.007*"square" + 0.007*"iii" + 0.007*"bones"'	Justice	Semantically Evident/Semantically Opaque
12	'0.023*"home" + 0.018*"trees" + 0.018*"sweet" + 0.018*"dream" + ' '0.016*"brown" + 0.016*"made" + 0.015*"africa" + 0.015*"knew" + ' '0.013*"great" + 0.011*"rich" + 0.011*"heard" + 0.011*"turn" + ' '0.010*"dreams" + 0.010*"grown" + 0.010*"hill" + 0.010*"sugar" + ' '0.010*"seeking" + 0.008*"love" + 0.008*"hell" + 0.008*"silence"'	Home	Semantically Opaque

Romantic LSA

Topic ID	Keywords and Weights	Label	Category
0	'0.152*"still" + 0.145*"eyes" + 0.145*"light" + 0.141*"day" + 0.138*"love" + ' '0.135*"heart" + 0.120*"night" + 0.114*"oer" + 0.108*"sweet" + 0.100*"came" ' '+ 0.098*"would" + 0.098*"life" + 0.096*"old" + 0.096*"made" + 0.093*"earth" ' '+ 0.093*"dark" + 0.091*"death" + 0.090*"bright" + 0.090*"world" + ' '0.087*"thus"'	Love and Time	Semantically Evident
1	'0.348*"lady" + 0.310*"christabel" + 0.197*"geraldine" + 0.169*"leoline" + ' '0.164*"sir" + 0.124*"maid" + -0.113*"dark" + 0.109*"sweet" + 0.106*"well" + ' '0.100*"hath" + 0.096*"ladys" + 0.088*"saw" + -0.087*"earth" + 0.081*"look" ' '+ 0.080*"eyes" + -0.079*"death" + 0.076*"child" + 0.076*"say" + ' '0.076*"tell" + -0.075*"ever"'	Christabel	Large Chunk Topic
2	'0.173*"dark" + 0.172*"eyes" + -0.161*"time" + -0.153*"seen" + ' '0.127*"christabel" + -0.125*"man" + -0.115*"could" + 0.113*"lady" + ' '0.096*"bright" + 0.094*"sleep" + -0.090*"lie" + 0.089*"sweet" + ' '-0.086*"know" + 0.081*"geraldine" + 0.080*"fled" + -0.078*"long" + ' '-0.074*"say" + -0.073*"little" + -0.073*"without" + -0.073*"young"	Christabel	Large Chunk Topic
3	'-0.169*"light" + 0.149*"dark" + 0.135*"seen" + -0.106*"came" + ' '-0.105*"dead" + -0.104*"sun" + -0.103*"adonais" + 0.087*"ever" + ' '-0.085*"tears" + 0.083*"well" + 0.083*"little" + -0.080*"weep" + ' '0.080*"still" + 0.078*"boat" + 0.077*"human" + -0.077*"shadows" + ' '0.077*"poet" + -0.076*"dew" + 0.075*"eyes" + -0.074*"cold"'	Light and Dark	Semantically Evident

4	'-0.219*"mind" + -0.159*"nature" + -0.135*"joy" + -0.119*"among" + ' '-0.117*"things" + 0.110*"light" + -0.105*"thus" + -0.105*"hills" + ' '-0.102*"soul" + -0.088*"hence" + 0.086*"death" + -0.085*"power" + ' '0.081*"world" + -0.080*"eye" + -0.077*"still" + -0.076*"objects" + ' '-0.074*"first" + 0.073*"night" + 0.072*"eyes" + 0.071*"others"'	Emotion in Nature	Semantically Opaque
5	'0.157*"death" + -0.147*"sun" + -0.132*"seemed" + -0.128*"moved" + ' '0.122*"heart" + 0.121*"adonais" + 0.120*"love" + -0.116*"shape" + ' '-0.114*"shadows" + 0.107*"tears" + -0.102*"old" + -0.099*"sea" + ' '-0.096*"within" + -0.092*"fell" + 0.087*"poor" + 0.084*"pale" + ' '-0.083*"ere" + -0.081*"world" + -0.080*"others" + 0.079*"grief"'	Death	Semantically Evident
6	'-0.157*"ship" + -0.142*"came" + -0.132*"sea" + 0.123*"mind" + -0.114*"wide" ' '+ -0.113*"heard" + 0.112*"world" + -0.110*"mist" + 0.110*"light" + ' '-0.102*"agnes" + -0.099*"never" + -0.091*"hand" + -0.090*"still" + ' '0.087*"christabel" + -0.087*"poor" + -0.087*"porphyro" + -0.085*"body" + ' '-0.084*"mariner" + -0.084*"moon" + -0.083*"sun"'	Repeated Words	Large Chunk Topic
7	'-0.181*"poor" + 0.127*"came" + -0.116*"oer" + 0.113*"death" + -0.113*"rude" ' '+ 0.104*"sea" + 0.100*"dead" + -0.099*"old" + 0.096*"adonais" + ' '-0.096*"care" + -0.095*"new" + -0.086*"age" + 0.084*"spirit" + ' '-0.084*"vain" + -0.083*"amid" + 0.080*"ship" + -0.077*"till" + 0.075*"mist" ' '+ -0.074*"joys" + -0.072*"shadows"'	Spirits	Semantically Opaque
8	'0.163*"agnes" + -0.154*"oer" + 0.149*"old" + 0.145*"st" + 0.139*"porphyro" ' '+ 0.116*"madeline" + 0.106*"eyes" + -0.104*"till" + -0.101*"sea" + ' '0.092*"soft" + 0.092*"wide" + -0.090*"sky" + 0.089*"saturn" +	Repeated Words	Large Chunk Topic

	-0.086*"rude" ' '+ -0.083*"wild" + -0.082*"death" + 0.080*"silver" + -0.080*"ship" + ' '-0.079*"song" + -0.078*"poor"'		
9	'-0.452*"wi" + -0.382*"tam" + -0.125*"auld" + -0.120*"night" + -0.117*"thro" ' '+ -0.116*"whare" + -0.116*"ae" + -0.102*"storm" + -0.100*"frae" + ' '-0.096*"na" + -0.096*"mony" + -0.094*"wad" + -0.092*"maggie" + ' '-0.086*"lang" + -0.082*"till" + -0.072*"bonie" + -0.067*"sae" + -0.065*"ah" ' '+ -0.062*"mind" + -0.061*"near"'	Scots	Dialect Features
10	(10, '0.165*"saturn" + -0.149*"love" + -0.124*"agnes" + -0.120*"sweet" + ' '0.113*"poor" + -0.110*"st" + -0.109*"joy" + 0.109*"still" + ' '-0.106*"porphyro" + 0.103*"sad" + -0.098*"rude" + -0.096*"heart" + ' '0.090*"thea" + -0.089*"madeline" + 0.088*"thus" + 0.083*"cannot" + ' '0.077*"space" + 0.072*"sorrow" + 0.071*"gods" + 0.070*"solemn"'	Repeated Words	Large Chunk Topic
11	'0.140*"saturn" + 0.114*"rude" + -0.111*"man" + 0.103*"new" + 0.099*"joys" + ' '0.093*"joy" + 0.089*"earth" + -0.089*"mind" + 0.084*"voice" + -0.082*"poor" ' '+ -0.082*"hand" + 0.078*"meadows" + 0.077*"come" + 0.076*"might" + ' '0.076*"wi" + 0.075*"thea" + 0.074*"gods" + -0.069*"eyes" + 0.069*"sky" + ' '0.068*"ear"'	Cosmos	Semantically Opaque
12	'-0.202*"could" + -0.201*"misery" + -0.184*"thorn" + -0.155*"little" + ' '-0.151*"moss" + -0.118*"day" + -0.116*"love" + -0.114*"pond" + ' '-0.114*"hill" + -0.108*"would" + -0.100*"mountain" + 0.093*"still" + ' '-0.093*"saw" + -0.090*"infants" + -0.089*"heard" + -0.089*"dungeon" + ' '-0.088*"grave" + -0.087*"last" + -0.080*"know" + 0.080*"saturn"'	Nature	Semantically Evident

Metaphysical LSA

Topic ID	Keywords and Weights	Label	Category
0	'-0.296*"love" + -0.251*"world" + -0.219*"man" + -0.188*"soul" + -0.173*"us" ' '+ -0.133*"first" + -0.131*"death" + -0.131*"doth" + -0.115*"long" + ' '-0.114*"new" + -0.113*"must" + -0.113*"made" + -0.111*"dead" + ' '-0.107*"thus" + -0.099*"men" + -0.097*"loves" + -0.096*"sweet" + ' '-0.094*"day" + -0.092*"well" + -0.092*"still"'	Love and Death	Semantically Opaque
1	'0.436*"love" + -0.316*"world" + -0.275*"man" + -0.169*"dead" + ' '0.138*"loves" + 0.133*"sweet" + 0.113*"souls" + 0.110*"die" + ' '-0.100*"mankind" + -0.095*"new" + -0.092*"gone" + -0.089*"long" + ' '0.087*"shalt" + 0.087*"must" + 0.084*"heavn" + -0.084*"worlds" + ' '-0.084*"mans" + 0.082*"bright" + -0.082*"prince" + 0.077*"life"'	Mortality	Semantically Evident
2	'0.325*"love" + -0.315*"saw" + -0.253*"chorus" + -0.247*"sweet" + ' '-0.185*"eyes" + -0.179*"bed" + -0.176*"thine" + -0.164*"day" + ' '-0.155*"sight" + -0.135*"earth" + -0.124*"birth" + -0.121*"thyrsis" + ' '-0.121*"tityrus" + -0.117*"cold" + -0.098*"east" + -0.097*"well" + ' '-0.097*"babe" + -0.090*"night" + -0.085*"light" + -0.084*"tell"'	Sight	Semantically Opaque
3	'0.326*"love" + -0.242*"death" + -0.183*"think" + 0.180*"us" + -0.153*"life" ' '+ -0.152*"shalt" + -0.134*"die" + 0.126*"compare" + 0.126*"meat" + ' '-0.125*"bright" + -0.110*"heavn" + 0.109*"doth" + -0.107*"soul" + ' '-0.106*"shell" + -0.106*"farewell" + 0.104*"sit" + -0.104*"live" + ' '0.092*"version" + -0.091*"fire" + -0.089*"breath"	N/A	Semantically Opaque
4	'-0.669*"think" + -0.203*"soul" + -0.157*"world" + 0.147*"man" + '	N/A	Semantically Opaque

	'-0.104*"give" + -0.102*"thine" + -0.098*"could" + -0.090*"go" + ' '-0.074*"compare" + -0.068*"trust" + 0.065*"long" + 0.065*"new" + ' '-0.062*"state" + -0.062*"meat" + -0.060*"brings" + -0.060*"broken" + ' '0.060*"name" + -0.059*"worm" + -0.059*"forget" + -0.059*"church"'		
5	'-0.192*"meat" + -0.190*"compare" + 0.180*"us" + -0.163*"sit" + ' '0.146*"still" + -0.141*"version" + -0.138*"welcome" + -0.117*"herberts" + ' '0.116*"doth" + -0.115*"serve" + -0.115*"world" + 0.110*"could" + ' '-0.102*"come" + -0.101*"eat" + -0.100*"man" + -0.099*"saw" + -0.096*"poem" ' '+ -0.095*"make" + -0.095*"lord" + -0.094*"1611"	Festivity	Semantically Evident
6	'-0.399*"love" + -0.240*"think" + 0.220*"soul" + -0.176*"us" + ' '0.170*"pleasure" + 0.122*"cannot" + 0.121*"compare" + 0.119*"meat" + ' '0.105*"lord" + -0.102*"loves" + -0.097*"saw" + 0.096*"heavens" + ' '-0.095*"world" + 0.094*"would" + 0.092*"sit" + 0.090*"version" + ' '0.087*"eat" + 0.086*"still" + 0.086*"roses" + 0.078*"guest"'	Love	Semantically Evident
7	'0.235*"soul" + -0.232*"could" + -0.162*"would" + -0.161*"lilies" + ' '0.155*"souls" + -0.136*"tears" + -0.131*"mine" + -0.130*"roses" + ' '0.129*"us" + -0.110*"heart" + 0.108*"none" + -0.107*"fawn" + -0.105*"die" + ' '0.104*"must" + -0.101*"sure" + 0.097*"pleasure" + 0.094*"loves" + ' '0.091*"darst" + 0.091*"religion" + 0.090*"power"	Souls	Semantically Evident
8	'0.416*"soul" + 0.291*"pleasure" + -0.131*"think" + 0.129*"love" + ' '-0.123*"must" + -0.111*"still" + -0.104*"truth" + 0.102*"heavens" + ' '0.101*"thine" + 0.094*"souls" + -0.091*"darst" + -0.091*"religion" + ' '-0.090*"power" + -0.088*"last" + 0.087*"show" + 0.081*"shalt" + '	Heaven	Semantically Evident

	'-0.080*"death" + 0.078*"lie" + 0.078*"us" + 0.075*"new"'		
9	'0.160*"day" + 0.151*"night" + -0.139*"would" + -0.125*"lilies" + ' '0.124*"days" + 0.116*"within" + -0.113*"cannot" + 0.111*"scythe" + ' '0.106*"wishes" + -0.106*"sure" + 0.102*"beauty" + -0.102*"die" + ' '0.095*"things" + -0.094*"thyrsis" + 0.093*"grass" + 0.091*"sun" + ' '-0.085*"roses" + -0.083*"fawn" + 0.083*"head" + -0.082*"thine"'	Cycles	Semantically Opaque
10	'0.482*"us" + -0.289*"love" + 0.182*"souls" + 0.179*"dorinda" + ' '0.164*"thyrsis" + -0.135*"loves" + -0.110*"pleasure" + 0.105*"death" + ' '0.090*"die" + 0.087*"doth" + -0.082*"new" + -0.081*"soul" + 0.079*"elysium" ' '+ 0.078*"things" + 0.077*"nothing" + 0.075*"hath" + 0.073*"makes" + ' '0.068*"go" + 0.068*"flow" + 0.067*"first"'	N/A	Semantically Opaque
11	'0.247*"scythe" + 0.217*"grass" + 0.182*"sun" + 0.177*"mower" + ' '0.143*"dorinda" + 0.139*"thyrsis" + -0.121*"night" + 0.116*"fair" + ' '-0.095*"wishes" + 0.094*"meadows" + 0.088*"dog" + -0.086*"nothing" + ' '0.084*"death" + 0.083*"thus" + 0.083*"way" + -0.082*"head" + -0.082*"souls" ' '+ 0.081*"heat" + 0.081*"juliana" + 0.079*"wound"	Harvest	Semantically Evident
12	'0.309*"dorinda" + 0.293*"thyrsis" + -0.214*"souls" + 0.133*"elysium" + ' '0.130*"away" + 0.121*"way" + 0.118*"till" + 0.117*"ill" + -0.112*"could" + ' '0.108*"day" + 0.099*"give" + 0.099*"love" + -0.095*"far" + -0.087*"still" + ' '-0.083*"men" + -0.081*"first" + 0.081*"go" + 0.077*"wishes" + -0.076*"saw" ' '+ 0.075*"sick"'	Thyrsis and Dorina	Large Chunk Topic

Harlem LSA

Topic ID	Keywords and Weights	Label	Category
0	'0.264*"man" + 0.253*"america" + 0.187*"black" + 0.164*"men" + 0.155*"world" ' '+ 0.138*"new" + 0.136*"white" + 0.114*"mans" + 0.113*"tom" + 0.102*"aunt" + ' '0.100*"god" + 0.099*"land" + 0.098*"eyes" + 0.096*"old" + 0.095*"dark" + ' '0.089*"look" + 0.084*"die" + 0.082*"blind" + 0.081*"day" + 0.080*"way"'	Humans	Semantically Opaque
1	'-0.387*"america" + -0.205*"men" + 0.151*"tom" + 0.133*"aunt" + ' '-0.119*"blind" + -0.102*"rendezvous" + -0.102*"dawns" + -0.100*"saw" + ' '0.096*"peggy" + 0.092*"new" + 0.092*"dark" + 0.089*"baby" + ' '-0.082*"plymouth" + 0.082*"nancy" + -0.082*"december" + -0.082*"pearl" + ' '-0.081*"mans" + 0.078*"way" + 0.078*"god" + -0.077*"look"'	N/A	Semantically Opaque
2	'-0.193*"tom" + 0.176*"doctor" + -0.166*"new" + -0.137*"aunt" + ' '-0.133*"world" + 0.133*"nkomo" + -0.123*"peggy" + -0.105*"nancy" + ' '0.102*"black" + 0.096*"come" + 0.089*"harlem" + 0.089*"since" + 0.083*"die" ' '+ 0.082*"man" + 0.078*"came" + -0.078*"time" + -0.073*"lincoln" + ' '-0.073*"hands" + -0.070*"year" + -0.070*"human"'	N/A	Semantically Opaque
3	'0.191*"race" + 0.189*"god" + 0.133*"soul" + 0.124*"death" + 0.122*"gone" + ' '0.116*"times" + 0.112*"idols" + 0.110*"town" + 0.105*"many" + 0.100*"us" + ' '-0.099*"tom" + -0.095*"america" + 0.091*"five" + -0.090*"aunt" + ' '0.086*"courthouse" + 0.085*"sun" + 0.085*"life" + -0.084*"man" + ' '0.084*"tribal" + 0.084*"tribe"'	Community	Semantically Opaque

4	'-0.300*"us" + -0.219*"song" + -0.140*"till" + -0.139*"sing" + -0.132*"come" ' '+ -0.115*"weary" + -0.103*"years" + -0.102*"went" + 0.099*"race" + ' '-0.093*"stand" + -0.091*"sung" + -0.089*"full" + -0.089*"taught" + ' '-0.086*"sun" + 0.084*"many" + 0.083*"times" + -0.083*"children" + ' '0.082*"town" + -0.081*"true" + -0.081*"land"'	Sing	Semantically Opaque
5	"-0.200*"us" + 0.151*"africa" + 0.139*"soul" + -0.121*"people" + ' '0.113*"came" + 0.113*"artist" + -0.113*"song" + 0.111*"color" + ' '0.111*"black" + -0.109*"god" + 0.108*"sun" + -0.108*"years" + -0.104*"sing" ' '+ -0.103*"new" + 0.102*"art" + 0.098*"knew" + 0.094*"old" + ' '-0.089*"children" + 0.084*"red" + -0.080*"full"'	Community and Culture	Semantically Evident
6	'0.175*"artist" + 0.157*"god" + -0.154*"seems" + -0.154*"lak" + 0.142*"art" ' '+ -0.133*"old" + -0.124*"went" + -0.115*"folks" + 0.110*"us" + ' '0.093*"black" + -0.089*"hard" + 0.088*"true" + 0.087*"blue" + 0.086*"far" + ' '0.084*"face" + -0.082*"bad" + 0.080*"white" + -0.077*"aint" + -0.074*"knew" ' '+ -0.073*"night"'	Art	Semantically Opaque
7	'0.445*"lak" + 0.445*"seems" + 0.187*"de" + 0.168*"went" + 0.166*"away" + ' '0.148*"sence" + 0.135*"folks" + 0.120*"people" + 0.113*"aint" + ' '0.105*"thing" + 0.099*"dont" + 0.096*"bad" + -0.083*"africa" + 0.083*"lost" ' '+ 0.079*"ever" + 0.079*"blue" + -0.078*"knew" + 0.078*"greatgodamighty" + ' '0.078*"feeling" + 0.077*"dat"'	Dialect Poetry	Dialect Features

8	'-0.378*"people" + 0.274*"seems" + 0.274*"lak" + -0.177*"never" + ' '-0.166*"years" + -0.125*"nobody" + 0.120*"song" + 0.115*"de" + ' '-0.109*"folks" + 0.100*"us" + 0.091*"sence" + -0.089*"lost" + 0.086*"away" ' '+ -0.080*"trying" + -0.079*"street" + -0.079*"laugh" + -0.078*"fashion" + ' '-0.078*"cooking" + -0.078*"take" + -0.077*"strength"'	N/A	Semantically Opaque
9	'0.156*"speaks" + 0.148*"race" + 0.139*"folks" + 0.130*"man" + 0.126*"thing" ' '+ -0.126*"lak" + -0.126*"seems" + 0.124*"chain" + 0.121*"ob" + ' '0.121*"wives" + 0.113*"brothers" + 0.110*"human" + 0.110*"good" + ' '0.106*"hunger" + -0.101*"people" + 0.101*"fire" + 0.095*"bad" + ' '0.086*"would" + 0.084*"hear" + 0.082*"take"	N/A	Semantically Opaque
10	'-0.252*"folks" + 0.225*"hunger" + 0.189*"dry" + 0.183*"would" + ' '-0.180*"bad" + 0.163*"oats" + -0.163*"hard" + -0.152*"ever" + ' '0.150*"people" + -0.141*"greatgodamighty" + -0.141*"feeling" + ' '-0.136*"thing" + 0.136*"set" + -0.135*"lost" + 0.133*"throat" + ' '0.126*"grain" + 0.121*"fear" + 0.121*"call" + -0.120*"times" + ' '0.114*"blood"'	Food	Semantically Opaque
11	'0.229*"folks" + 0.181*"hunger" + 0.160*"bad" + -0.159*"race" + 0.139*"hard" ' '+ 0.138*"dry" + 0.138*"ever" + 0.130*"oats" + -0.129*"lak" + -0.129*"seems" ' '+ 0.126*"feeling" + 0.126*"greatgodamighty" + 0.120*"set" + 0.120*"would" + ' '-0.113*"people" + -0.112*"speaks" + 0.111*"lost" + 0.108*"throat" + ' '0.104*"fear" + 0.103*"grain"'	Food	Semantically Opaque
12	'-0.277*"blood" + -0.244*"flowing" + -0.228*"wood" + -0.192*"pocket" + ' '-0.192*"whizzing" + 0.187*"hunger" +	N/A	Semantically Opaque

	-0.153*"set" + -0.141*"would" + ' '0.134*"oats" + 0.118*"fear" + -0.116*"black" + 0.110*"grain" + ' '0.110*"throat" + -0.101*"seventh" + -0.101*"money" + -0.098*"silken" + ' '-0.097*"street" + -0.097*"washington" + -0.096*"n*****" + ' '-0.096*"crudeboned"'		

Appendix 3 - Consent Form and Survey Questions

Topic Model Thematic Evaluation

This questionnaire will consist of eight multiple choice questions as well an optional space for general feedback towards the end. Each multiple choice question will ask you to attempt to identify a fake, 'intruder' topic in a list of topics generated by topic modelling algorithms, each topic containing keywords the algorithm detects as being related to each other. This 'fake' topic will have been generated from a separate dataset and chosen at random from a list.

The content of the topics may contain profanity which has been partially censored.

The results collected from this questionnaire will be stored electronically and viewed only by this project's main researcher, supervisor and markers.

Your personal information will not be collected or published, and this questionnaire will be completed entirely anonymously.

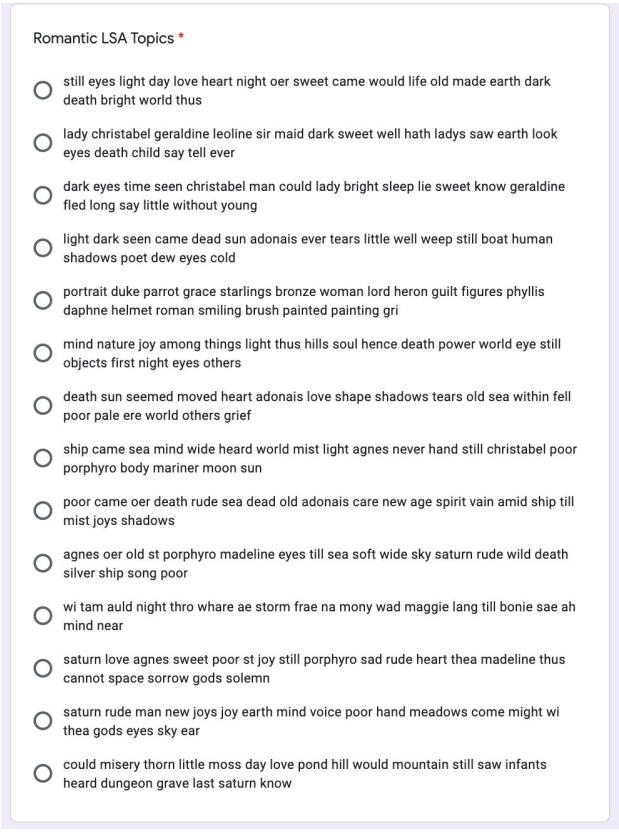
Any questions about this research can be directed by email to 17639930@students.lincoln.ac.uk.

* Required

By selecting 'I agree', you give your consent that the results of this questionnaire can be used for research purposes as part of a Computer Science undergraduate final year project and dissertation. *

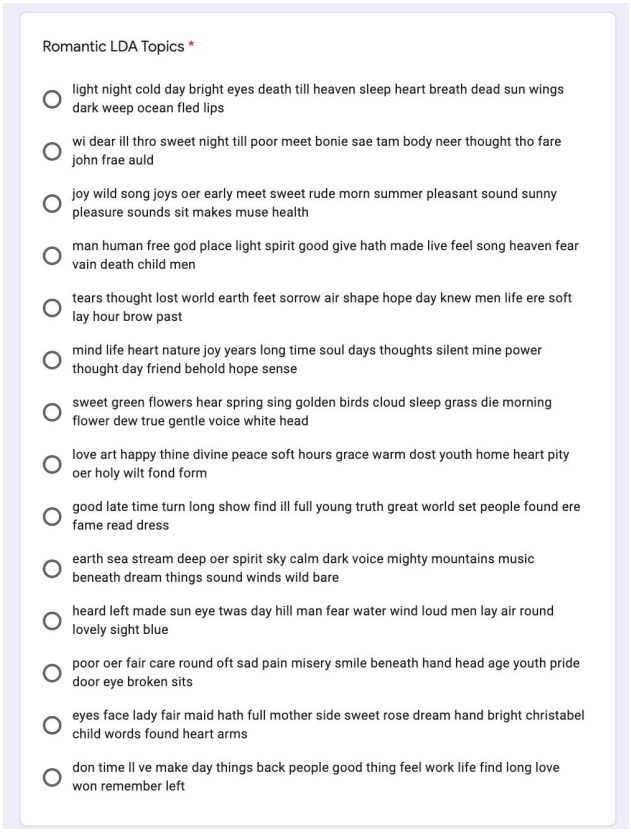
◯ I agree

Note - the order of the keywords in the questionnaire's topics vary slightly from the table in Appendix 2 due to a different random seed being used in the models. The themes covered in the topics remain consistent.



Romantic LSA Topics *

- still eyes light day love heart night oer sweet came would life old made earth dark death bright world thus
- lady christabel geraldine leoline sir maid dark sweet well hath ladys saw earth look eyes death child say tell ever
- dark eyes time seen christabel man could lady bright sleep lie sweet know geraldine fled long say little without young
- light dark seen came dead sun adonais ever tears little well weep still boat human shadows poet dew eyes cold
- portrait duke parrot grace starlings bronze woman lord heron guilt figures phyllis daphne helmet roman smiling brush painted painting gri
- mind nature joy among things light thus hills soul hence death power world eye still objects first night eyes others
- death sun seemed moved heart adonais love shape shadows tears old sea within fell poor pale ere world others grief
- ship came sea mind wide heard world mist light agnes never hand still christabel poor porphyro body mariner moon sun
- poor came oer death rude sea dead old adonais care new age spirit vain amid ship till mist joys shadows
- agnes oer old st porphyro madeline eyes till sea soft wide sky saturn rude wild death silver ship song poor
- wi tam auld night thro whare ae storm frae na mony wad maggie lang till bonie sae ah mind near
- saturn love agnes sweet poor st joy still porphyro sad rude heart thea madeline thus cannot space sorrow gods solemn
- saturn rude man new joys joy earth mind voice poor hand meadows come might wi thea gods eyes sky ear
- could misery thorn little moss day love pond hill would mountain still saw infants heard dungeon grave last saturn know



Romantic LDA Topics *

- light night cold day bright eyes death till heaven sleep heart breath dead sun wings dark weep ocean fled lips
- wi dear ill thro sweet night till poor meet bonie sae tam body neer thought tho fare john frae auld
- joy wild song joys oer early meet sweet rude morn summer pleasant sound sunny pleasure sounds sit makes muse health
- man human free god place light spirit good give hath made live feel song heaven fear vain death child men
- tears thought lost world earth feet sorrow air shape hope day knew men life ere soft lay hour brow past
- mind life heart nature joy years long time soul days thoughts silent mine power thought day friend behold hope sense
- sweet green flowers hear spring sing golden birds cloud sleep grass die morning flower dew true gentle voice white head
- love art happy thine divine peace soft hours grace warm dost youth home heart pity oer holy wilt fond form
- good late time turn long show find ill full young truth great world set people found ere fame read dress
- earth sea stream deep oer spirit sky calm dark voice mighty mountains music beneath dream things sound winds wild bare
- heard left made sun eye twas day hill man fear water wind loud men lay air round lovely sight blue
- poor oer fair care round oft sad pain misery smile beneath hand head age youth pride door eye broken sits
- eyes face lady fair maid hath full mother side sweet rose dream hand bright christabel child words found heart arms
- don time ll ve make day things back people good thing feel work life find long love won remember left

Metaphysical LSA Topics *

- love world death soul man sweet us loves souls eyes must first life thine die saw light heavn day shalt
- saw chorus love sweet eyes bed thine death sight day earth birth tityrus cold thyrsis well life east babe die
- world man love dead sweet new heavn long die worlds shalt gone bright loves mankind think doth saw souls mans
- love us death world doth man heavn still loves two shalt bright life sweet shell farewell fire live day must
- think soul man world thine give could go pleasure new love long eye dead round state forget trust brings broken
- goat mr fly horowitz mrs tenure goats elephant buzz sheep milk trunk carlyle apricots stack nice cleft devil rushes nervous
- love soul think pleasure cannot world heavens still know fair would roses lie saw rest none stand truth lilies must
- could lilies would souls soul roses us tears mine fawn loves die sure heart power must time darst religion none
- soul pleasure think love souls heavens thine must truth show power darst still religion shalt new last lie tears rest
- tear eye sun day would lilies days scythe wishes die night drop cannot sure bright could make diamond rose soul
- us love dorinda thyrsis souls pleasure loves tear eye soul way heart elysium death thus go sun god bed tell
- tear day thine eye scythe us souls drop grass must heaven love mower true shalt make dorinda sad thyrsis bed
- scythe sun grass souls mower us night wishes dorinda fair thyrsis love first play thence wish made dares lovers nothing
- dorinda thyrsis souls away elysium way love ill day far give heart till scythe could cannot wings sick two men

Harlem Renaissance LDA Topics *

- eyes dark face race back long water eye bone spirit arm wild speaks law fire memories dying chain low fast
- sun blood set air wood blue strong time coal tree dare flowing pour fire music turn soil poets pocket southern
- die man doctor sound nkomo leaves mans everlasting eagle wheel ill talk rid bible ice space weeds thin sighed sigh

Metaphysical LDA Topics *

- love loves doth souls make makes body grow move whilst spring made book shadows twere spheres loving beauties bodies theyare
- mine time stay alas roses tears art oft land lilies false grave sighs call lines sighst country image wind fish
- nature find earth mind air green twas tree care trees winds fed root height cruel strange plants gods hay grew
- soul doth heavens meet sea rest fall pleasure show reach state worlds plain kind thought storms fate flesh lay women
- death love souls bright shalt joys light blood thine strong scarce men heavn grace face angels breath farewell lord learn
- things god age thousand years lovd mee lose hundred graves power worst honour dwell rage force wild past scorn tomorrow
- sun grass thoughts wound poor sing scythe meadows juliana feet pain heat golden voice sought hair sound grief fires alas
- life die heart live leave fair soul loves fire make deaths sweet home flame shell art large great wounds hearts
- eyes sweet day chorus light bed thine cold head birth earth pass lie fit east night sight sleep tityrus young
- till true eye fair tear heaven eyes thine sun place lie water make wear high drop shalt wine soft ere
- ll buy laura lizzie goblin forest dear marsh eat fruits sir tender gun freud blades grow beat rapture minnehaha brookdog
- day play night long dares days full made wishes kisses glory set flowers blood vain delight beauty mind cheek lips
- wilt run ill white fear hands sin hast doubt hand hope dorinda dost thyrsis sigh lies day art foot bear
- world man dead long hath men made give worlds part great thing good lost laid house spent prince mans grown

Harlem Renaissance LSA Topics *

- man america black men world new white mans tom aunt god land eyes old dark look die blind day way
- america men tom aunt blind rendezvous dawns saw peggy new dark baby plymouth nancy december pearl mans way god look
- tom doctor new aunt world nkomo peggy nancy black come harlem since die man came time lincoln hands year human

Performance Comparison

How difficult was the process of identifying the fake topic? *

	1	2	3	4	5	
Very Easy	○	○	○	○	○	Very Difficult

In your opinion, which of the algorithms produced the most thematically coherent topics? Which algorithm seemed to 'read' the poetry better and group the best keywords together? *

- ○ Latent Dirichlet Allocation (LDA)
- ○ Latent Semantic Analysis (LSA)

Which differences between topics did you primarily consider when choosing the 'fake' topic? *

- ☐ Linguistic differences (e.g. fake topic seemed to have more archaic or more modern vocabulary)
- ☐ Semantic differences (e.g. the fake topic seemed to mention terms and themes that seemed inconsistent with the rest)
- ☐ Both
- ☐ Other: _____

Please add any additional comments you may have about the performance and results of the two topic modelling algorithms.

Your answer