Human-in-the-loop Topic Modelling

Assessing topic labelling and genre-topic relations with a movie plot summary corpus

Paul Matthews, Computer Science Research Centre, UWE Bristol
April 2019

ABSTRACT

A much-used but not yet mainstream text analysis approach, topic modelling allows the identification of lexical themes for a document collection. Against principles for interpretable AI and sociotechnical design, there are definite strengths from its speed and ability to discover structure, but remain challenges in how results can be interpreted whether this be by analysts, domain experts, or potential end users. Automated coherence and labelling measures go some of the way toward bridging the understanding and trust gap, and user empowerment through visualisation and design intervention is starting to show how the remaining ground might be made up. This study uses topic modelling on a corpus of Wikipedia movie summaries to illustrate challenges and potential. Topic labelling for naive users was found to only be easy in a quarter of cases, and difficulty increased markedly with 100 topics compared to 50. While automated measures suggested 88 topics, the number manageable by users was closer to 50. The unsupervised topic model was compared to the movie genre labels and indicated that the two might work together well to complement genres, match content across genre and highlight within-genre variability. It is suggested that unsupervised models might work better for creativity and discovery than semi-supervised versions.

1 Introduction

Topic modelling is a machine learning technique for inferring structure across a corpus of documents by detecting a number of characteristic topics each containing typical words. While a range of sapproaches and accompanying algorithms now exist, many of those used are based upon

the modelling part being "unsupervised", that is the lexical patterns across documents are detected without reference to any prior human- or content-determined constraints. This has both potential pitfalls and advantages. A naive clustering fails to take account of prior knowledge about the material and about other metadata that can have a clear bearing on its topical variability. Also, being purely statistically-driven, the method can lead to results that are hard for humans to interpret as resulting topics to not necessarily correspond to coherent concepts or add value to an analysis (Bakharia et al. 2016). That said, the approach can be seen in some ways to be objective and to have the potential to uncover useful hidden patterns which might be exploited to improve classification, discovery or theory-building.

> "The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection—a task that is painstaking to perform by hand—and these annotations can be used to aid tasks like information retrieval, classification, and corpus exploration. In this way, topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts." - (Blei 2012)

As a "bag of words" approach that focuses on word occurrence and usually ignores word order (Blei 2012), topic modelling has been seen as too scattergun by linguists:

> "The value of the technique for genuine discourse analysis in thus very limited because the 'topics' are either too general or too incoherent to be useful" (Brezina 2018)

This paper will elaborate some of the tensions described above and ways in which they are being explored by bringing humans and prior knowledge together with the algorithmic results. Decisions such as the number of topics to detect, the labelling of topics and correspondence to human categorisation will be illustrated with reference to a labelled corpus of film plot summaries from Wikipedia.

1.1 Topic model algorithms and typology

Topic models assume that the documents in a corpus contains a number of discoverable topics, and that those topics are distributed across the corpus in certain proportions (Figure 1). The

topics themselves are a probability distribution of words, where a high probability of appearing in the topic make it a characteristic or "key" word for that topic. One of the classic processes in the field, Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), starts by assuming the above and that the words and topics have an estimable distribution. A number of algorithms can then be used to build the posterior likelihoods of words being associated with topics and topics with documents. As an unsupervised model, usually the only inputs to LDA are the number of topics one wants to identify (the K value), and some "hyperparameters" that encapsulate some a priori assumptions or goals for the resulting topic and document assignments. The result of a model run is a list of documents with their topic distributions, and a list of topics with their word distributions.

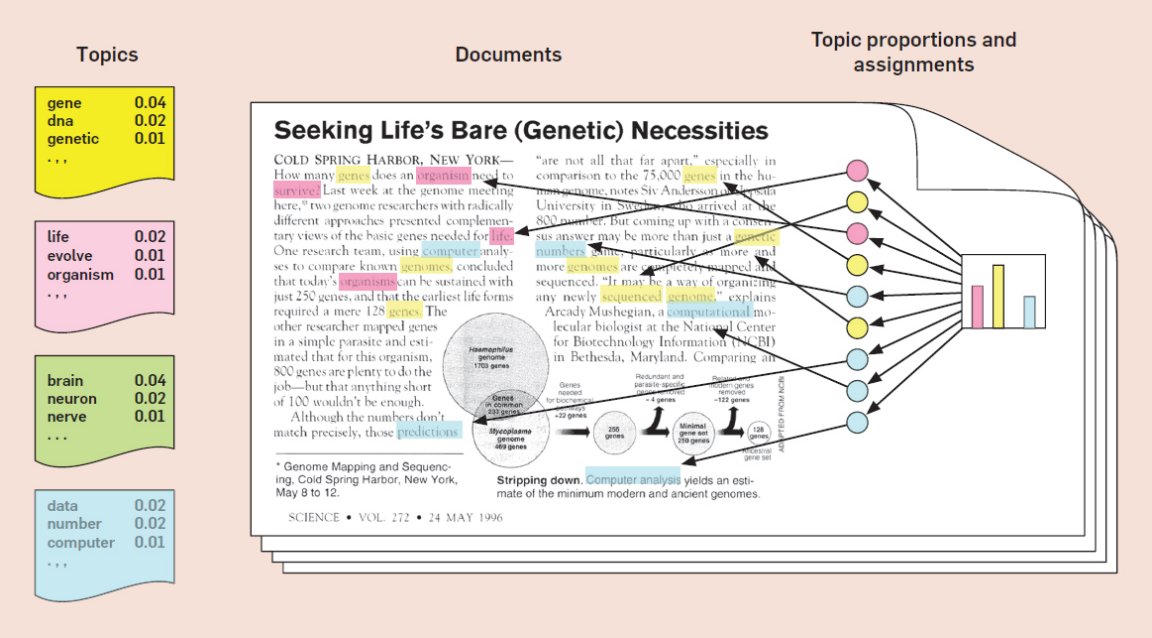

Figure 1: Topics built up from word distributions in a document set (Blei 2012)

While already having a good deal of power in surfacing lexical themes from text, the LDA algorithm has been developed in a number of ways for specific applications and to address certain limitations. The problem of not knowing the optimal topic number, for example, was addressed through the development of the Hierarchical Dirichlet Process (HDP) model, which adds a slot for the topic number and outputs this together with word-topic and topic-document distributions. Notably, the process for optimising the topic number is based on lexical features rather than any

particular user goals. A different hierarchical approach, hLDA, is used to progressively output increasing number of topics with each run.

A further family of models make use of metadata and prior knowledge to guide the process. The labelled LDA, the author-topic model and an increasing number of domain-specific models do this by having the LDA as just one subsystem within the model. These are therefore semi-supervised models as they provide as guidance some labelled training examples of different document types. The Stuctural Topic Model (STM) (M. E. Roberts, Stewart, and Tingley 2018; Roberts et al. 2013) is used in the analysis used in this study and enables both pure unsupervised modelling and also the analysis and visualisation of interactions between metadata and topic distributions.

So while more "supervised" versions of topic modelling exist, this paper will focus largely on the unsupervised versions, in order to better draw out the pitfalls and potential advantages for human users and to compare the results with categories applied by people to document collections.

1.2 Interpretable AI and human factors computing

Artificial intelligence and specific machine learning approaches are often criticised for their opacity of operation and the lack of human agency involved. Doshi-Velez and Kim (2017) argue that machine learning explanation is needed when algorithmic and application-oriented objectives are not aligned or there is a need for decisions to be made over the trade offs between such objectives. They provide a continuum of evaluation approaches toward interpretability that range from formal proxies for interpretability not involving users, through the use of simplified tasks for real users to test, to the most desirable approach of real humans using the technology for real tasks (figure 2), though cost and time increase along this continuum.

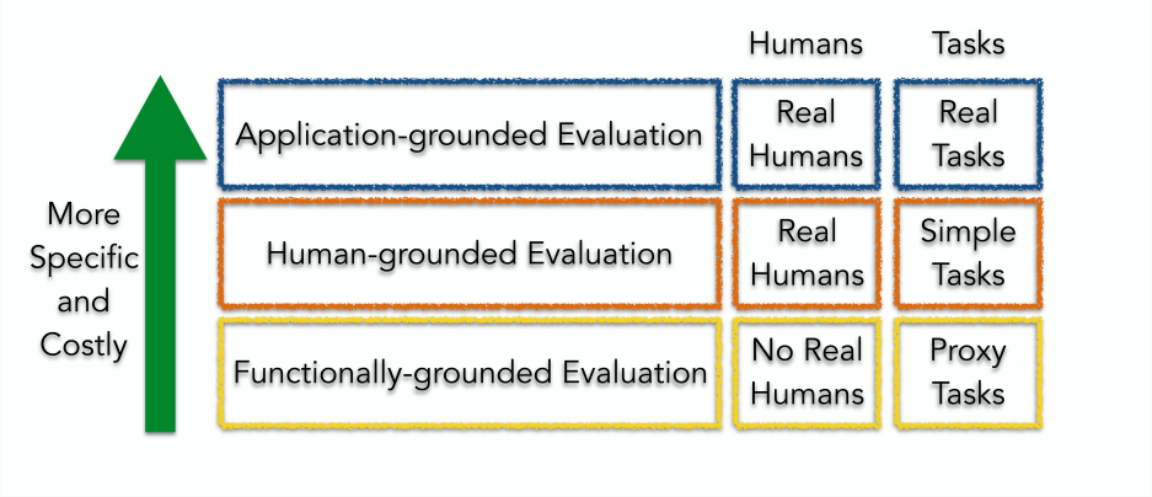

Figure 2: Taxonomy of evaluation approaches for interpretable AI (Doshi-Velez and Kim 2017)

For humans to work effectively with AI-based tools, it is important to develop design guidelines aimed to optimise the mutual relationship. Read et al (2015) sought to develop design principles for sociotechnical systems based on a unification of sociotechnical systems theory and cognitive work analysis. Their derived methodological attributes were ranked by subject matter experts and the top five resulting attributes were: creative, holistic, structured, efficient, iterative, integrated and valid. Systems need to allow the scope for opportunistic and exploratory analysis. The outputs need to justify the expenditure of time and energy. System design should be coherent, with structure enabling communication and means-end accountability. At the same time systems should accommodate changes over time as understanding grows.

1.3 Functional evaluation: Automated measures

At the automated end of topic model evaluation, approaches started at the model-centric level but have moved more recently toward measures which better approximate human judgement. For some time the standard evaluation methods for topic modelling was “held out likelihood” (or log-likelihood), wherein a subset of documents which were not used to train the model are used to predict topical composition. A “perplexity” metric can then be used to score how well the existing model fits these new texts. This tends to give a lower score for higher numbers of topics, which provide a closer fit to the training set. However, it does not correspond well to interpretability

and may in fact be antagonistic to human understanding (Jacobi, Van Atteveldt, and Welbers 2016).

Coherence is a more recent and important interpretability metrics for topics (Röder, Both, and Hinneburg 2015). A popular method of evaluating coherence is pointwise mutual information (PMI), which scores words within a resulting topic to their likelihood of co-occurrence in a reference corpus such as Wikipedia (or within the corpus text itself). The overall coherence of a topic is then a sum of these scores. The score may be further improved by normalisation to a score in the range from -1 to 1, where -1 means the words are never found together and 1 means they always are.

While coherence may tend toward more homogeneous topics, other metrics sucyg as FREX can be used to rate the exclusivity of topics based on the mean of composite word exclusivity and frequency (M. E. Roberts, Stewart, and Tingley 2018). This is quite an important balancing metric as the topic modeller usually wants the topics to be in some way distinctive.

1.4 Human evaluation: Which humans and where in the loop?

People may be involved in the topic modelling process at generation, evaluation and end use stages. At the generation stage, the typical actor is the data scientist / information specialist who has access to and an understanding of the data, tools and methods needed to run the model. Modelling consists of importing text into the corpus, some processing and preparation of the text and then fitting the model (Figure 3). There are then some further steps to evaluate, understand and visualise the model output.

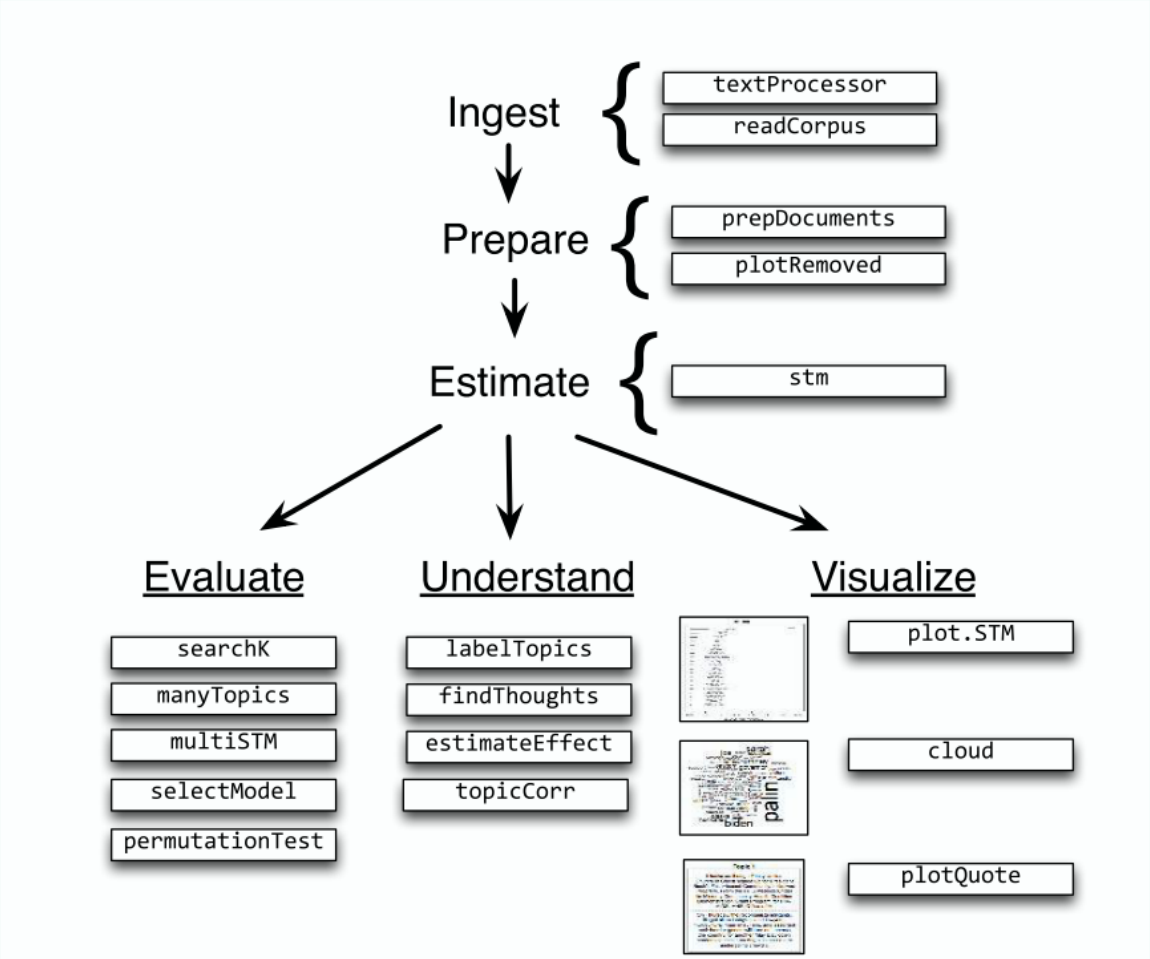

Figure 3: Steps in a topic modelling workflow (M Roberts, Stewart, and Tingley 2018)

Further evaluation may take place with people, where evaluators might be subject matter experts or more general users. The same groups may then form the intended target audience for the product or service that the model is applied to, although often modelling is done only during a research project and not implemented into consumer products or services.

In practice, generation and evaluation usually proceed iteratively, with repeated runs of the modelling process based on an evaluation of the results. The data scientist and domain expert may be part a research team working together on this. With large data sets, the model runs may take some time (as the algorithms are themselves iterating over the documents repeatedly to derive topics) and this stage forms something of a black box, as most algorithms to not have the ability to absorb parameter changes during a run, though with some flavours of algorithm the

analyst can provide a priori constraints as to how words and topics are treated (Bakharia et al. 2016).

A further though more indirect way of combining prior domain knowledge comes though combining modelling outputs with human-generated metadata or document labels. This provides some triangulation between the human and machine classifications and is where interesting insights can be obtained. Such labels may be incorporated into the model a priori as a semi-supervised aspect, or compared a posteriori after modelling. In some ways, as the work reported below suggests, the post-hoc analysis might be more interesting, as incorporating prior knowledge into the model at the start may bias it only toward expected outcomes.

The majority of topic model-based studies in the literature, such as those in Table 1, use automated evaluation measures such as those described in the previous section. Lee et al (2017) however provide a rare attempt to bridge the gap between machine-based topic building and human interpretability and usefulness. They provided a custom interface to non-specialist users to evaluate 20 and 30 topics and document allocations on a corpus of newspaper articles with the option to select refinement operations on the results. They found that the most common desirable operations were to remove words from topics, to remove documents from topics and to change the order of words appearing in topics. Notably, with the first two operations evaluators tended to pick lower probability words and documents, indicating some implicit agreement with the machine-based ranking. In another study of more interactive modelling algorithms, Bakharia et al (2016) noted that the "Topic Creation Rule" was most widely applied by users, allowing them to supply seed words for new topics of interest to their analysis.

1.5 Visualisation for understanding and interaction

While the immediate output of a topic model process are matrices of probability assignments for words and topics, these are not useful or easily interpretable by humans. Work has therefore been required to provide visual summaries of the model outputs to enable evaluation and action by users. Many visualisation interfaces are part of relatively ephemeral research projects, but some are available within open source ecosystems.

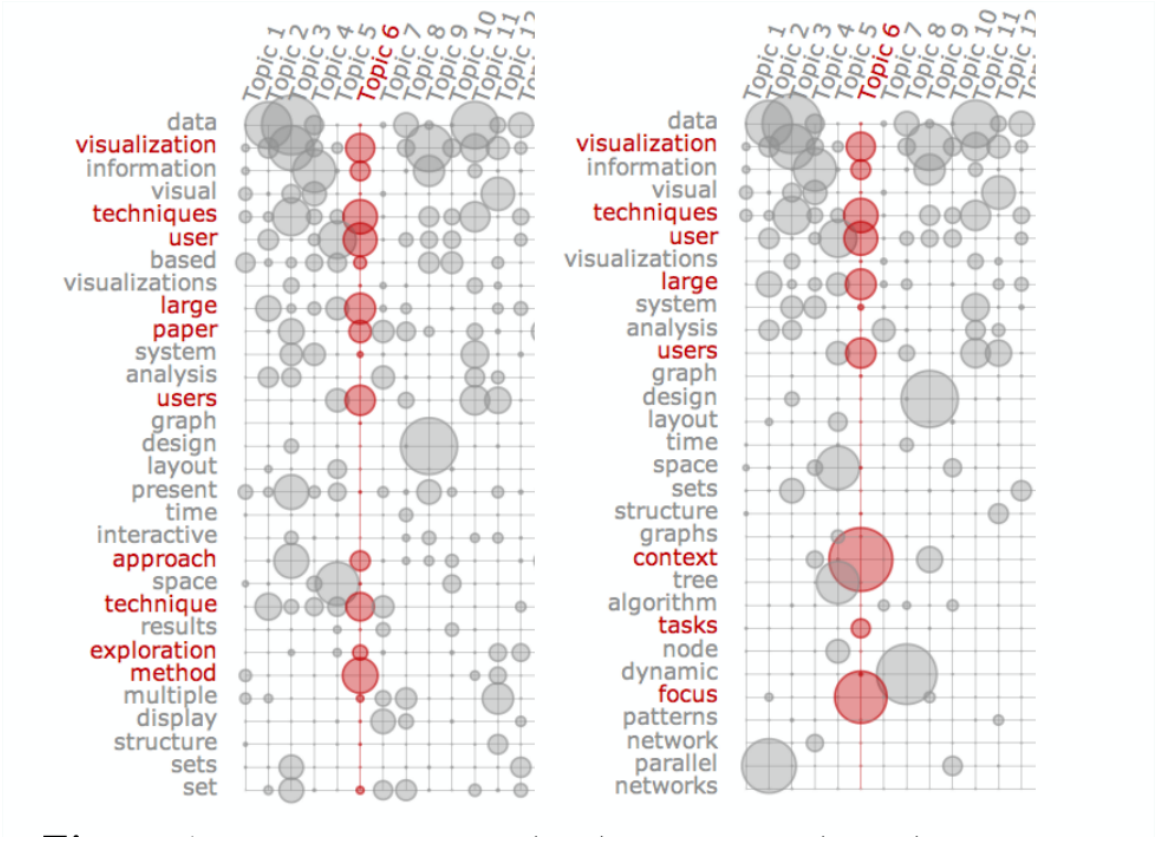

Figure 4: Termite term comparison (Chuang, Manning, and Heer 2012)

Chuang et als' "Termite" tool (2012) enabled visualisation of term salience (filtering out less informative terms within a topic) and seriation (better highlighting of clustered key terms in a topic). An example is shown in Figure 4. LDAExplore, presented by Ganesan et al (2015) enabled the visualisation of word and topic distributions via interactive graphs and treemaps. Users could filter on and examine topics alongside their representative documents and keywords. A small scale evaluation indicated that the visualisations improved understanding of the keyword composition of topics. Another tool, LDAvis (Sievert and Shirley 2015), works with the R statistical programming environment and, like the other tools above, works with the output of an LDA modelling run. It provides ways to view the relative size of topics and inter-topic distance, measured as Jensen-Shannon divergence. It also provides a view of representative terms for a topic and these are ranked both by probability and the use of the "lift" weighting that promotes less common terms characteristic of the topic (Figure 5). LDAvis also allows topics to be clustered to help make sense of a model using larger numbers of topics. A further R tool, stmInsights (Schwemmer 2018), used in the analysis below, takes the output of Structural Topic

Modelling (STM) and provides several sensemaking and visualisation functions. These include topic labelling, topic distributions and visualisation of model diagnostics and effects.

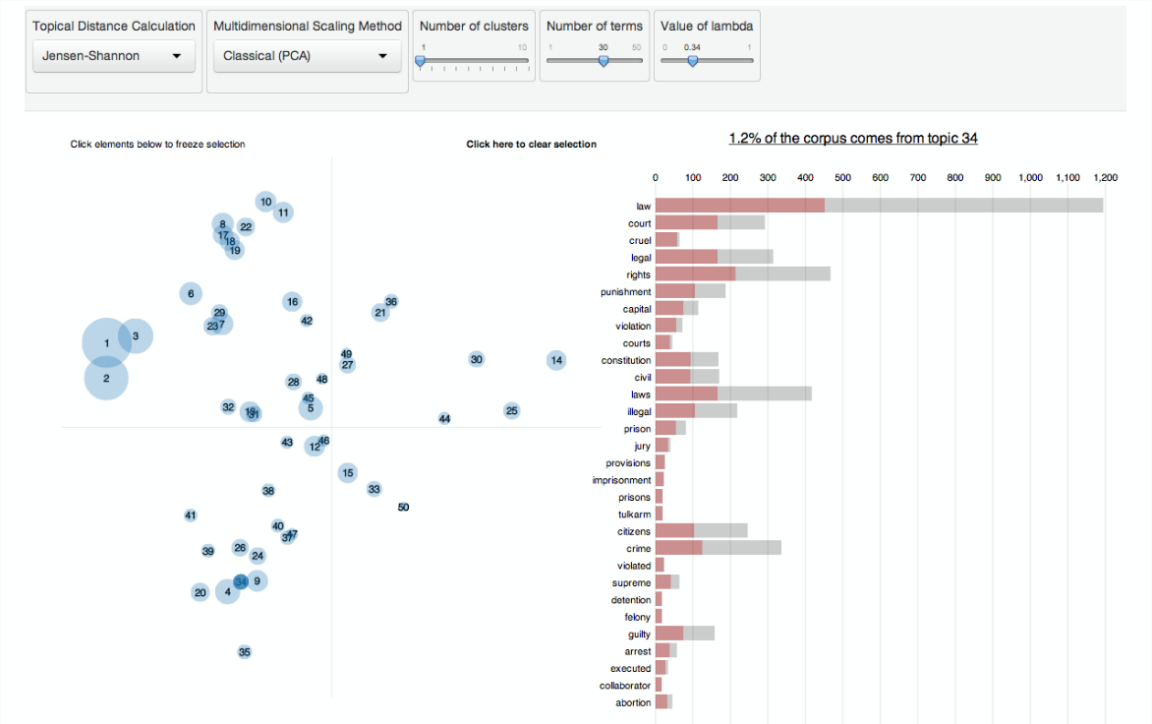

Figure 5: LDAvis topic relationships (left) and term distributions (right) (Sievert and Shirley 2015)

All of the above techniques are post hoc and do not allow the user to affect the actual topic modelling process, only to better understand the results. It is much harder for human interactions to feed back into the algorithm as it is running. That said, recent work by El-assady and colleagues (2019) provides some interesting approaches into how this might be achieved. They use "speculative execution" and model quality feedback from the user to influence and predict the affects of alterations to the topic clusters produced with a hierarchical topic model (Figure 6). Their analytics dashboard allows the user to step through the model run and make alterations, or view the affect of alterations on the model quality.

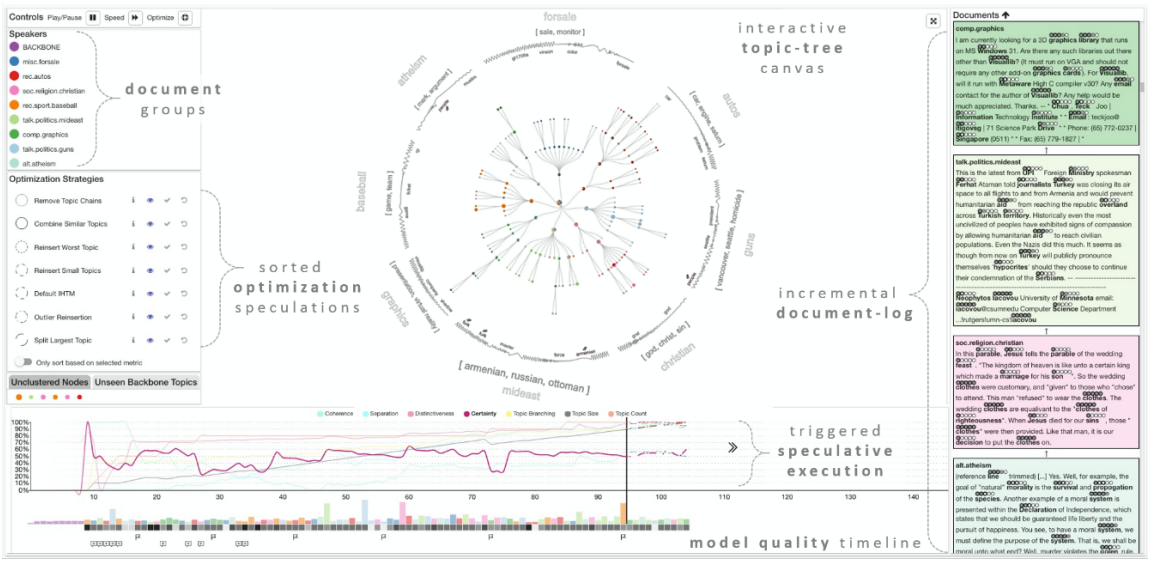

Figure 6: Visual analytics workspace for user-led hierarchical modelling (El-assady et al. 2019)

1.6 Human-grounded decisions and evaluation

1.6.1 Topic numbers

As mentioned above, the choice of topic numbers (commonly denoted as the K value) is often a tradeoff between model accuracy and human interpretablility. As with other unsupervised methods, the received wisdom is that there is no "correct number" as this is an application and context-dependent issue (M. E. Roberts, Stewart, and Tingley 2018). In a sample of recent studies, researchers have tended to select relatively small K of 10-50, independent of corpus size (Table 1)

Reference	Application	No..of.Documents	No.of.Topics	Method	Evaluation
Park et al (2017)	Medical prescriptions	180,000 and 2,000,000	15	Extended LDA (with diagnosis and medication)	Perplexity (5-fold Cross validation)
Wang et al (2017)	Herbal Medicine Cases	3800	Not selected (though ~60 optimal from evaluation)	LDA v Extended LDA (with symptoms and herbs)	Perplexity

Reference	Application	No..of.Documents	No.of.Topics	Method	Evaluation
Giaquinto2018 & Banerjee (2018)	Personal medical journals	9,010,623 journals written by 200,388 authors	50 (15 author personas)	Rapid Dynamic Author-Persona (DAP), compared to other methods	Per-word log likelihood
Ren et al (2016)	Suicide blogs	907	Unclear: top 20 reported (5 emotion intensities)	CET – Topics + emotion and emotion intensities	Perplexity
Lee & Kang (2018)	Journal articles	12000	50	LDA	Subjective
Grajzl & Murrwll (2018)	Bacon's works	282	16	STM	Held-out likelihood, coherence, exclusivity
Faisal & Petoniemi (2018)	Video games wikis	15000	31	Multitask, Non parametric LDA	
Pomeda et al (2018)	Interviews	25	10	LDA	

Table 1: Comparison of recent studies showing topic numbers, model types and validation approaches

Arnold et al (2016) found 100 topics to work best on a word intrusion-based evaluation for specialist and non-specialists on their corpus of medical reports. While claiming that a greater number of topics allows more interesting topics to emerge, Lee and Kang settled on a relatively conservative 50 (Lee and Kang 2018)

1.6.2 Topic interpretability, coherence and labelling

The word intrusion method has been used to good effect to evaluate topic interpretability (Lau, Newman, and Baldwin 2014). Here, an intruder term is added to the top words for a topic and a

human evaluator is asked to identify words that appear out of place. The successful identification of intruders is an (indirect) indication of the coherence of the remaining topic terms.

Topic coherence evaluation on topic models of medical reports by Arnold et al (2016) showed that clinical experts (primary care physicians) performed better than students at identifying mismatched words and topics, indicating for the authors that the models were successfully capturing specialised concepts that the subject experts were better at identifying. The authors note that the system was still far from complete in being able to flexibly absorb new documents with new potential topics, where relearning would be needed.

Whilst topic intrusion was designed as a human oriented evaluation measure, Lau et al (2014) show that this approach can be successfully used as a machine learning task to predict which words are likely to be detected as intruders given a labelled training set and the ability to automate coherence measures using pointwise mutual information (PMI) and conditional probability (CP). For the PMI measure, the topic terms are compared to a reference corpus, but for CP the coherence can be calculated from the documents being used to generate the model. Lau et al also noted that the correlations between human judgements of overall topic coherence and word intrusion are only mild, revealing subtle differences in how the tasks are approached.

Selection of labels is another task that can be fully human controlled, machine-assisted or machine-determined. “Eyeballing” is perhaps most commonly used (Morstatter, Rey, and Ave 2018), whereby labels are selected by the user / analyst based on a visual inspection of the most common words for a topic. This is nevertheless considered a non-trivial task (Lee and Kang 2018) and in their work these authors used expert consensus to derive agreed labels, where 39 of 50 proved easy but the remaining 11 required “in-depth discussion”. In tools such as STM, the most common words can be accompanied by the most exclusive words to give an idea of the distinctiveness of the topic. Machine-oriented approaches include choosing hypernyms and representative words from the topic terms or comparing the terms using reference knowledge bases or ontologies (Boyd-graber, Mimno, and Hu 2017)

2 Analysis using the movie summaries corpus

What do topics mean to a naive audience and how (easily) do they label them? What size of K provides the most useful output for a given corpus? How does unsupervised topic output relate to received and widely used and shared human-generated categories? In order to investigate these questions, the following section outlines the application of topic modelling to a corpus of movie plot summaries.

2.1 Dataset

In order to combine unsupervised modelling with a posteriori analysis a labelled dataset was used. This was the CMU Movie Summaries corpus from Carnegie Mellon University (Bamman, O'Connor, and Smith 2014, 2013). This consists of movie plot summaries (1900-2014) from Wikipedia together with matched movie and character metadata from Freebase (now part of Wikidata). Once imported and cleaned, there were 42,206 records in the dataset, though for some of the analysis this was further subsetted or sampled as mentioned below.

2.2 Methods

The dataset was imported into R and topic models developed using the STM (M Roberts, Stewart, and Tingley 2018) package. The standard workflow recommended by the STM documentation (M Roberts, Stewart, and Tingley 2018) was followed, with text processing followed by document preparation before modelling commenced. At the processing step, a custom name filter was applied to remove as many first names as possible, as these were found to lead to a number of "junk" topics based around names. The first names were retrieved from the dataset of US baby names 1880-2008 (Wickham 2009) and the top 0.05% and above used as the filter list. At the document preparation step, a lower threshold of 10 word occurrences across the corpus was used in order to make the processing more efficient.

To investigate automated topic interpretability measures, two modelling runs were used, firstly with K = 20,30,50,100 and 200 with a random subsample of 1000 movie plots (for K diagnostics, below), then topic models with K = 0, 20,30,50 and 100 were generated for the entire corpus for the user testing and genre analysis (200 proved too computationally expensive). The STM default

(spectral initialisation) was used, with a maximum iteration of 75 per model. Hyperparameters were set as default. The zero K option is particular to the STM package and enables an estimate of optimal topic numbers using dimensionality reduction on the word co-occurrence matrix followed by an algorithm to find the minimum number of points to encompass this reduction (M Roberts, Stewart, and Tingley 2018).

For user evaluation, generated topics at each level of K from 20-100 were presented to users on the Amazon Mechanical Turk platform, with the top 12 most probable words from each topic being presented to 3 distinct evaluators. People were asked to provide a label that best represented the group of words in the topic and then to rate how easy they found the topic to label (from 1-very easy to 6-impossible). They were advised that the label could be one of the topic words if that represented the group well, otherwise it could be their own word or short phrase. They were allowed to use the label "impossible" if they could not see any link between the terms. Evaluators were not made aware that the words came from movie summaries in order not to bias the results toward established movie genre terminology.

For investigation of genres and their relation to output topic distributions, modelling output (the theta distribution corresponding to the proportion of each topic predicted for the document or movie) was joined back to the movie metadata. In most cases the K=50 model run was used for this analysis.

3 Results

3.1 Automated interpretability and number of topics

The STM diagnostics allows the comparison of models for topic semantic coherence and exclusivity (Figure 6). Semantic coherence is related to how often words appear together in the same source document and tends to increase when there are fewer topics with more common words. Exclusivity is calculated from a weighted mean of the constituent words' frequency and exclusivity (ie. tending to appear in only one topic). Figure 7 clearly illustrates the nature of the tradeoff in terms of topic numbers, with a reasonable optimal estimate being between 50 and 100 topics, where there is still good coherence but topics are also sufficiently different to be useful.

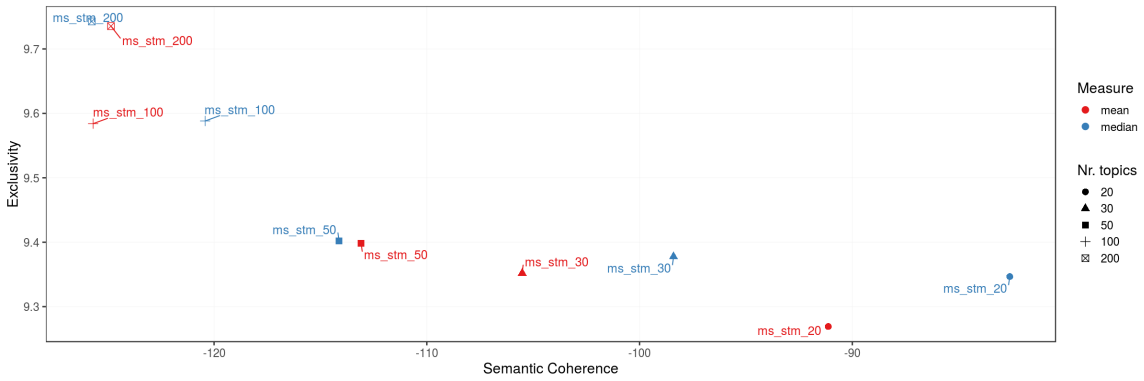

Figure 7: Topic coherence and exclusivity for different K

This was supported by the K=0 modelling run mentioned above, where K is estimated based on the dimensionality of the document-term matrix. This indicated a K of 88, within this range.

3.2 Topic labelling

Evaluators found most of the topics of average difficulty to label. Of those judged very difficult or impossible, more came from the K=100 set, with K=20 containing the fewest (Table 2)

Topics	Labelling - % hard or impossible
20	12.963
30	18.391
50	14.074
100	23.273

Table 2: Topic labelling - percentage of topics judged to be hard or impossible to label

At the other end of the difficulty spectrum, there was less differentiation in the ease of labelling across K (Table 3), though K=30 was marginally reported as the easiest.

Topics	Labelling - % easy or very easy
20	25.926
30	24.138

Topics	Labelling - % easy or very easy
50	28.889
100	27.636

Table 3: Topic labelling - percentage of topics judged to be easy or very easy to label

Table 4 shows some examples of topics that were classed as relatively easy to label, together with the labels applied by different evaluators. While we see quite good agreement in labels, there is clearly variability in the emphasis given to the topic terms by evaluators (particularly in the second topic, labeled as "hobbies", "writing" and "work" respectively).

Topics	No.	Keywords	Difficulty (1=very easy, 6=impossible)	Label
20	1	war, soldier, armi, american, german, men, kill, forc, unit, general, order, british	2	world war II
20	1	war, soldier, armi, american, german, men, kill, forc, unit, general, order, british	2	war
20	1	war, soldier, armi, american, german, men, kill, forc, unit, general, order, british	3	World war
20	2	book, max, letter, write, read, paint, jenni, find, publish, tell, art, work	3	hobbies
20	2	book, max, letter, write, read, paint, jenni, find, publish, tell, art, work	2	writing
20	2	book, max, letter, write, read, paint, jenni, find, publish, tell, art, work	3	work
20	3	school, student, high, teacher, colleg, class, friend, girl, univers, parti, becom, professor	3	college
20	3	school, student, high, teacher, colleg, class, friend, girl, univers, parti, becom, professor	3	school

Topics	No.	Keywords	Difficulty (1=very easy, 6=impossible)	Label
20	3	school, student, high, teacher, colleg, class, friend, girl, univers, parti, becom, professor	3	teachers and education
20	4	ship, island, captain, boat, crew, sea, water, find, gold, rescu, take, fish	3	maritime
20	4	ship, island, captain, boat, crew, sea, water, find, gold, rescu, take, fish	3	being on sea
20	4	ship, island, captain, boat, crew, sea, water, find, gold, rescu, take, fish	3	Ships

Table 4: Examples of topics judged to be easy to label, with labels suggested

In Table 5 are some examples of topics that evaluators found very difficult or impossible to label. Here, topics seemed to me more likely to contain noise from character names used disproportionally highly in the dataset. Interestingly, even topics containing some linked terms e.g. topic 13, containing "magic", "witch" and "spell" was judged by one evaluator as impossible, perhaps due to the distraction of the names and verbs also in the topic keywords.

Topics	No.	Keywords	Difficulty (1=very easy, 6=impossible)	Label
20	6	love, get, marri, come, father, take, son, fall, friend, kill, meet, villag	6	Dramatic relationships
20	10	dave, camp, lake, ted, buddi, find, phil, get, josh, willi, ned, elli	6	Camping
20	12	tell, get, leav, see, back, say, car, ask, find, hous, goe, call	5	commands and places
20	13	magic, castl, simon, stone, find, witch, sir, back, franki, return, spell, take	7	impossible
20	14	dog, get, back, bug, cat, run, tri, see, fall, head, come, tree	6	Pets and outdoors
20	16	king, babi, queen, princ, princess, tell, take, find, love, palac, lord, duke	7	impossible

Topics	No.	Keywords	Difficulty (1=very easy, 6=impossible)	Label
20	17	alien, earth, plane, fli, use, crash, destroy, space, pilot, one, control, bomb	5	strome
30	1	van, miller, mile, junior, nina, willi, pink, davi, anderson, ransom, panther, bishop	5	Character names
30	1	van, miller, mile, junior, nina, willi, pink, davi, anderson, ransom, panther, bishop	7	impossible
30	1	van, miller, mile, junior, nina, willi, pink, davi, anderson, ransom, panther, bishop	6	people
30	2	max, freddi, abbi, puppet, neil, tell, mauric, valentin, lenni, philipp, hugo, alli	7	impossible
30	2	max, freddi, abbi, puppet, neil, tell, mauric, valentin, lenni, philipp, hugo, alli	7	impossible

Table 5: Examples of topics judged to be difficult or impossible to label, with labels suggested (or 'impossible' if no suggestion feasible)

3.3 Correspondence to received genres

As documents are considered to be composed of a combination of topics derived from the unsupervised process, we can see how the results compare to general category groupings, notably the genre to which the film has been assigned in Wikipedia. Most of the films in the dataset were associated with at least one genre, many with several, with overall 363 genres used in the dataset. That said, many were quite niche with only a few films assigned to them.

Taking the learned document-topic distribution matrix (theta) for the K=50 model, we clustered the matrix into 20 centres using K-means. This results in 20 distinct topic combinations for the full movie corpus. These were then compared to the top (32) movie genres and visualised as a

heat map (Figure 8). Here, A high value or deep colour, indicates that a large number of films in that cluster were associated with a particular genre.

Taken together, the results show that the topic signature of the clusters correspond quite well to the genres. For many genres there is a distinct cluster associated. This is especially the case were the topic is likely to have quite distinctive associated vocabulary (e.g. science fiction). The heatmap also identifies genres were there are no particular distinct topic signatures (e.g. world cinema, film adaptation) and that intuitively tallies with the knowledge that these are somewhat "catchall" genres.

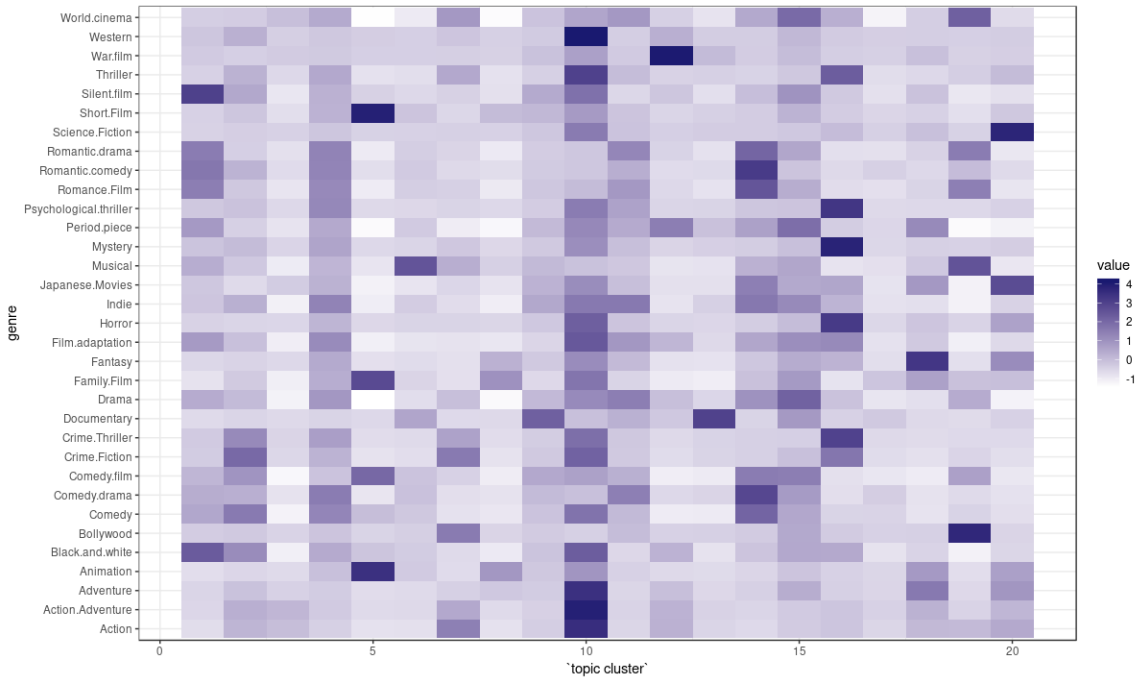

Figure 8: Heat map of topic clusters by genre (topic model K=50, k-means=20)

By visualising the distribution of films over topics for a particular genre, we are able to see "typical" or "prototypic" examples of a genre, together with outliers which might be either misclassified or to some extent "genre busters". Figure 9, for instance, visualised the topic distributions for westerns, with two examples highlighted: "A Fistful of Dollars", perhaps a classic example of a western, and "Brokeback Mountain" a potential genre buster, which was criticised following its release for being stereotypically characterised as a "gay cowboy movie", this characterisation detracting from its importance as a standalone work (Spohrer 2009). We see from

Figure 8 that "A Fistful of Dollars" is itself an outlier, on topic 46 (perhaps due to its Mexican focus) and topic 9 (it does contain an excessive amount of shooting). "Brokeback Mountain" is also an outlier on a number of topics, notably topics 19, 30 and 45 (relationship and romance related). It is in the bottom quartile for the leading topic of westerns overall (topic 42). It is fairly median for violence on topic 9, though notably the violence in the film is directed toward the main characters as homosexuals rather than being instigated by them (Wikipedia 2019).

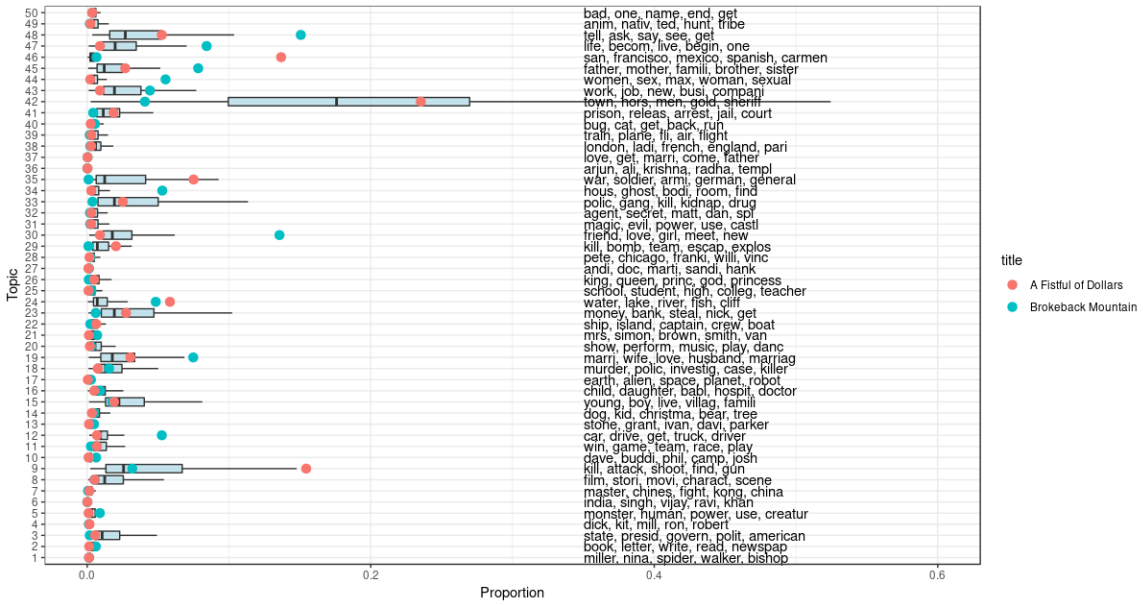

Figure 9: Topic distribution for the 'Western' Genre with two instances compared

Another interesting aspect of the genre is that new genres emerge over time despite earlier works being very much in the mold. Thereafter, movies are made with the specific genre in mind. Road movies are one example of such a genre (Hurault-Paupe 2015). Figure 10 shows the topic distributions for Road movies in the corpus. We see a characteristic emphasis on roads and transport (topic 12) but also on relationships (30) and aspects of self-actualisation (topic 47).

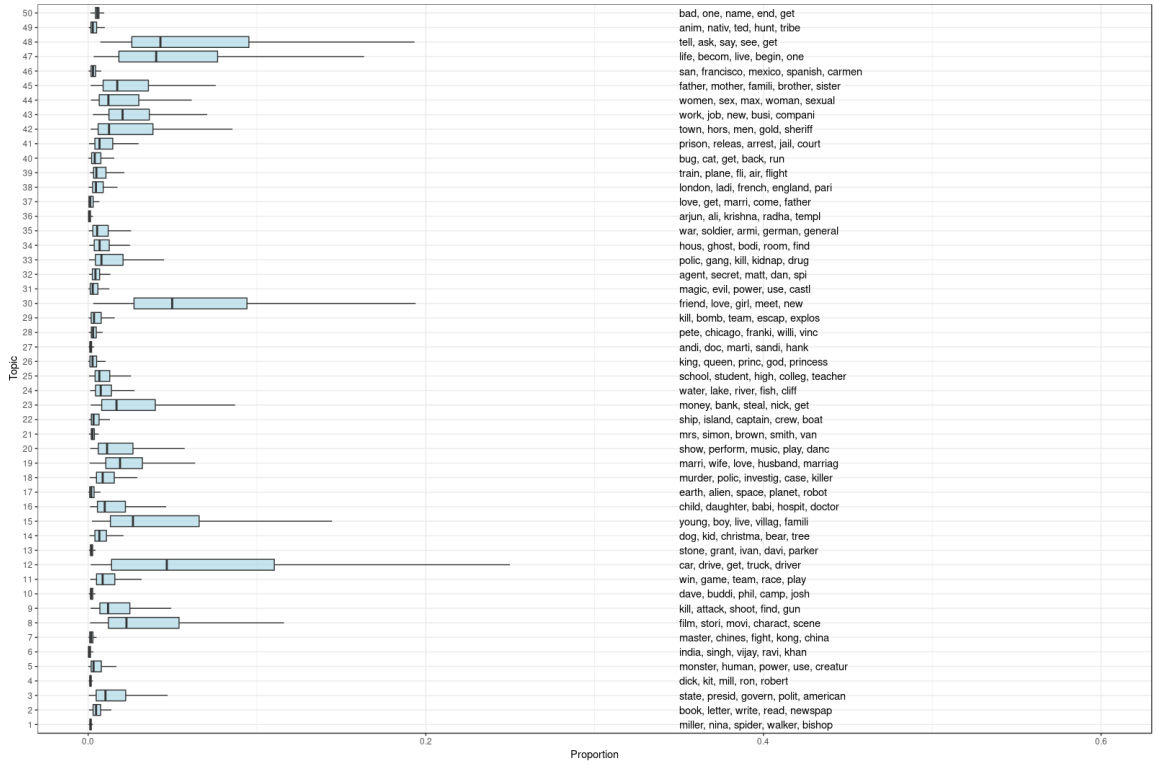

Figure 10: Topic distribution for the 'Road Movie' Genre

The time-bound and perhaps quite flexible assignation of movies to the "Road Movie" genre provides potential to identify works that contain these elements but which have not been explicitly labelled as such. We might define a prototypical movie as the median topic distribution for the genre, with distance from the median calculated as a squared difference weighted by the topic proportion. To remove the effect of short summaries, we only include summaries of 4000 characters or more. Table 6 shows the most prototypical examples of the genre in the left column with other films from the corpus in the right column that have not been explicitly labelled with the genre. Many of these are closer to the median distribution than the road movies themselves. While some time would be needed to ascertain whether these truly qualify as road movies, some have clear elements of travel and transport as well as self-actualisation, though others have been identified by virtue of multiple mentions of motorcycles, cars and diners.

In road movie genre	Distance from median	Not in genre	Distance
My Name is Khan	0.000	Gadar: Ek Prem Katha	0.000
O Brother, Where Art Thou?	0.000	Alien from L.A.	0.000

In road movie genre	Distance from median	Not in genre	Distance
A Canterbury Tale	0.000	The Stepford Children	0.000
Little Miss Sunshine	0.000	Arthur and the Vengeance of Maltazard	0.000
Blues Brothers 2000	0.000	Wake in Fright	0.000
Singh Is Kinng	0.000	Hum Ek Hain	0.000
The Last Detail	0.000	Kim Possible: A Sitch in Time	0.000
Space Truckers	0.000	Cat's Eye	0.000
Sullivan's Travels	0.000	Just Imagine	0.000
Five Dollars a Day	0.001	The Petrified Forest	0.000
The Darjeeling Limited	0.001	Stand by Me	0.000
Cannonball Run II	0.001	The Business	0.000
To Wong Foo, Thanks for Everything! Julie Newmar	0.001	Role Models	0.000
The Brave Little Toaster	0.001	My Beautiful Laundrette	0.000
Adventures of Power	0.001	Passion Play	0.000
The Motorcycle Diaries	0.001	Handle With Care	0.000
Tashan	0.001	Highlander	0.000
Kabul Express	0.001	Epic Movie	0.000
Until the End of the World	0.001	The Happening	0.000
Serving Sara	0.001	Garfield: The Movie	0.000

Table 6: Road movies - labelled and unlabelled candidates

3.4 Further analysis using covariates and visualisation

As was indicated by the human evaluation task, topic labelling is not straightforward or necessarily intuitive. For movie plot summaries it nonetheless tended to group themes relating to activities and settings. This has the interesting side effect of identifying significant action sequences that may be outside the genre classification. Figure 10, below, shows the topic distribution for Bridget Jones's Diary (labels by author, K=20, random sample of 1000 movies). The topics were labelled using the STMInsights package (Schwemmer 2018), which shows examples of document instances with high proportions of that topic. As the "Rocky" topic was mostly inspired by the Rocky series it was so named. Nevertheless, there is a pivotal fight scene

in Bridget Jones between the two male protagonists, played by Colin Firth and Hugh Grant. This example indicates the potential of this approach to provide content-based visualisation and also perhaps could be used for recommendation (e.g. "show me romantic comedies with a bit of fighting!").

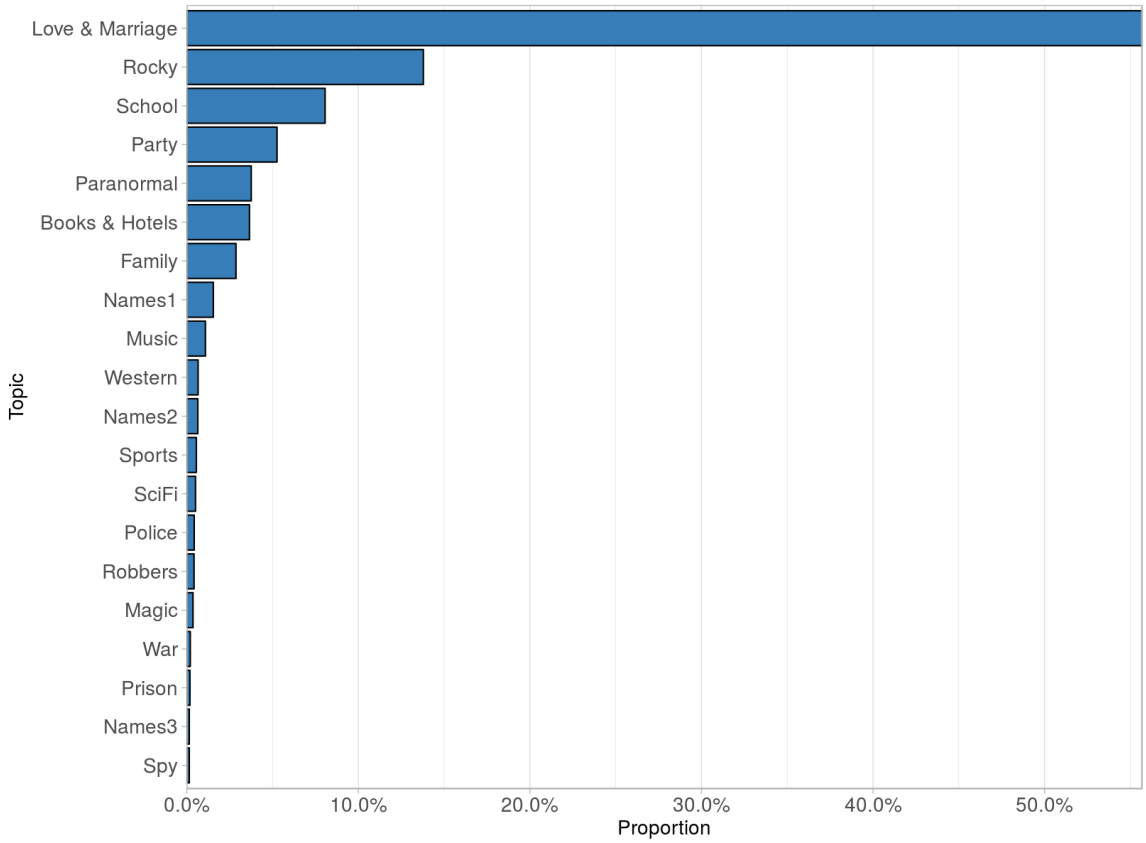

Figure 11: Topic proportions for Bridget Jones' Diary

Covariates can be applied before or after modelling and can indicate patterns and temporal trends in topic distribution of use to the human analyst. In the examples shown in Figures 12 and 13, two author-named topics are visualised with corpus prevalence over time (K=20, random sample of 1000 movies). The graphs show the growth in the party topic over time and the relative shrinkage of the war topic in the sample with time. Certainly for the movie corpus represented on Wikipedia at least, then, there is some indication that hedonistic themes may be overtaking warfare-related themes.

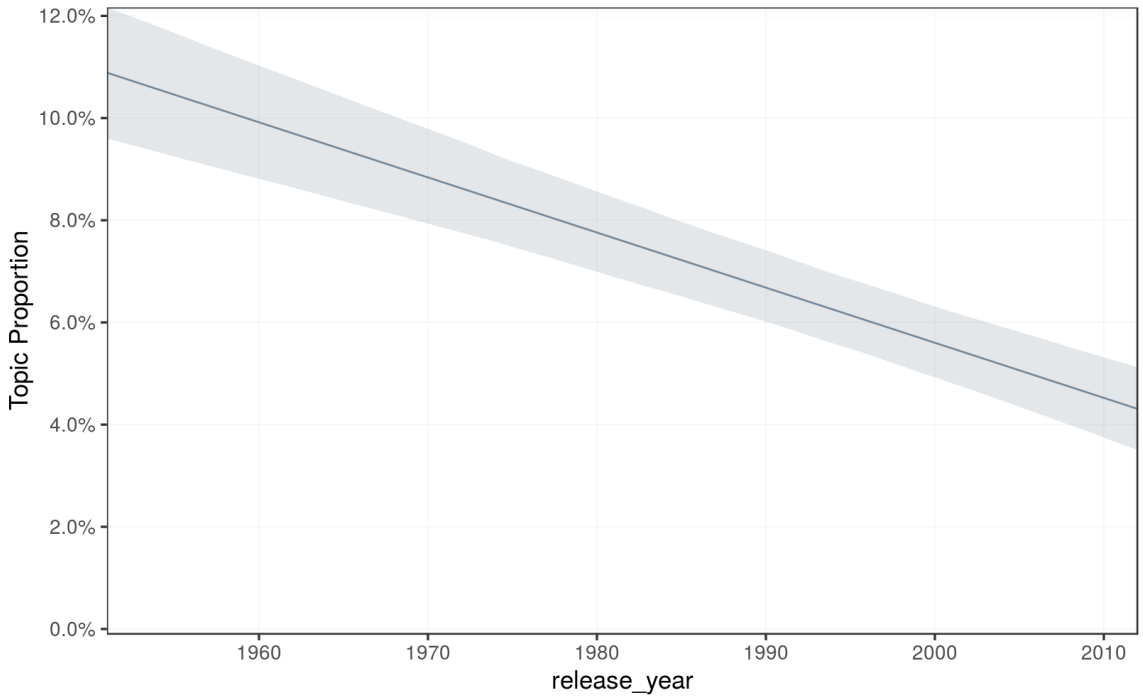

Figure 12: Changes in 'war' topic over time

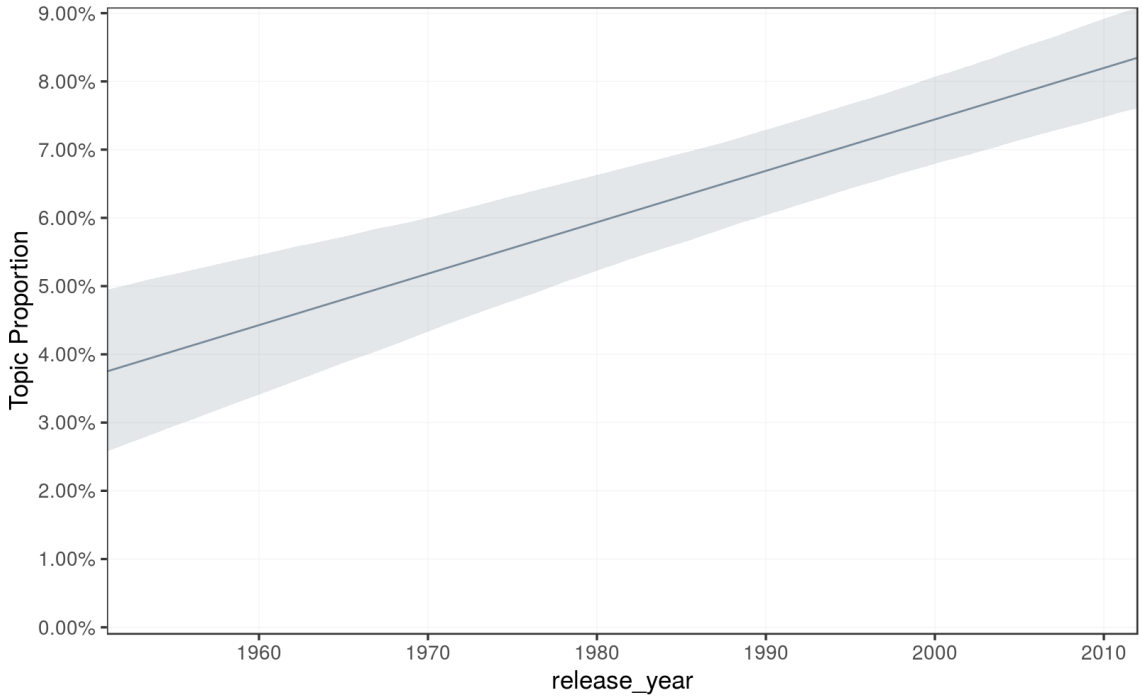

Figure 13: Changes in 'party' topic over time

4 Discussion

In terms of the interpretability levels of Doshi-Velez and Kim [(2017); Figure 2] topic modelling to date is often confined to the first two, certainly in machine learning and computer science literature. The role of human evaluation is often to develop or validate automated evaluation metrics, or is confined to discrete research studies. Information analysts, domain experts and general users tend to be progressively more shielded from the modelling process, though the review of recent work above does suggest some interesting ways forward. In terms of the analyst/researcher this means more control over model parameters and better understanding or the implications of model choice. For users more generally, this may mean the ability to take the model results as a starting point and "clean up" the resulting topics, or provide new constraints for additional iterations. For both groups, better visualisation of model run results enables a greater understanding of how topics have been arrived at and can provide a range of views to allow meaningful labelling. Such tools start to approach a broader view of explainability that takes account of rich prior research and thinking in the philosophy and cognition of causal logic and social psychology (Miller 2017). This work has shown how people can struggle to link together high probability words into coherent themes and this effect worsens with a greater number of topics, even if this greater number leads to a better fit in a purely modelling definition.

Looking further at the sociotechnical design principles of Read et al (Read et al. 2015), there are clear strengths conferred by topic modelling, notably the efficiency and power to apply a structuration process to large text collections. Some iteration is also feasible, though this tends to be limited to repetition of model runs rather than direct, real-time intervention in the modelling process. In terms of validity, there is some doubt over the stability and reproducability of any particular modelling run (Agrawal, Fu, and Menzies 2018). To be useful, a model should be able to produce similar outputs on data from the same source (Greene, O'callaghan, and Cunningham 2014). While this study did not explicitly test this, it seems likely that better estimation processes provide better reproducability (M. E. Roberts, Stewart, and Tingley 2018). Choice is also an important meta-principle of sociotechnical design (Clegg 2000), but existing topic modelling approaches are not particularly good at communicating model options and their potential consequences (Boyd-graber, Mimno, and Hu 2017).

In terms of topic modelling as part of a holistic, integrated process, there are limited tools and systems available that are doing this:

> "..a holistic approach needs to be employed in the development of software that implements interactive Topic Modeling algorithms as qualitative content analysis aids. While the algorithm is the central element and interactivity plays an important role in helping the analyst to answer their research questions, the incorporation of tools to help the analyst interpret the derived topics are equally important and affect trust." (Bakharia et al. 2016)

For analysts with some statistical / coding knowledge, a promising open source toolset is the R Statistical STM package together with the STMInsights web application used in this study, and similar model-visualisation tools are available in Python.

A final design principle from Read's work, and the one judged most important by the experts polled, was creativity. It is here that there seems to be great potential for topic modelling to be part of an exploratively, creative user-led process. What this study indicates is that the unsupervised nature of the approach can lead to interesting empirical findings that perhaps would not emerge so readily from the use of document-level metadata or a single set of keywords. It is in the sub-document patterns and distributions that we can start to look across documents in new and interesting ways.

5 Conclusions and further work

This work brought together some overall design principles for explainable machine learning with specific considerations for topic modelling and used the example of movie summary texts to investigate these. It seems that automated evaluation measures can get us some way toward one of the main modelling decisions, that of topic numbers. That said, the numbers suggested by automated process are at the top end of what seems to be manageable and comprehensible for human users. In our example, while the modelling suggested something around 80 topics, our user testing and further analysis seemed better suited with a maximum of 50. Topic labelling—particularly with lack of context—was confirmed as an often challenging task and there was some

indication that out of place words can easy detract from identifying connections, even where some can be discerned.

Using some exmaples of genre analysis against the topic distributions from the modelling stage, we see that the unsupervised nature of the technique can add potential value to existing, popular social categorisation. While a general topic signature corresponds well with several popular genres, we can also gain insight from examining particular instances and how they vary along different topic axes. A further step might be to develop tools that allow easier exploration of these and the comparison of outliers with prototypical examples.

There is also good potential to take this work further to look at how unsupervised might compare to semisupervised approaches when topic modelling is integrated into discovery and recommendation services. We might hypothesise that the use of labelled text might improve local coherence but limit creative exploration.

6 References

Agrawal, Amritanshu, Wei Fu, and Tim Menzies. 2018. "What is Wrong with Topic Modeling? (and How to Fix it Using Search-based Software Engineering)." *Information and Software Technology*, no. February. https://doi.org/10.1016/j.infsof.2018.02.005.

Arnold, Corey W., Andrea Oh, Shawn Chen, and William Speier. 2016. "Evaluating topic model interpretability from a primary care physician perspective." *Computer Methods and Programs in Biomedicine* 124: 67–75. https://doi.org/10.1016/j.cmpb.2015.10.014.

Bakharia, Aneesha, Peter Bruza, Jim Watters, Bhuva Narayan, and Laurianne Sitbon. 2016. "Interactive Topic Modeling for aiding Qualitative Content Analysis." *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16*, 213–22. https://doi.org/10.1145/2854946.2854960.

Bamman, David, Brendan O'Connor, and Noah Smith. 2014. "CMU Movie Summary Corpus." http://www.cs.cmu.edu/{~}ark/personas/.

Bamman, David, Brendan O'Connor, and Noah A Smith. 2013. "Learning Latent Personas of Film Characters." In *Proceedings Ofthe 51st Annual Meeting Ofthe Association for Computational Linguistics*, 352–61.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77. https://doi.org/10.1145/2133806.2133826.

Blei, David, Andrew Ng, and Micchael Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Boyd-graber, Jordan, David Mimno, and Yuening Hu. 2017. "Applications of Topic Models." *Foundations and Trends in Information Retrieval* 11 (2): 143–296. https://doi.org/10.1561/1500000030.

Brezina, Vaclav. 2018. "Statistical choices in corpus-based discourse analysis." In *Corpus Approaches to Discourse*, 259–80. Milton Park, Abingdon, Oxon; New York: Routledge, 2018.: Routledge. https://doi.org/10.4324/9781315179346-12.

Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. 2012. "Termite : Visualization Techniques for Assessing Textual Topic Models." *International Working Conference on Advanced Visual Interfaces*, 74. https://doi.org/10.1145/2254556.2254572.

Clegg, Chris. 2000. "Sociotechnical principles for system design." *Applied Ergonomics* 31 (5): 463–77. https://doi.org/10.1016/S0003-6870(00)00009-0.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." *Arxiv.org*, no. Ml: 1–13. https://doi.org/10.1016/j.bbrc.2004.04.155.

El-assady, Mennatallah, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. 2019. "Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution." *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 374–84. https://doi.org/10.1109/TVCG.2018.2864769.

Faisal, Ali, and Mirva Peltoniemi. 2018. "Establishing Video Game Genres Using Data-Driven Modeling and Product Databases." *Games and Culture* 13 (1): 20–43. https://doi.org/10.1177/1555412015601541.

Ganesan, Ashwinkumar, Kiante Brantley, Shimei Pan, and Jian Chen. 2015. "LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation." *Arxiv.org*, no. March. https://doi.org/10.1007/BF00268510.

Giaquinto, Robert, and Arindam Banerjee. 2018. "DAPPER: Scaling Dynamic Author Persona Topic Model to Billion Word Corpora." *arXiv Preprint*, 971–76. https://doi.org/10.1109/icdm.2018.00120.

Grajzl, Peter, and Peter Murrell. 2018. "Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas." *Journal of Comparative Economics*, no. October: 1–25. https://doi.org/10.1016/j.jce.2018.10.004.

Greene, Derek, Derek O'callaghan, and Pádraig Cunningham. 2014. "How Many Topics? Stability Analysis for Topic Models." In *ECML Pkdd 2014*. https://arxiv.org/pdf/1404.4606.pdf.

Hurault-Paupe, Anne. 2015. "The paradoxes of cinematic movement: is the road movie a static genre?" *Miranda*, no. 10: 0–15. https://doi.org/10.4000/miranda.6257.

Jacobi, Carina, Wouter Van Atteveldt, and Kasper Welbers. 2016. "Quantitative analysis of large amounts of journalistic texts using topic modelling." *Digital Journalism* 4 (1): 89–106. https://doi.org/10.1080/21670811.2015.1093271.

Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality." In *Proceedings Ofthe 14th Conference Ofthe European Chapter Ofthe Association for Computational Linguistics, Pages 530–539, Gothenburg, Sweden, April 26-30 2014. C?2014 Association for Computational Linguistics*, 38:530–39. 1. https://doi.org/10.11860/j.issn.1673-0291.2014.01.006.

Lee, Hakyeon, and Pilsung Kang. 2018. "Identifying core topics in technology and innovation management studies: a topic model approach." *Journal of Technology Transfer* 43 (5): 1291–1317. https://doi.org/10.1007/s10961-017-9561-4.

Lee, Tak Yeon, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. "The human touch: How non-expert users perceive, interpret, and fix topic models." *International Journal of Human Computer Studies* 105 (July 2016): 28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38. https://doi.org/arXiv:1706.07269v1.

Morstatter, Fred, Marina Del Rey, and S Mill Ave. 2018. "In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics." *Journal of Machine Learning Research* 18 (169): 1–32.

Park, Sungrae, Doosup Choi, Minki Kim, Wonchul Cha, Chuhyun Kim, and Il Chul Moon. 2017. "Identifying prescription patterns with a topic model of diseases and medications." *Journal of Biomedical Informatics* 75: 35–47. https://doi.org/10.1016/j.jbi.2017.09.003.

Pomeda, J. Rodríguez, F. Casani Fernández de Navarrete, L. A. Sandoval Hamón, F. Sánchez Fernández, and Ceci E. Bayas Aldaz. 2018. "a Probabilistic Topic Model on Energy and Transportation Sustainability Perceptions Within Spanish University Students." *European Journal of Sustainable Development* 5 (4): 367–74. https://doi.org/10.14207/ejsd.2016.v5n4p367.

Read, Gemma J. M., Paul M. Salmon, Michael G. Lenné, and Neville A. Stanton. 2015. "Designing sociotechnical systems with cognitive work analysis: putting theory back into practice." *Ergonomics* 58 (5): 822–51. https://doi.org/10.1080/00140139.2014.980335.

Ren, Fuji, Xin Kang, and Changqin Quan. 2016. "Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model." *IEEE Journal of Biomedical and Health Informatics* 20 (5): 1384–96. https://doi.org/10.1109/JBHI.2015.2459683.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2018. "Journal of Statistical Software stm : R Package for Structural Topic Models" VV (Ii): 1–41. https://doi.org/10.18637/jss.v000.i00.

Roberts, Margaret, Brandon M Stewart, Dustin Tingley, and Edoardo Airoldi. 2013. "The Structural Topic Model and Applied Social Science." In *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation.*

Roberts, M, M Stewart, and D Tingley. 2018. "stm: R Package for Structural Topic Models." https://doi.org/10.18637/jss.v000.i00.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures," 399–408. https://doi.org/10.1145/2684822.2685324.

Schwemmer, Carsten. 2018. "Package 'stminsights' - A 'Shiny' Application for Inspecting Structural Topic Models." 4. Vol. 58. https://doi.org/10.1111/ajps.12103.

Sievert, Carson, and Kenneth Shirley. 2015. "LDAvis: A method for visualizing and interpreting topics," 63–70. https://doi.org/10.3115/v1/w14-3110.

Spohrer, Erika. 2009. "Not a gay cowboy movie? Brokeback mountain and the importance of genre." *Journal of Popular Film and Television* 37 (1): 26–33. https://doi.org/10.3200/JPFT.37.1.26-33.

Wang, Lidong, Keyong Hu, and Xiaodong Xu. 2017. "Discovering Symptom-herb Relationship by Exploiting SHT Topic Model." *IPSJ Transactions on Bioinformatics* 10 (0): 16–21. https://doi.org/10.2197/ipsjtbio.10.16.

Wickham, Hadley. 2009. "hadley/data-baby-names: Distribution of US baby names, 1880-2008." https://github.com/hadley/data-baby-names.

Wikipedia. 2019. "Brokeback Mountain." https://en.wikipedia.org/wiki/Brokeback{_}Mountain.