Imperial College London

Imperial College London

Department of Life Sciences, Faculty of Natural Sciences

The Analysis of Topic Bias in Four Academic Funding Agencies using Machine-Learning Approaches

Author:
Jingkai SUN

CID:
01991822

Date:
26 August 2021

Word Count:
5882

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Science at Imperial College London

Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

I can declare that I am the only person who is responsible for data collection, data cleaning, the development of methodology for the topic analysis, text mining and machine learning models used in this project. I herewith certify that all works in my project are done by myself. David Orme provided solutions to some statistical problems, James Rosindell shared the idea of using topic modelling techniques with me and provided some scientific papers related to my project. He also gave feedback on the draft of my project. Samraat Pawar provided supports for the technical questions in terms of Python, R and Latex and also offered useful suggestions for my draft of the project.

Acknowledgement

Firstly, I would like to thank my supervisors for solving all my questions during the weekly meeting. David Orme offered me an opportunity to study this interesting topic and many supports on statistical analysis in the project. Samraat Pawar provided lots of computing modules and solutions to my technical questions so that I have the ability to process the data and create machine-learning models using Python, Shell and R. James Rosindell provided me numerous ideas from the start of the project and suggestions for assigning topic labels in my topic model. Additionally, he arranged multiple meetings for providing me many useful feedbacks based on the draft of the project. In addition, I also would like to thank Yewshen Lin, who is one of my colleagues, for sharing his suggestions and experiences of using text-mining techniques. Moreover, a big thank you to Marina Papadopoulou, a student who graduated with an MSc in Computational Methods in Ecology and Evolution at Imperial College London, for the methods of the selection and assessment of topic models she introduced in the dissertation. Last but not least, many thanks to everyone in James' lab meeting for listening to my presentation and giving suggestions to me.

Contents

List of Figures **4**

1 Abstract **5**

2 Introduction **6**

3 Methods **8**

3.1 Experimental Data 8

3.2 Latent Dirichlet Allocation 9

3.3 Coherence Measurement 10

3.4 Model Selection 11

3.5 Prediction of Funding Amount 12

3.6 Computing Tools 13

3.7 Data Availability 14

4 Results **14**

4.1 Topic Selection 14

4.2 LDA Model Fitting and Topic Visualisation 16

4.3 Topic Analysis 18

4.4 Prediction of Funding Amount 22

5 Discussion **22**

6 Conclusion **24**

Support Information **26**

References **37**

List of Figures

1 The result of the coherence measurement 14

2 Mean similarity and stability between models 15

3 Interactive visualisation for topic distributions 16

4 Probability distributions of all topics from all documents 17

5 Eight wordclouds of top 25 stemmed words 18

6 Heatmap between topic and agencies 19

7 Heatmaps between topics and 50 intervals of discrete funding amount 20

8 The accuracy of prediction of funding amount using LightGBM 22

9 Visualisation of topic 1 by pyLDAvis 30

10 Visualisation of topic 2 by pyLDAvis 30

11 Visualisation of topic 3 by pyLDAvis 31

12 Visualisation of topic 4 by pyLDAvis 31

13 Visualisation of topic 5 by pyLDAvis 32

14 Visualisation of topic 6 by pyLDAvis 32

15 Visualisation of topic 7 by pyLDAvis 33

16 Visualisation of topic 8 by pyLDAvis 33

17 The probability distribution of topic probabilities after cut small probabilities . . 34

18 t-SNE clustering for visualisation of eight topics 34

19 Word count and importance of topic keywords 35

20 Sentences coloring of dominant topic in part of documents 36

1 Abstract

Academic funding plays an essential role in research projects. Due to the finite amount of academic funding and an increasing number of emerging research fields, funding agencies could provide more funding to some specific topics, resulting in a topic bias in funding agencies. Combining topic modelling with a supervised machine-learning algorithm, I examined the topic bias from the information of funded projects in the National Institutes of Health (NIH), European Research Council (ERC), National Science Foundation (NSF) and UK Research and Innovation (UKRI). Then, I used a supervised classifier to predict the amount intervals of funding, with topic distributions generated from topic models and project duration as predictors. I collected 145,787 funded project data between 2015 and 2021 from 4 academic funding agencies and extracted topic distributions using latent Dirichlet allocation. Then, I evaluated the topic distributions in each agency and each amount interval of funding per day in project duration. Afterwards, I created Light Gradient Boosting Machine to classify three intervals of funding amounts (i.e low-, medium- and high-amount classes) and made comparisons to models with different predictors in different agencies. The results show that (1) topic modelling is helpful for topic assignment of project abstracts; (2) The topic "Computer Systems and Commercial Application" dominated NSF, ERC and UKRI agencies, whereas the topic "Cancer Treatment and Immunology" dominated NIH agency; (3) With combining the topics "Computer Systems and Commercial Application" with "Cancer Treatment and Immunology" (e.g. computational biology or medical engineering), a project could have a higher probability to obtain more funding in all agencies; (4) Topic distributions are found to improve the accuracy of prediction of the funding amount by 5% – 20% once I combined them with project duration as predictors together, even though the model fails to have sufficient prediction power when the model solely considered topic distributions as predictors.

Keywords: Text Mining, Bibliometrics, latent Dirichlet allocation, Multiclassification, Funding agencies, Academic funding, National Institutes of Health, European Research Council, National Science Foundation, UK Research and Innovation, LightGBM, Machine Learning

The Analysis of Topic Bias in Four Academic Funding Agencies using Machine-Learning Approaches

Jingkai Sun (CID: 01991822)

August 26, 2021

2 Introduction

The advancement of academic research is not only dependent on the advancement of technology, but also on funding support. Academic funding can provide essential support for scientists, especially for medicine and natural science studies [1]. However, with an increase in the number of researchers, scientists and emerging research fields all over the world [2], the competition of the academic market is increasingly fierce so that funding fail to cover all excellent research projects. This indicates a decrease in opportunities for scientists to obtain enough funding to complete their research projects, many scientists worry about their research careers [3]. Therefore, it is important to understand determinants of funding success for research projects and how they will be biased due to different reasons.

In 2011, Allesina *et al.* studied whether nepotism can affect scientific success in Italian Academia using conventional statistical analysis [4]. They leveraged the Monte Carlo approach to show there are fewer different last names of principal investigators (PIs) compared with the expected situation at random for each discipline in Italian research institutions. This revealed the existence of nepotism in most disciplines, even though the results are underestimated due to the exclusion of cases of nepotism of the mother-child type. Furthermore, they also used logistic regression models to examine how the probability of sharing the last name changes with factors such as geography, institutions, sub-disciplines and latitude, concluding a significant latitudinal influence on nepotism.

However, at that time, in most of the research that studies the factors of scientific success such as what Allesina *et al.* did, they used traditional statistical analysis to uncover the factors that determine the funding success. In recent years, with the development of machine-learning techniques and the computational power of computers, big data are more influential for us in

recent decades than in the past [5]. It is possible to find the relationship between things and hence predict what will happen in the future through big data in many research fields. For example, Mahajan *et al.* utilised the Naïve Bayes model with a large amount of data in the stock market to predict the movement of the stock market and conclude that the model is useful for predicting the future prices of shares [6].

Therefore, more researchers utilised machine-learning and text mining techniques to extract information and patterns from numerous text data such as scientific papers. In 2014, Emre *et al.* analysed 100,000 publications from Computer Science to examine the centrality in the coauthorship network and use it with a random forest algorithm to predict whether a paper can be highly cited five years after publication [7]. The final outcome showed a high accuracy of prediction. Additionally, Michael *et al.* applied several machine-learning classifiers to predict the funding success of projects from a crowdfunding website – Kickstarter. They used the information of projects such as project duration, the number of Twitter followers and the number of Facebook connections *etc.* as features, obtaining 68% of the accuracy of prediction [8]. Instead of using the basic information of projects exclusively, Tanushree and Eric (2014) extracted phrases in the project description by text-mining techniques from Kickstarter and explored whether they are the factors determining the funding success [9].

To explore the impact factors of academic funding success using text-mining approaches, Magua *et al.* (2017) studied the problem of gender bias in the National Institutes of Health (NIH), an academic funding agency. They evaluated the impact on the funding process caused by gender differences in NIH by combining Latent Dirichlet Allocation (LDA) and qualitative analysis [1]. They first used the regression models to examine the sex differences by assigned scores for male (PIs) and female PIs. The outcome suggested, compared with female PIs, that male PIs' applications have higher priority, approach and significance scores with having the same experience level and research quality. With latent Dirichlet allocation (LDA), they also found that funded male PIs are usually described as "pioneers" in specific fields with "highly innovative" and "significant research", whereas funded female PIs are described as "expertise" or "excellent". The outcome revealed a distinct criterion that may lead to different funding success rates of funding applications between male and female PIs.

The difference of research topics of the projects is also another one of the most important impact factors of funding success [10]. By finding the 'hot' topics that can be more funded, researchers can study their own fields by combining these 'hot' topics to increase the probability of obtaining more funding. However, using big data approaches to study the relationship of topics and academic funding are not found yet.

Thus, the objective of this project is to explore the relationship between topics and academic

funding in the National Institutes of Health (NIH), National Science Foundation (NSF), European Research Council (ERC) and UK Research and Innovation (UKRI), which are four well-known academic funding agencies in the world. At the end of the project, I try to answer the following questions:

- Is text-mining techniques useful for finding 'hot' topics in different agencies?
- Whether different topics could affect the funding amount in four different academic funding agencies?
- What are some topics that could improve the probability of obtaining more funding in four agencies?
- Are academic funding can be predicted using Machine Learning Approach?
- Can topics distributions improve the accuracy of prediction of academic funding amount?

3 Methods

3.1 Experimental Data

Data was collected from the websites of UKRI, NIH, NSF and ERC. After removing projects of which abstracts were unavailable, data from 145,787 funded projects awarded between 2015 and 2021 were collected, including 27,035 project abstracts from the UKRI funding agency collected by the GtR API program [11], 41,796 abstracts from the NSF database of funded projects [12], 71,008 abstracts from NIH collected using RePORTER database and finally 5,948 abstracts from ERC collected directly using its website database [13, 14].

To obtain more accurate results, project abstracts need to be preprocessed in ways that are required by any text mining technique. The first step was to remove all HTML character references (e.g. & or <), which is a way to display special characters on a website page since some of the abstracts in my dataset are scraped from websites. Then, I removed all punctuations, non-English words and English stop words such as "You", "I", "necessary" and "on" *etc.* Some of the stop words are from the NLTK package in Python and some are identified manually (see Support Information). In addition, to improve the accuracy of the LDA model, bigram and trigram models were utilised to identify two words or three words that have a high probability to occur together [15]. For example, if the word set {chimeric, antigen, receptor} could have a higher probability to show up together in the genetic- or protein-related topics than physics-related topics, then the trigram will treat them as a single word with an underscore (e.g. chimeric_antigen_receptor).

Also, these co-occurring words are helpful for describing topics produced by the LDA model, even though a few co-occurring words are not ideal such as 'erc_start' [16]. Afterwards, all remaining words were stemmed, meaning that all inflected words were cut back to a common root [17]. For instance, all words like "damage", "damaged" and "damaging" were transformed into "damag". Next, all high-frequency (occur in more than 80% of abstracts) and low-frequency (occur in less than 0.1% of abstracts) words were removed. Finally, 7060 unique words are left to fit the topic models.

The final dataset consists of project title, project abstracts, funding amount of projects, the project duration of projects. Because four funding agencies are not founded in the same country, meaning the currency are distinct from each other. For consistency, the currency of funding amount in four agencies is converted to U.S. Dollars ($) with real-time exchange rates using the Fixer API program [18].

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is one of the most popular models in topic analysis, which is an unsupervised machine learning technique. First, I introduce D as the number of documents, V as the number of words in corpus and K as the number of topics. LDA treats D documents in the dataset as a probability distribution of K topics. LDA is used to infer K latent topics that can summarise V words in text documents [16]. From this perspective, the LDA model can be also regarded as a dimensionality reduction tool, transforming V words into K clusters (i.e. topics). Two matrices are inferred in the LDA model for obtaining topic distributions for D documents [19]. The first matrix is a document-topic matrix with D rows of documents and K columns of topics, the values in the matrix are probabilities of each document belonging to each topic. Hence, the sum of probabilities of K topics for each document should be one. Similarly, the topic-word matrix enables K topics to have words with different weights representing how important each word is. For instance, if the high weights are generated by the LDA model for some words such as 'climate', 'emission', 'change', 'environment' and 'temperature', then this might be a topic related to "Climate Change" or "Environment Protection".

In this project, the LDA model is implemented by collapsed Gibbs sampling (see Support Information section). The collapsed Gibbs sampling method has the following basic process: (1) Select a document and randomly assign topics for each word in the document. (2) Repeat the same process for every document in the corpus and then summarise local and global (i.e. for all documents) statistics for each topic. For example, if a document has words such as 'epilepsy', 'dynamic', 'bayesian', 'EEG' and 'model', one first assign topics for each word in the document and summarises the number of words each topic included in the document (local statistics) and

the number of times for each word grouped into each topic in all documents (global statistics). (3) Topics are reassigned randomly again and the probability of a new assignment is estimated. (4) The same process is iterated throughout the words and documents. The probability of a new assignment can be estimated by [20]:

$$P = \frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \frac{m_{t,k} + \beta}{\sum_{w \in V} m_{t,k} + V\beta} \tag{1}$$

where n_{ik} is the number of current assignments to topic k in doc i, N_i is the number of words in document i, α is a hyperparameter that controls the prior distribution over topic weights in each document, K represents the number of topics, $m_{t,k}$ means the number of assignments of given word t to topic k, V is the size of vocabulary and β is a hyperparameter for the prior distribution over word weights in each topic. Tuning two hyperparameters also affect inferences of the LDA model. Higher α can make document preference "smoother" over topics, enabling the model to generate more topics with higher probability; higher β also enables more terms to be important for topic description and vice versa, i.e. topics preferences are "smoother" over their keywords. Therefore, for simplicity, my research used default values of α and β in Gensim [21], which both are one over the number of topics. Additionally, because collapsed Gibbs sampling is based on Markov Chain Monte Carlo (MCMC) algorithm, leading to the stochastic output when the same LDA model with fixed parameters is run for multiple times, explaining why I need to conduct a *stability* test for multiple topic models in the following section [22, 23].

3.3 Coherence Measurement

The coherence measurement is an approach to examine how the words in the same topic support each other [24]. Many metrics have been developed to measure how coherent a set of words in each topic such as C_{uci}[25], U_{mass} [26], C_{npmi}[27], C_v [28] and C_p [29]. In this project, the C_v measure is applied since it has been tested that it outperformed all other metrics in coherence measurement [28]. C_v is based on a sliding window [30], which uses one-set segmentation (i.e. any single word within a set is paired with every other single word for comparison) between topic words as well as an indirect confirmation measure [24] to calculate normalized pointwise mutual information (NPMI) for each word pair [28]. Afterwards, the cosine similarity for each NPMI of word pairs is calculated [30]. The equation of NPMI can be expressed as:

$$NPMI(w_i, w_j)^\gamma = \left(\frac{PMI(w_i, w_j)}{-log(P(w_i, w_j) + \epsilon)} \right)^\gamma \tag{2}$$

where

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \tag{3}$$

where $P(w_i, w_j)$ is the probability of words w_i and w_j co-occur, ϵ is a parameter that can avoid the logarithm of zero, γ is a parameter that can give more weights on higher NPMI values as γ increases.

3.4 Model Selection

To obtain an efficient topic model, I first run topic models with a given number of topics k with 10,000 iterations [31], where each model M^k are repeated by 10 times and each repetition has a single C_v value. Therefore, I obtained 10 different coherence values for each model M^k. Mathematically speaking, each model M^k has a vector of coherence values, expressed as:

$$\begin{bmatrix} Coh_1^k & Coh_2^k & \dots & Coh_n^k \end{bmatrix} \tag{4}$$

where $n = 10$ and Coh_i^k means the i-th coherence measurement (i.e. C_v value) for a k-topic model (i.e. the topic model with k topics).

In this project, I tested 32 topic models whose the k starts from 2 to 64 with increment 2 (i.e. $k \in \{2, 4, 6, ..., 64\}$). Therefore, I gained 32 vectors of coherence values, expressed as a $(p \times n)$ coherence matrix:

$$\begin{bmatrix} Coh_{11}^{k=2} & Coh_{21}^{k=2} & \dots & Coh_{n1}^{k=2} \\ Coh_{12}^{k=4} & Coh_{22}^{k=4} & \dots & Coh_{n2}^{k=4} \\ \vdots & \vdots & \ddots & \vdots \\ Coh_{1p}^{k=64} & Coh_{2p}^{k=64} & \dots & Coh_{np}^{k=64} \end{bmatrix} \tag{5}$$

where n is the number of coherence values in each k-topic model, p is the number of k-topic models.

Afterwards, I computed the median C_v (Cv_{med}^k) values for each row in the coherence matrix,

transforming the coherence matrix into a $(p \times 1)$ vector, shown as follows:

$$\begin{bmatrix} Cv_{med}^{k=2} \\ Cv_{med}^{k=4} \\ \vdots \\ Cv_{med}^{k=64} \end{bmatrix} \tag{6}$$

where each element in the vector represents the median C_v value for a k-topic model M^k. Then I choose a subset of M^ks with relatively high Cv_{med}^k and make a comparison with each other using *stability* and mean *similarity* approach (see Support Information) to find the final topic model.

3.5 Prediction of Funding Amount

To find out whether the topic distributions I obtained using the LDA model have prediction power or improved prediction compared to the project duration, I used a supervised machine-learning tool to forecast the discrete funding amount of projects. First, I divided the four agencies into two groups: the NIH group (the health-specialist agency) and all the other agencies including NSF, ERC and UKRI. Two targets were predicted, which are the total funding amount for each project and the funding amount per day, called daily funding amount in this project, which is calculated by:

$$A_{daily} = \frac{A_{total}}{Duration} \tag{7}$$

where A_{daily} is the daily funding amount, A_{total} is the total funding amount, $Duration$ is the project duration in the unit of days.

Then, I treat the prediction as a multiclassification problem. Thus, all funding amounts are discretized to three different classes (i.e. low-amount zone, medium-amount zone and high-amount zone). For the multiclassification problem, I utilised a popular gradient boosting framework known as Light Gradient Boosting Machine (LightGBM) [32]. LightGBM is based on the gradient boosting decision tree (GBDT), which is a prevalent machine-learning algorithm [33] that can be used to solve multiclassification problems [34] and learning to rank [35]. However, different from GBDT, LightGBM leverages approaches of Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS method retains only data instances with larger gradients. It can provide more information gain, obtaining more accurate outcomes. Meanwhile, the EFB approach can optimize the efficiency of feature selection, enabling the LightGBM algo-

rithm to handle data with less computational complexity [32]. The LightGBM has been proved that it can outperform XGBoost and SGB algorithms concerning the efficiency of computation of memory consumption [32]. Therefore, the prediction was done by solely LightGBM.

I created separate LightGBM models according to datasets of two groups (NIH and all other agencies). For each group, I built three models with distinct predicting variables. The first one is the LightGBM model with all topic distributions as predicting variables (defined as M_{topic}), the second model is created with only project duration as a feature, defined as $M_{duration}$. The third model has all topic distributions and project duration as predicting variables, defined as M_{all}. All predicting variables used for the three models are normalised by using the "StandardScaler" function in the "sklearn" package in Python [36]. Finally, these two groups comprised 70,842 and 74,141 observations respectively, with a balanced number of classes and six models in total (i.e. three models for each group). 70% of observations are used as the training set, the remaining observations are used as the test set for prediction in both two groups.

There are three steps to make a prediction: (1) I fitted all models described above by using normalised predictors. (2) I chose parameters for each model, where the learning rate is 0.02, boosting type is GBDT (i.e Gradient Boosting Decision Tree), the maximum depth of the tree is 80, the number of leaves is 200 and the number of iterations is 150. The parameters of all models are the same. (3) The probabilities of three classes for each data point were gained. I then chose a class with the highest probability as a label for each observation. The accuracy of each model was calculated by the expression:

$$Accuracy = \frac{tp}{tp + fp} \tag{8}$$

where tp is the number of true positives and fp is the number of false positives, indicating the ability of the model not to assign positives to the data that is negative.

3.6 Computing Tools

In this project, all scraped data are stored in a MySQL database. ALL analytical techniques are conducted using Python 3, where the "pymysql" package is used for extracting data from MySQL database, packages "numpy", "pandas", "NLTK", "pyinflect" and "sklearn" are used for data preprocessing, the LDA model used in the project is implemented by the "Gensim" package, the "pyLDAvis" package is used for model visualisation.

3.7 Data Availability

All scripts of data preprocessing, text-mining and machine-learning modelling are available at `https://github.com/Kyle-J-Sun/CMEE_Final_Project`

4 Results

4.1 Topic Selection

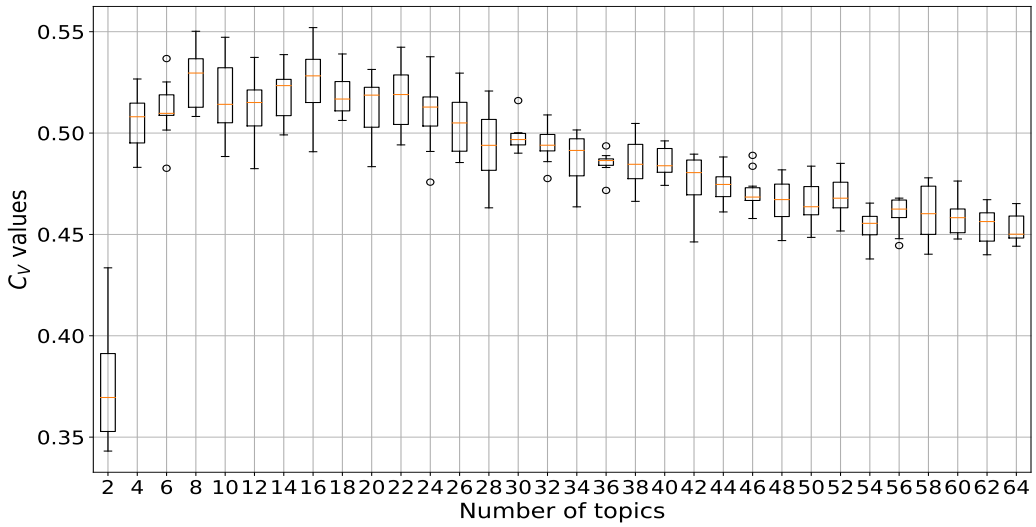

Figure 1: **Coherence measurement:** The figure shows the result of the coherence measure (C_v) for topic models with 2 to 64 topics with increment by 2, every observation is the boxplot of coherence vector of each k-topic model. The orange line of every boxplot represents the median C_v value of each 10-run LDA model with given topics. Higher median C_v values are preferred.

Fig 1 shows an increase in C_v values from 2 and 8 topics, while a decreasing trend occurred from 22 to 64 topics. Therefore, the topic models with 8 to 22 topics (8 topic models in total) were selected for the *stability* test since their median C_v values are significantly higher than others, where the 8-topic models have maximum median C_v value (Median $C_v = 0.53$) and 10-topic models have minimum median C_v value (Median $C_v = 0.51$).

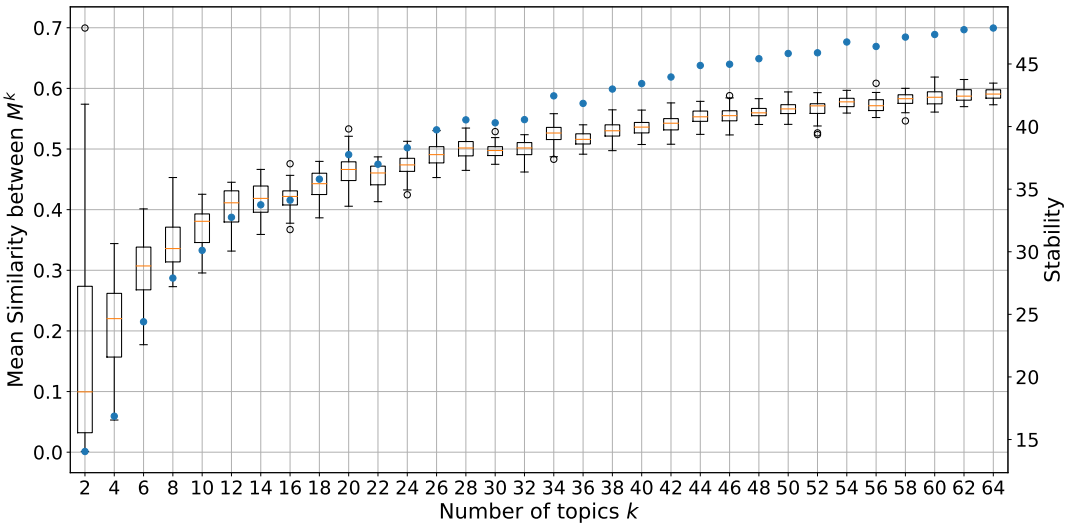

Figure 2: **Mean similarity and stability between models**: The main y-axis is the averaged *similarity*. Each boxplot is a vector of averaged *similarity* values. The secondary y-axis represents *stability* values, which are blue points in the figure. The best model should have the lowest mean *similarity* and *stability* values (see Support Information section) in the subset of k-topic models selected by C_v metrics.

Furthermore, the averaged *similarity* and model *stability* both show an increasing trend from topic 2 to topic 64 in general. The 8-topic model has the lowest averaged *similarity* ($Sim(M^k) = 0.34$) and *stability* ($Stab(M^k) = 27.9$) between eight models selected by coherence measurement. In conclusion, by figures above, the 8-topic LDA model has the highest median C_v value while its median mean *similarity* and *stability* values are also lowest compared to other selected topic models. Therefore, the 8-topic LDA model is considered as the final model throughout the analysis.

4.2 LDA Model Fitting and Topic Visualisation

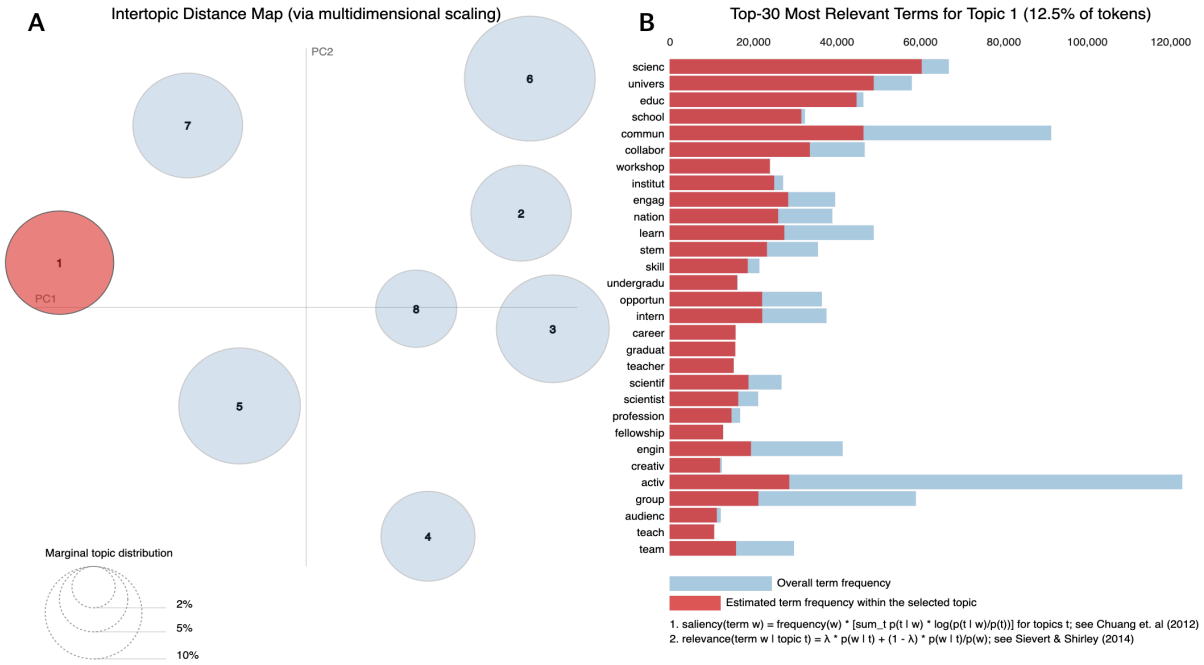

Figure 3: **Topic Visualisation:** An example of interactive visualisation for topic 1 produced from the 8-topic LDA model. Figure A projects eight topics onto a two-dimensional plane by Principal Component Analysis (PCA), where centres of topic circles represent the distance between topics. The areas of circles represent the frequency of topics. Figure B provides a bar chart, where red bars are the frequency of the most relevant stemmed words in topic 1, light blue bars show the corpus-wide word frequency of the individual topic words.

For the interpretability of topics, λ is adjusted to 0.61 (see Support Information section). From Fig.3 A, there is no overlap between the areas of 8 topics, indicating that 8 topics are well-distributed on the two-dimensional plane, supporting that the topic model with 8 topics could be the best-fit model in my dataset. For Fig.3 B, the most five relevant words are 'science', 'university', 'education', 'school' and 'collaboration' respectively. Thus, the first topic might be relevant with "Higher Education". I used the same logic to assign labels for all eight topics. The labels of eight topics are: "Higher Education", "Brain Science", "Protein Chemistry and Microbiology", "Material Science and Physics", "Computer System and Commercial Application", "Cancer Treatment and Immunology", "Social Policy and Public Healthcare" and "Climate Change and Environment".

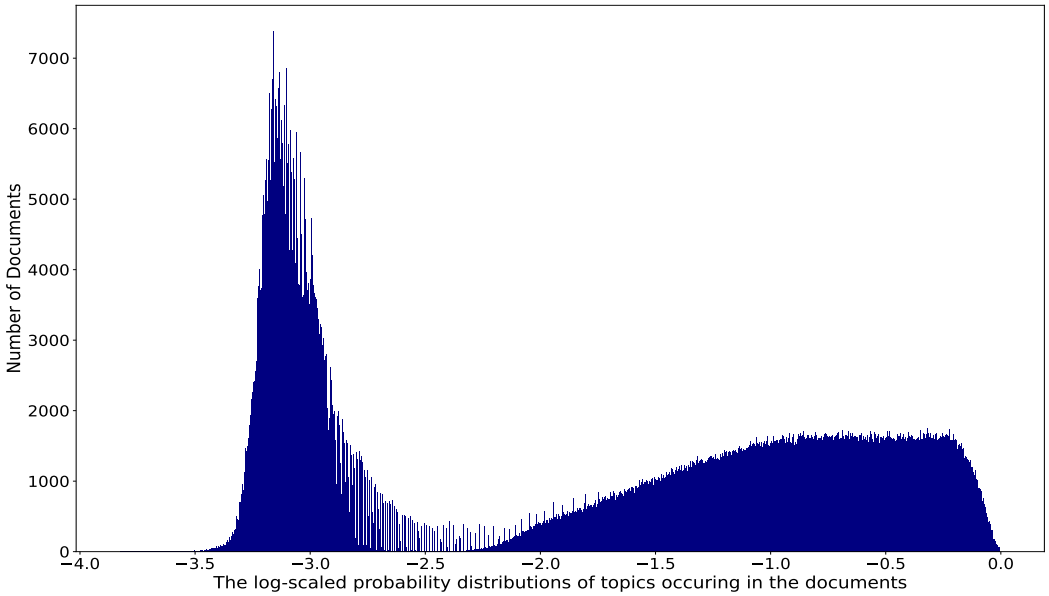

Figure 4: **Topic Distribution:** The figure shows log-scaled (with base 10) probability distributions of all topics from all documents.

The probability distribution of topics in each document might not always be helpful for topic identification since all topics can have non-zero probabilities occurring in each document [19], indicating that the probabilities for some topics are too small to be informative. Therefore, I plotted the topic probabilities throughout documents based on base 10 logarithm. Two peaks can be seen in Fig. 3, where one occurred as the probabilities are between $10^{-3.5}$ and $10^{-2.5}$, another peak is distributed from approximately 10^{-1} to $10^{-0.2}$. Hence, in order to remove all uninformative topics for each topic, any topics of which the probability is less than the minima between two peaks, at $10^{-2.3}$ (i.e. 0.005), are ignored.

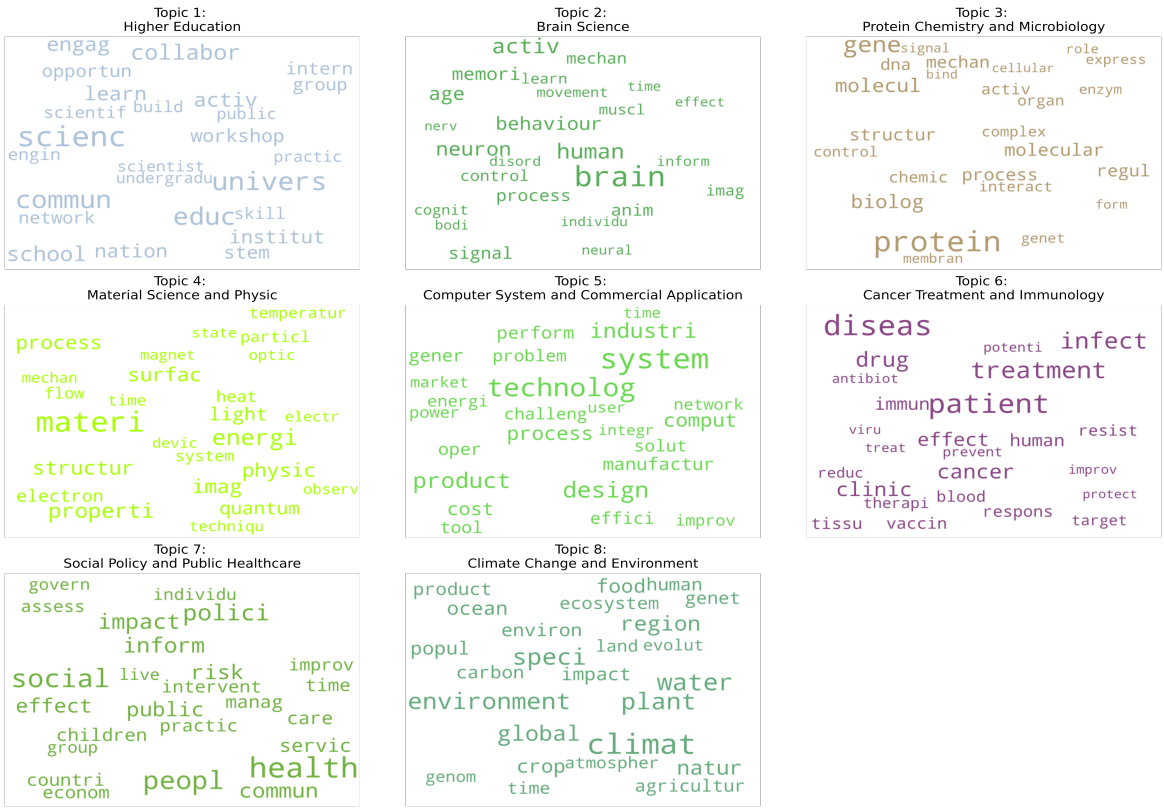

Figure 5: **Topic Wordclouds:** The figure shows the wordclouds including top 25 stemmed words of each topic derived from the LDA model, titles for the wordclouds are topic labels assigned manually.

4.3 Topic Analysis

In general, excluding the "Higher Education" Topic in ERC, the topic occurrence probabilities of topic 5 (Computer System and Commercial Application), topic 4 (Material Science and Physics) and topic 1 (Higher Education) are relatively high in NSF, ERC and UKRI, whereas other topics occurred relatively more rarely. For NIH agency, "Cancer Treatment and Immunology" (topic 6), "Protein Chemistry and Microbiology" (topic 3), "Brain Science" (topic 2), and "Social Policy and Public Healthcare" (topic 7) have higher mean topic probabilities in total, which is not surprising as NIH is a funding agency mainly focusing on medical and health-related research, supporting that my LDA model can obtain topic distributions for projects accurately.

For the group of all other agencies (Fig. 6), they are all dominated by "Computer System and Commercial Application" with the mean probability approximately $10^{-0.56}$, $10^{-0.61}$ and $10^{-0.54}$ in NSF, ERC and UKRI respectively. For the agency NIH, it is dominated by "Cancer Treatment and Immunology" whose mean probability is $10^{-0.53}$. Moreover, the "Climate Change and Environment" and "Brain Science" topics are prevalent (i.e. larger than or equal to 10^{-1}) solely in NSF and NIH respectively, whereas other topics are prevalent in 2 agencies at least.

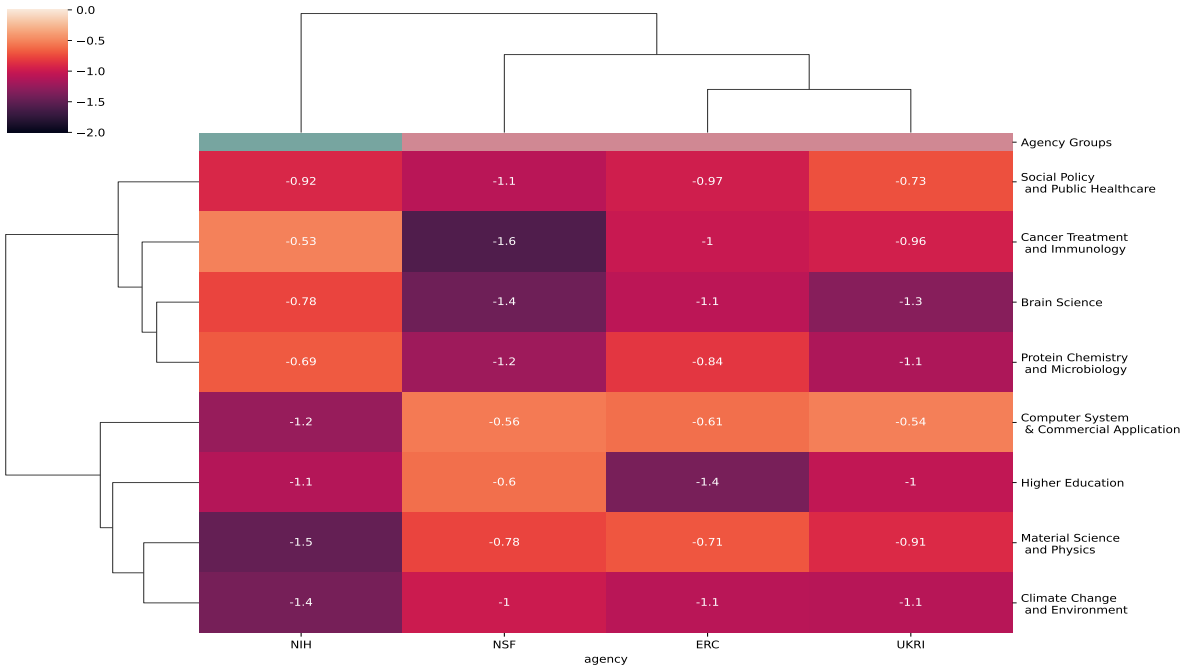

Figure 6: **Topics-Agencies Heatmap:** The heatmap between topics and four agencies. The colours inside the heatmap represent the base 10 log-scaled mean occurrence probabilities of topics. The darker is the colour, the fewer occurrence probabilities for the topic. The column and row dendrograms represent the results of hierarchical clustering on agencies and topics respectively. The two colours at the top heatmap represent different amount areas of daily funding amount: dark-green is the NIH group, dark-pink is the group of all other agencies.

For the group of all other agencies (Fig. 7A), the mean probabilities of topic "Brain Science" that occur at the low-amount and medium-amount areas are low, but becoming larger at the high-amount area of the daily funding amount. For the topics "Protein Chemistry and Microbiology" and "Cancer Treatment and Immunology", they are also less probable of occurring at the low-amount area and increasingly likely to be at the areas of daily medium- and high-amount of funding, even though they have the lowest occurrence probabilities occurring in the group of all other agencies. For the "Social Policy and Public Healthcare" topic, it is more likely to be funded at the range from 0 to 69.6 dollars and the range from 344.5 to 388.3 dollars, comparing with other intervals of the funding amount. The "Climate Change and Environment" topic has a relatively high probability of occurrence between 193.7 and 203.1 dollars, at the highest ranges of the low-amount area. Except for the interval between 0 to 36.5 dollars, the probability of appearance of the topic "Computer System & Commercial Application" is evenly distributed. The funded projects related to "Higher Education" are more probable to be funded at low-amount funding area than at the medium- and high-amount areas. For the topic "Material Science and Physics", the projects in the group of all other agencies are more likely to be funded at the medium-amount area, compared to be at the low- and high-amount areas.

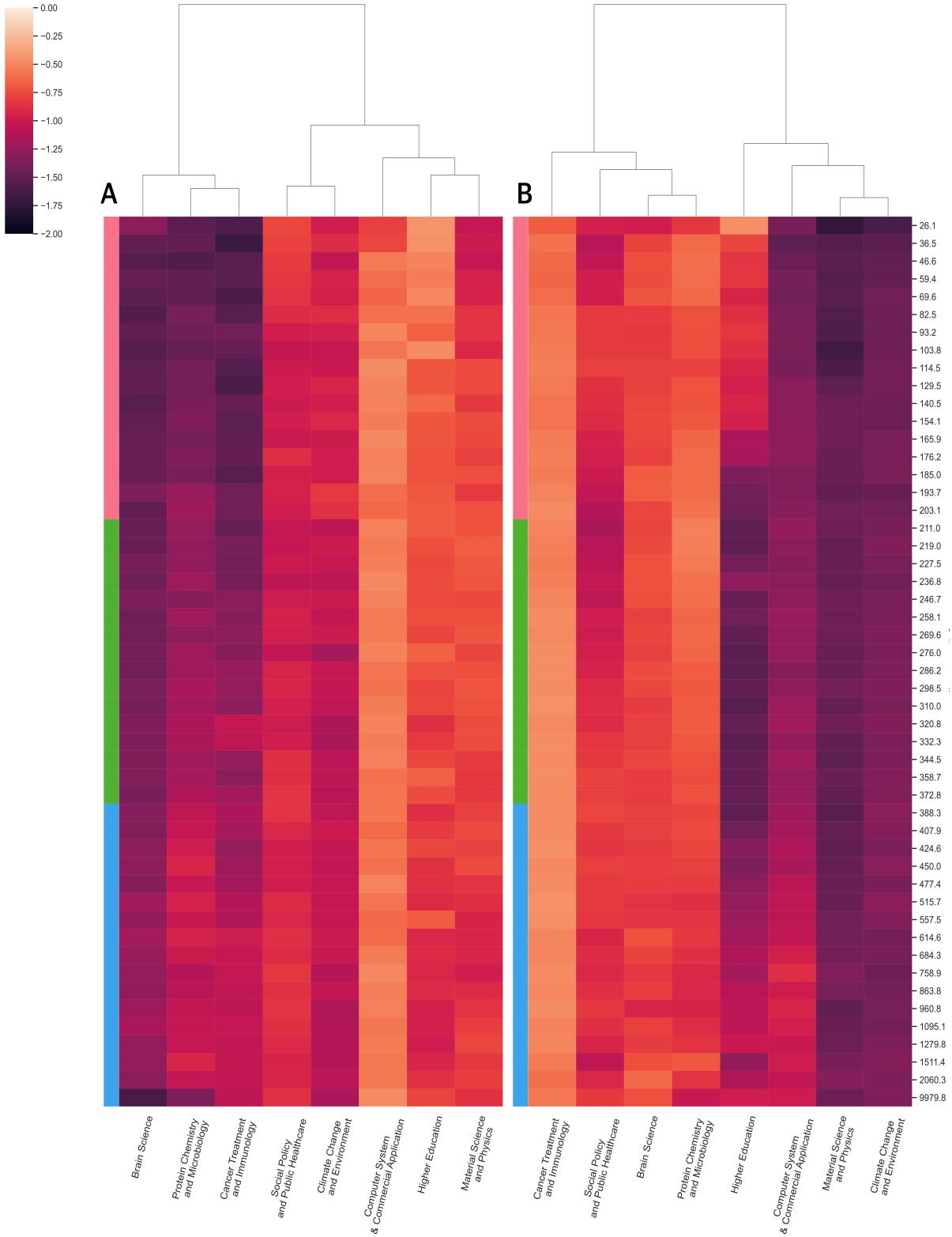

Figure 7: **Topics-Amount Heatmap:** The heatmap A and B gives the base 10 log-scaled mean occurrence probabilities of eight topics between 50 intervals of daily funding amount (see section 2.5) in NSF, ERC and UKRI agencies and in the NIH agency respectively. The vertical axis represents the discrete daily funding amount, the horizontal axis represents the agencies. Three colours on the left side of A and B describe three areas of the daily funding amount: Pink is the low-amount area, Green is the medium-amount area and Blue is the high-amount area. The column dendrogram gives the result of hierarchical clustering on topics.

For the NIH, a health-focused agency (Fig. 7B), topics "Climate Change and Environment"

and "Material Science and Physics" have equal distributions but the lowest occurrence probabilities over three areas of daily funding agencies. The probability of occurrence of the "Cancer Treatment and Immunology" topic is evenly distributed over the intervals of daily funding amount and, compared with other topics, is highest in NIH agency. There is an opposite trend that can be identified between the topics "Computer System & Commercial Application" and "Higher Education". The projects related to the former topic is more probable to be funded at a high-amount funding area between 684.3 and 9979.8 dollars each day, whereas projects in NIH agency that are relevant to the latter topic are more likely to be funded at the low-amount area between 0 dollar and 154.1 dollars per day. Furthermore, the NIH funded projects related to "Brain Science" have a higher probability of obtaining funding at ranges of 176.2–193.7 dollars and 1279.8–2060.3 dollars per day. The "Protein Chemistry and Microbiology" appears with a higher probability at the range from 26.1 and 69.6 dollars and the range from $203.1 to $227.5 per day. Additionally, the topic "Social Policy and Public Healthcare" is more likely to appear in some low intervals of daily funding amount at the low-amount and high-amount areas.

In general, the occurrence probabilities of two dominated topics ("Computer System & Commercial Application" and "Cancer Treatment and Immunology") in the NIH agency group and the group of all other agencies are both distributed evenly at the three areas of the daily funding amount. The probability of occurrence of these two topics increases over the intervals of daily funding amount if the two topics exchange groups with each other. In addition, in the group of all other agencies, with the unchanged distribution of the probability that the topic "Computer System & Commercial Application" occurs, the occurrence probabilities of "Protein Chemistry and Microbiology" and "Brain Science" topics rise with increasing daily funding amount.

4.4 Prediction of Funding Amount

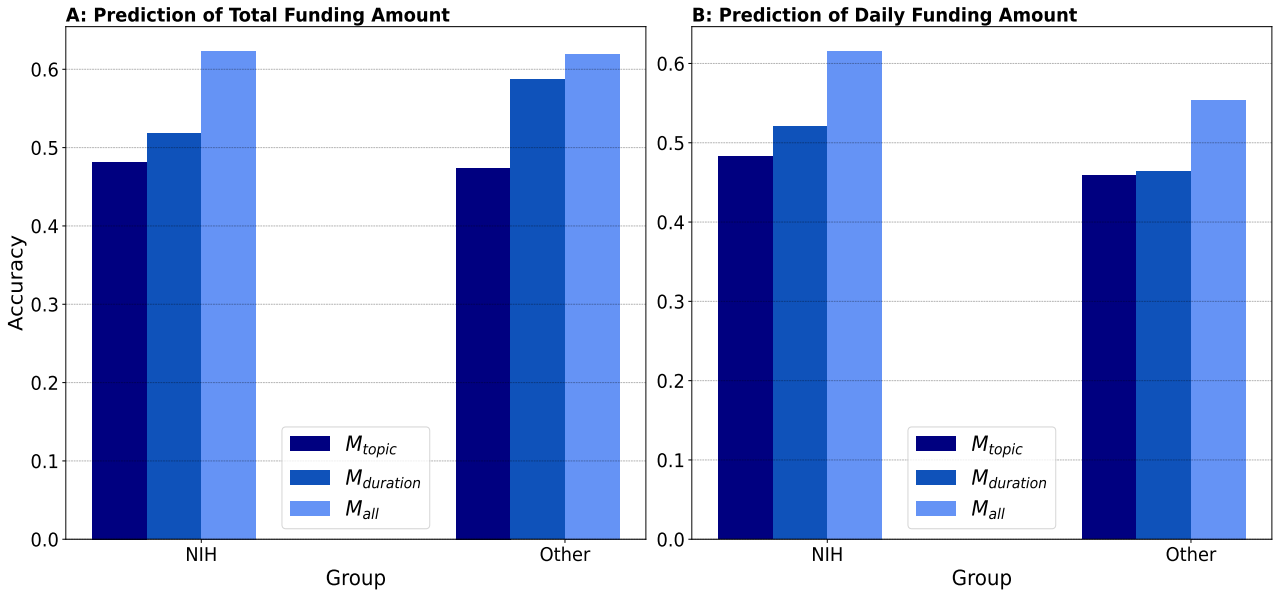

Figure 8: **Prediction Accuracy:** The accuracy of prediction of the total and daily funding amount using LightGBM in NIH group and the group of all other agencies.

As predicting the total funding amount, the M_{topic} in both the NIH group and all other agencies obtained the lowest accuracy, at approximately 0.482 and 0.474. Having only project duration as a feature, the accuracy of $M_{duration}$ in all other agencies is 0.587, 13.3% higher than that of $M_{duration}$ in the NIH group. As combining all topic distributions and project durations together as predictors, both the accuracy of M_{all} are beyond 0.6, boosting 20.4% in the NIH group and 5.5% in all other agencies compared with the accuracy of M_{topic}. When the prediction target is the daily funding amount, a different situation is seen in the group of all other agencies. The accuracy of $M_{duration}$ and M_{all} in all other agencies decreases by 10.5% and 8.5% respectively, compared with that of corresponding models in the NIH group. However, regardless of whether the prediction target is either total or daily funding amount, the accuracy of prediction is improved in all groups with the topic distributions as predictors.

5 Discussion

By using combined machine-learning approaches, I found that topic modelling techniques are helpful for topic assignment in these four funding agencies because the LDA model correctly assigned medicine- and health-related projects into NIH agency. Also, I looked into the part of project abstracts in the other three agencies and the topics also can be found interpretably. Moreover, the LightGBM might be useful for the prediction of funding amount since the accuracy of models M_{all} can beyond 0.6 as I used both the topic distributions and project duration

as prediction variables, even though the accuracy of M_{all} is poorer as predicting daily funding amount than total funding amount in the group of all other agencies. However, the accuracy is below 0.5 if I use only topic distributions as predictors to forecast either total or daily funding amount, indicating that the topic distributions generated from the LDA model don't show the prediction power.

From heatmap analysis on topics, a significant topic bias can be seen in four prevalent academic agencies. Except for NIH, which is a health-focused funding agency, the funding of the other three agencies is occupied by topics related to Computer Systems, Commercial Applications, Material Science and Physics. In addition, protein chemistry and microbiology are also prevalent topics in the ERC agency. Thus, NSF, ERC and UKRI are biased by STEM-related topics. The topic "Climate Change, Environment" are ignored significantly by all four popular funding agencies. Only two dominated topics in NIH ("Computer System and Commercial Application") and in all other agencies ("Cancer Treatment and Immunology") are equally distributed in almost 50 intervals of the daily funding amount.

However, a positive or negative relationship with daily funding amount also can be seen in two groups. For example, "Cancer Treatment and Immunology", dominating the group of all other agencies, increases its occurrence probability with funding amount in the NIH group, the occurrence probability of "Computer System & Commercial Application", dominating the NIH group, also rises with daily funding amount in the group of all other agencies. This suggests that a project that is mainly relevant to "Cancer Treatment and Immunology" and "Computer System and Commercial Application" (e.g. Medical Engineering) will have a higher probability of obtaining a higher daily funding amount in any of four agencies. Similarly, a project will be more likely to be funded with a high amount in the group of all other agencies if it includes a large portion of "Computer System & Commercial Application" with "Cancer Treatment and Immunology" or "Protein Chemistry and Microbiology" or "Brain Science". (e.g. possibly Computational Biology or Computational Neuroscience).

So far, the quantitative analysis combining the LDA model and LightGBM is first introduced in this field. Thus, there are some limitations to the project and possible solutions to be completed in the future. In this project, the dataset of four agencies is obtained in different ways, resulting in the difficulty of data collection. Therefore, one can develop a program that can automatically scrape up-to-date data from the websites of the agencies without the breach of data privacy statement, allowing the models used in the project to update Regularly.

Moreover, I used an 8-topic model by comparing models that are run multiple times with different k. However, because of time and computational restriction, only 32 topic models are included and each model is repeated only 10 times, leading the final topics to the lack of

specificity. For example, for the topic "social policy and public healthcare", I can not give a precise conclusion that the prevalence of the topic is due to "Social Policy" or because of "Public Healthcare". The problem can be solved by testing a wider range of k and more repetition times for each model if time and computational power allow.

Second, the parameters of all LightGBM models are all set to be the same on a rule-of-thumb basis. However, a better approach to parameter optimisation could be implemented in order to mitigate the risk of overfitting. For example, the Bayesian hyper-parameter optimization algorithm was utilised by Wang *et al.* to improve the performance of LightGBM [37]. Next, I used $M_{duration}$ and M_{all} to study the improvement of the model accuracy by topic distributions. However, the only project duration can not always provide sufficient information of the data, leading to poor accuracy. This might be the reason why $M_{duration}$ and M_{all} in the group of all other agencies give worse performance as predicting the daily funding amount than predicting the total funding amount. Therefore, one can add more factors that could affect funding amount to improve the accuracy of the models such as the h-index of PI and the rank of PI's university *etc.*

Then, to assess the performance of the LightGBM models, only the accuracy metric is used (see section 2.5). However, the accuracy (or precision) score can be reasonable only if the cost of false positives is high. Therefore, more approaches can be applied to assess the model performance such as recall, F1 Score and Receiver Operating Characteristic Curve. Furthermore, prediction on funding amount has been treated as a multiclassification problem. So, other prediction methods are worthy of consideration in order to acquire more specific outcomes such as regression on funding amount.

Last but not the least, due to the impossible collection of unfunded project data, the relationship between topic distributions and funding failure can not be explored yet. So, if the information of unfunded projects is available, one can use the same approaches in the research to forecast funding success and failure. For example, instead of predicting academic funding amount, a similar method was developed by Mitra *et al.* (2014) for predicting crowdfunding success and failure by penalized logistic regression with phrases produced by unigram, bigram and trigram models as features [9].

6 Conclusion

In conclusion, I answer the five questions mentioned in the introduction section. (1) With topic modelling approach, it enables us to obtain the interpretable topics from four academic funding agencies. (2) Except for the topics "Cancer Treatment and Immunology" in NIH agency

and "Computer System and Commercial Application" in all other agencies, the eight topics have different occurrence probabilities over discrete daily funding amount, meaning that the existence of topic bias in different intervals of daily funding amount and they may provide some information for prediction of the funding amount. (3) The research projects that are relevant to "Computer System and Commercial Application" and "Cancer Treatment and Immunology" possibly like computational epidemiology or medical engineering *etc.* have a higher probability of obtaining more funding in all funding agencies. However, some projects that are related to Environment, Ecosystem and Environment with Education or Social Policy are more likely to be funded with the less daily amount. (4) I found it useful for LightGBM to classify the three classes in both of total and daily funding amount, even though the accuracy is poorer as predicting the discrete daily funding amount in the group of all other agencies. (5) Using topic distributions as predictors is able to boost the accuracy of prediction by approximately 5% to 20%.

Support Information

All stop words used in this project

{", 'november', 'shouldn', 'necessary', 'inc', 'twice', 'whether', 'from', "you'll", 'aspect', 'proposed', 'measure', 'as', 'significant', 'took', 'w', 't', 'his', 'thru', 'almost', 'proposal', 'vi', 'out', 'say', 'someone', 'd', 'www', "didn't", 'mostly', 'gone', 'edu', 'f', 'never', 'h', 'chinese', 'october', 'down', 'hither', 'those', 'serious', 'specific', 'our', 'is', 'are', "you've", 'poorly', 'himself', 'program', 'i', 'become', 'awfully', 'o', 'something', 'near', 'come', 'detected', 'pursue', 'less', 'february', 'able', 'consider', 'just', 'gotten', 'until', 'going', 'changes', 'nine', 'away', 'three', 'unless', 'somebody', 'zero', 'progress', 'doi', 'like', 'upon', 'although', 'seemed', 'together', "it's", 'enable', 'behind', 'recent', 'e', 'itself', 'august', 'because', "isn't", 'iv', 'far', 'considering', 'often', 'made', 'therein', 'indeed', 'need', "shan't", 'x', 'inasmuch', 'all', 'so', 'same', 'rather', 'obviously', 'others', 'no', 'theirs', 'associated', 'l', 'thoroughly', "that'll", 'four', 'beside', 'thereby', 'needed', 'herself', 'training', 'latter', 'let', 'knows', 'main', 'exactly', 'aren', 'third', 'following', 'linkage', 're', 'beyond', 'enough', 'thats', 'not', 'll', 'application', 'got', 'lately', 'herein', 'respectively', 'cover', 'lest', "couldn't", 'but', 'value', 'focus', 'before', 'large', 'he', 'seem', 'insofar', 'truly', 'english', 'do', 'yourself', 'nor', 'beforehand', 'vii', 'toward', 'certainly', 'containing', 'second', 'been', 'sure', 'academic', 'q', 'ought', 'applicant', 'willing', 'example', 'about', 'becoming', 'currently', 'between', 'everything', 'liked', 'very', 'many', 'study', 'one', 'anyhow', 'across', 'relate', 'etc', 'forth', 'seeming', 'hopefully', 'aim', 'km', 'next', 'specifying', 'tell', 'during', 'increase', 'since', 'fifth', 'still', 'plus', 'against', 'though', 'small', 'quite', 'durham', 'oh', 'certain', 'approach', 'baseoxford', 'too', 'name', 'causes', 'theres', 'such', 'after', 'course', 'through', 'count', 'might', 'indicates', 'neither', 'future', 'examine', 'face', 'five', 'either', 'k', 'bed', 'nobody', 'wherever', 'had', 'cannot', 'kept', 'concerning', 'formerly', 'here', 'go', 'once', 'way', 'understand', 'lead', 'available', 'however', 'uses', 'alone', 'specify', 'well', 'whither', 'off', 'easily', 'last', 'thanx', 'whereby', 'below', 'immediate', 'please', 'how', 'at', 's', 'taken', "needn't", 'a', 'without', 'then', 'only', 'induce', 'seen', 'ourselves', 'six', 'http', 'take', 'further', 'give', 'moreover', 'africa', 'reasonably', 'wasn', 'least', 'okay', 'unlikely', 'everyone', 'known', 'myself', 'afterwards', 'elsewhere', 'seven', 'therefore', 'contains', "don't", 'needn', 'later', 'local', 'according', 'she', 'good', 'most', 'little', 'which', 'closely', 'highly', 'see', 'sup', 'comes', 'much', '_', 'important', 'anyone', 'japanese', 'getting', 'recently', 'always', 'p', 'david', 'ok', 'using', 'et', 'somewhat', 'entirely', 'ma', 'became', "you'd", 'dataset', 'continued', 'c', 'field', 'they', 'great', 'sometime', 'trying', 'essential', 'work', 'won', 'mustn', 'relatively', 'look', 'know', 'along', 'amongst', 'apart', 'west', 'now', 'possible', 'somewhere', 'seeing', 'definitely', 'descriptionabstract', 'viz', 'have', 'within', 'tried', 'overall project summary', 'throughout', 'any', 'nothing', 'test', 'perhaps', 'programme', 'my', 'march', 'under', 'better', 'given', 'whole', 'sheet', 'thereafter', 'goes',

'shan', 'sensible', 'channel', 'and', 'in', 'consequently', 'asking', 'each', 'greetings', 'description provided by applicant:', 'december', 'july', 'anything', 'onto', 'isn', 'extend', 'whereupon', 'having', 'to', 'vast', 'lt', 'likely', 'while', 'inward', 'meanwhile', 'april', 'hereby', 'usually', 'you', 'co', 'than', 'survey', 'why', 'be', 'outside', 'even', 'different', 'we', 'develop', 'also', 'knowledge', 'howbeit', 'successful', 'provides', 'indicated', 'y', 'nowhere', 'normally', 'went', 'whenever', 'whereas', 'investigate', 'un', 'hers', 'soon', 'affect', 'range', 'first', 'think', 'believe', 'award', 'seriously', 'latterly', 'particular', 'questions', 've', 'them', 'instead', 'secondly', 'whereafter', 'th', 'other', 'was', 'project description', 'india', 'hello', 'mainly', 'used', 'by', 'anybody', 'whatever', 'done', 'needs', 'help', 'whoever', 'various', 'this', 'couldn', 'above', 'r', 'did', "doesn't", 'cell', 'does', 'gets', 'decide', 'follows', 'somehow', 'namely', 'there', 'unto', 'haven', 'another', 'include', 'september', 'description', 'hi', 'line', 'more', 'maybe', "hasn't", 'came', 'says', 'based', 'besides', 'nevertheless', 'cause', 'g', 'thus', 'don', 'with', 'placed', 'should', 'due', 'doing', 'specified', 'own', 'yes', 'among', 'mightn', 'themselves', 'ignored', 'sorry', 'again', "shouldn't", 'nice', 'saw', 'anyway', "hadn't", "mustn't", 'june', 'show', 'successfully', 'yours', 'mean', 'right', 'contain', 'students', 'especially', 'shall', 'underly', 'china', 'its', 'can', 'looks', 'everybody', "weren't", 'ms', 'clearly', 'us', 'hence', 'anywhere', 'hasn', 'selves', 'brief', 'ltd', 'tries', 'inner', 'up', 'appear', 'sent', 'remains', 'big', 'provide', 'an', 'use', 'around', "wasn't", 'except', 'regarding', 'for', 'probably', 'receive', 'uk', 'appropriate', 'sector', "you're", 'their', 'method', 'appreciate', 'south', 'presumably', 'could', 'want', 'b', 'greatly', 'get', 'ii', 'few', "should've", 'u', 'the', 'really', 'didn', 'apply', 'hadn', 'research', 'some', 'project', 'seems', 'her', 'looking', 'tends', 'becomes', 'furthermore', 'nearly', 'despite', 'goal', 'may', 'towards', 'funded', 'ex', 'unfortunately', 'support', 'thanks', 'cant', 'try', 'none', 'whom', 'into', 'former', 'me', 'thereupon', 'ask', 'happens', 'keep', 'said', "haven't", 'thence', 'hardly', 'doesn', 'wherein', 'downwards', 'merely', 'call', 'british', 'two', 'm', 'whence', 'seminar', 'on', 'z', 'sometimes', 'yet', 'these', 'com', 'where', 'several', 'london', 'yourselves', 'subject', 'nd', 'qv', 'both', 'corresponding', 'org', 'input', 'rd', 'new', 'dr', 'otherwise', 'purpose', 'your', 'would', 'everywhere', 'best', 'particularly', 'thorough', 'make', 'supports', 'regards', "she's", 'sub', 'predict', 'area', 'actually', 'wonder', 'being', 'easy', 'lack', 'propose', 'must', 'ie', 'easet', 'data', 'him', 'will', 'thank', 'hereupon', 'wouldn', 'hereafter', 'has', 'abstract', 'of', 'https', 'indicate', 'per', 'worldwide', 'regardless', 'que', 'shown', "aren't", 'that', 'north', "wouldn't", 'vs', 'over', 'whose', 'model', 'areas', 'novel', 'weren', 'uucp', 'when', 'useful', 'gives', 'every', 'ones', 'january', 'were', 'high', 'described', 'v', "mightn't", 'accordingly', 'unique', 'n', 'old', 'am', 'ours', 'if', 'what', 'already', 'or', 'iii', 'else', 'objective', 'function', 'eg', 'anyways', 'followed', 'summary', "won't", 'it', 'estimate', 'j', 'allow', 'summaryabstract', 'saying', 'require', 'lot', 'non', 'ain', 'funding', 'overall', 'keeps', 'amountdetail', 'contact', 'wants', 'run', 'identify', 'ever', 'viii', 'self', 'experience', 'who', 'wish', 'via', 'eight', 'noone', 'aside', 'welcome', 'participation' }

Topic Similarity and Model Stability

To explore how the two topics in two sub-models at the same k-topic model differ from each other. I calculated Jensen-Shannon Divergence (JSD) between best-matched topic pairs since it has been proven to outperform other divergence-based similarity metrics [38]. Due to the stochasticity of the LDA model, topics with similar words won't be in the same order between sub-models with different random states. I applied the Kuhn-Munkres algorithm in the 'scipy' package in Python [39], also known as the Hungarian algorithm, in order to solve this minimum weight bipartite matching problem [40]. Afterwards, the JSD distance between two best-matched topics from two different sub-models of a k-topic model (see Section 2.5 Model Selection) can be calculated by using the 'Gensim' package in Python [21]. Then, from these calculated distances between topics, the averaged *similarity* between two sub-models in the same k-topic model can be expressed as follows:

$$Sim(M_u^k, M_v^k) = \frac{1}{k} \sum_{x=1}^{k} JSD(Z_{ux}, \phi(Z_{vx})) \tag{9}$$

where M_u^k and M_v^k are $u-th$ and $v-th$ sub-models in the model set with the topic indicator k; $JSD(Z_{ux}, \phi(Z_{ux}))$ are the JSD measure between the best-matched topics Z_{ix} and $\phi(Z_{ix})$, where Z_{ix} is the $x-th$ topic in M_u^k and $\phi(Z_{ux})$ is the most similar topic (i.e. best-matched) with Z_{ux} in M_v^k.

Afterwards, the *stability* between two k-topic models is expressed as:

$$Stab(M^k) = \frac{2}{n \times (n-1)} \sum_{u,v,u<v}^{r} Sim(M_u^k, M_v^k) \tag{10}$$

where n is the number of sub-models in a single k-topic model, which is 10 in my project. The topic model with the lowest *stability* and median *similarity* of pair-wise sub-models are considered as my final topic model.

Hierarchical Clustering

To explore how similar topics change with different funding agencies and funding amounts. The hierarchical clustering with the Ward linkage method, known as the minimal increase of sum-of-squares (MISSQ), is applied for clustering similar topics. Ward linkage are more explicable way to cluster noisy data [41]. The distance between two clusters can be calculated by the sum of squares of the deviations from mean or the centroid (i.e. the error of sum squares or ESS)[42].

The Ward approach tries to find the minimal increase of ESS at every iteration. The distance between clusters using the Ward approach can be expressed as [41]:

$$Dist(X, Y) = ESS(X \cup Y) - [ESS(X) + ESS(Y)] \tag{11}$$

where X, Y are two different clusters, $X \cup Y$ is a combined cluster of X and Y.

Topic Label Assignment

To assign more interpretable labels for topics derived from the LDA model, pyLDAvis is utilised, which is a Python package providing a method of interactive visualisation for the LDA model [43]. In the approach, the *relevance* measure of words of topics is defined, allowing to rank topic terms by not only the probability under a topic but also its lift. The *relevance* can be described as [43]:

$$r(w, k|\lambda) = \lambda log(\phi_{kw}) + (1 + \lambda) log(\frac{\phi_{kw}}{p_w}) \tag{12}$$

where ϕ_{kw} is the probability of a word w for topic k, p_w is defined as the marginal probability of term w, λ ($0 \leq \lambda \leq 1$) is a parameter that can tune the weight between logged probability (ϕ_{kw}) and its logged lift ($\frac{\phi_{kw}}{p_w}$). If topic terms are ranked by only logged probability estimated by the LDA model, it could affect interpretability of topics due to globally frequent tokens; if the topic terms are ranked by solely the $lift$, which could weaken the impact of globally frequent tokens but also could result in uninterpretable rare words [44]. Hence, a suitable λ value can give a solution for the problem of the non-interpretability of topics brought by either globally frequent words or field-specialised terminologies.

Topic Inference of the LDA Model

Two common approaches are utilised to infer the coefficients for document-topic and topic-word matrices [20]: variational Bayesian approach and collapsed Gibbs sampling approach. The former method is a variational Bayesian method with an Expectation-Maximization (EM) algorithm proposed by Blei et al. (2003) [19] in their original paper of the LDA model. By the variational approach, the computational process is faster but may produce inaccurate inferences [45]. The latter approach is to use collapsed Gibbs sampling to infer the coefficients of matrices, which is introduced by Griffiths and Steyvers(2004). For this method, theoretically, the accuracy of inference can be improved but with the cost of more intensive computation, which could

therefore affect the efficiency of the model as the size of datasets are larger [20].

Topic space and top-30 words of eight topics by pyLDAvis

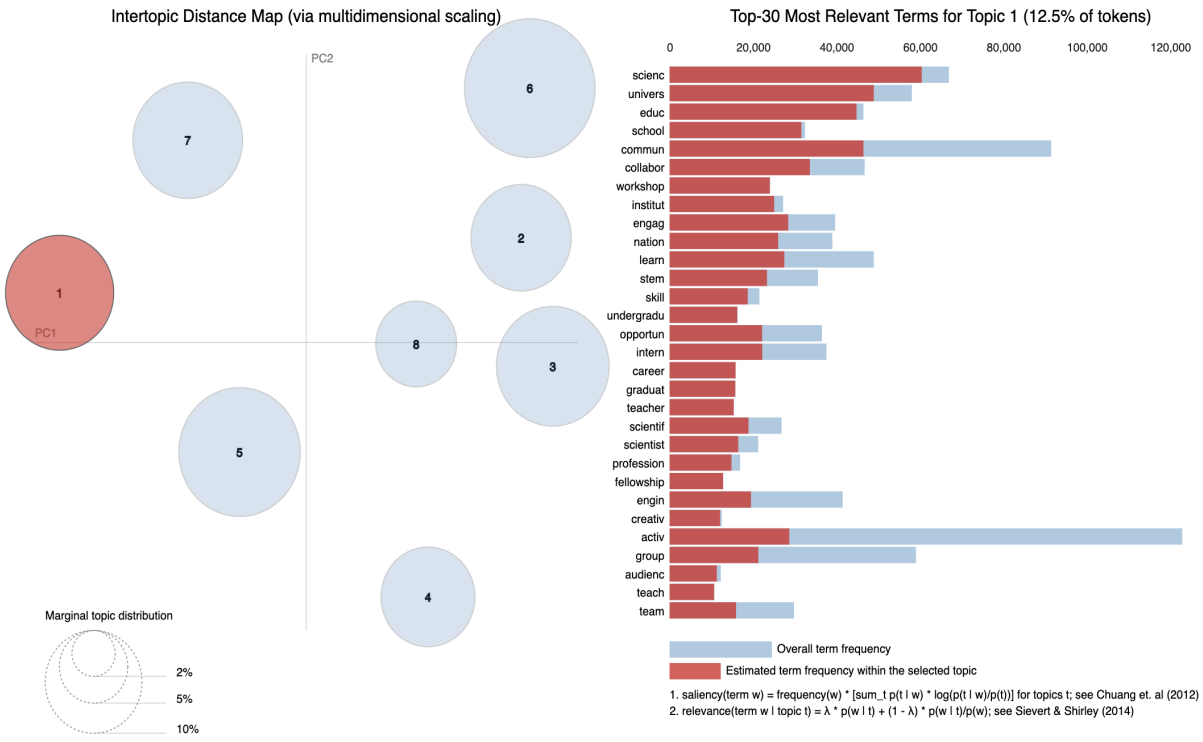

Figure 9: Visualisation of topic 1 by pyLDAvis

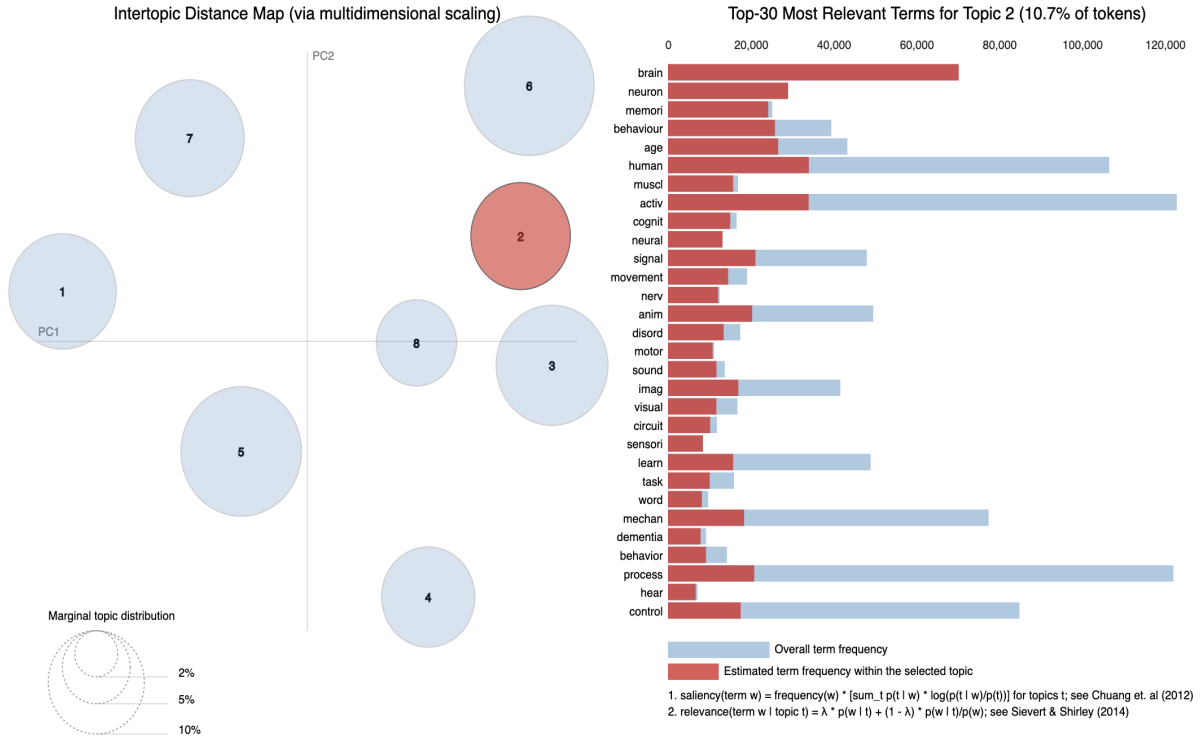

Figure 10: Visualisation of topic 2 by pyLDAvis

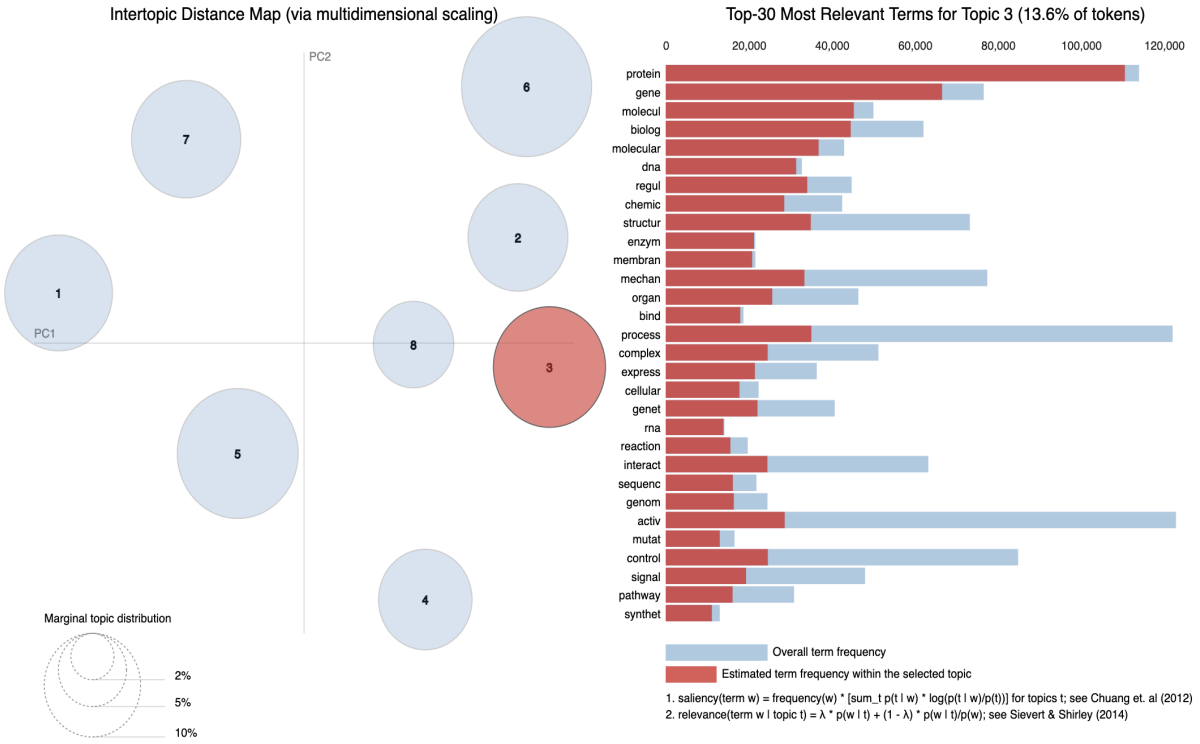

Figure 11: Visualisation of topic 3 by pyLDAvis

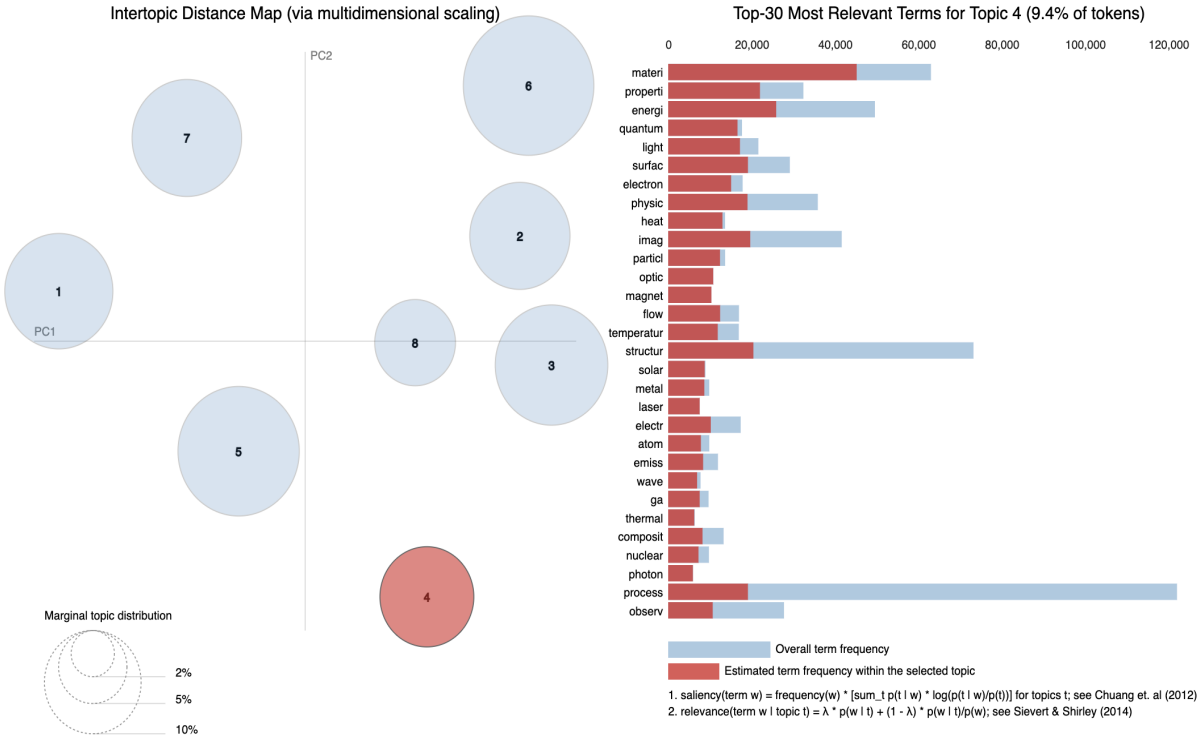

Figure 12: Visualisation of topic 4 by pyLDAvis

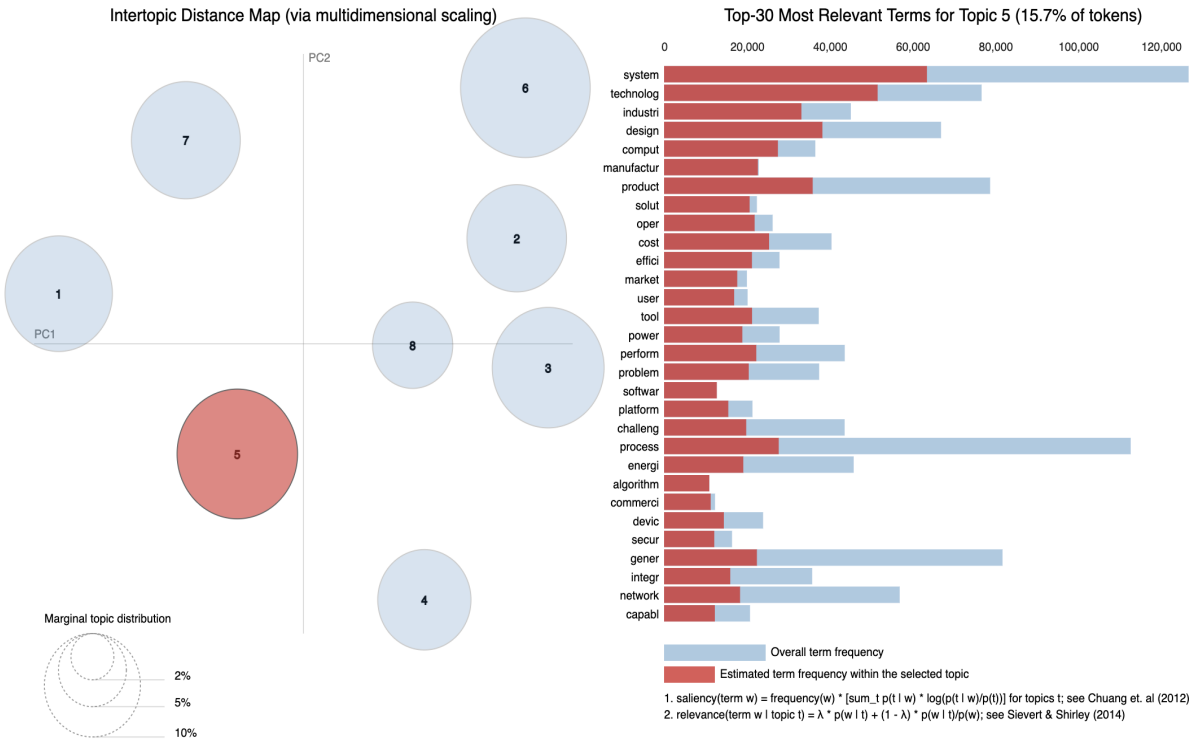

Figure 13: Visualisation of topic 5 by pyLDAvis

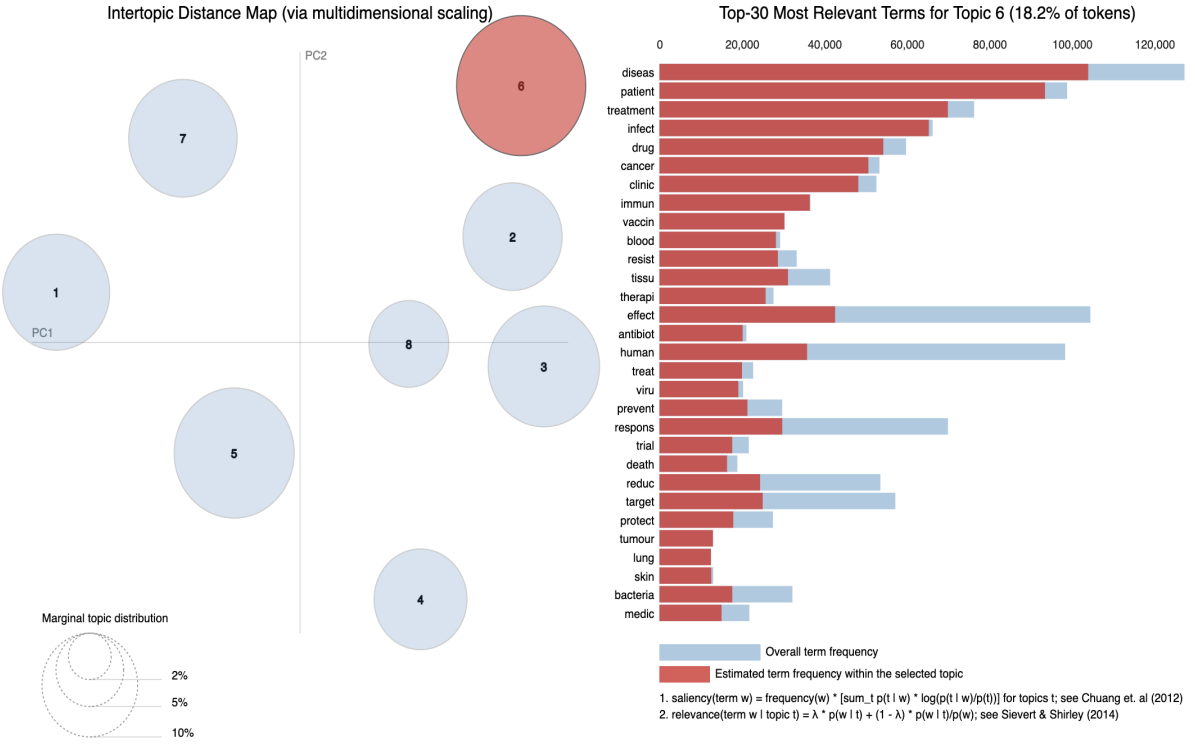

Figure 14: Visualisation of topic 6 by pyLDAvis

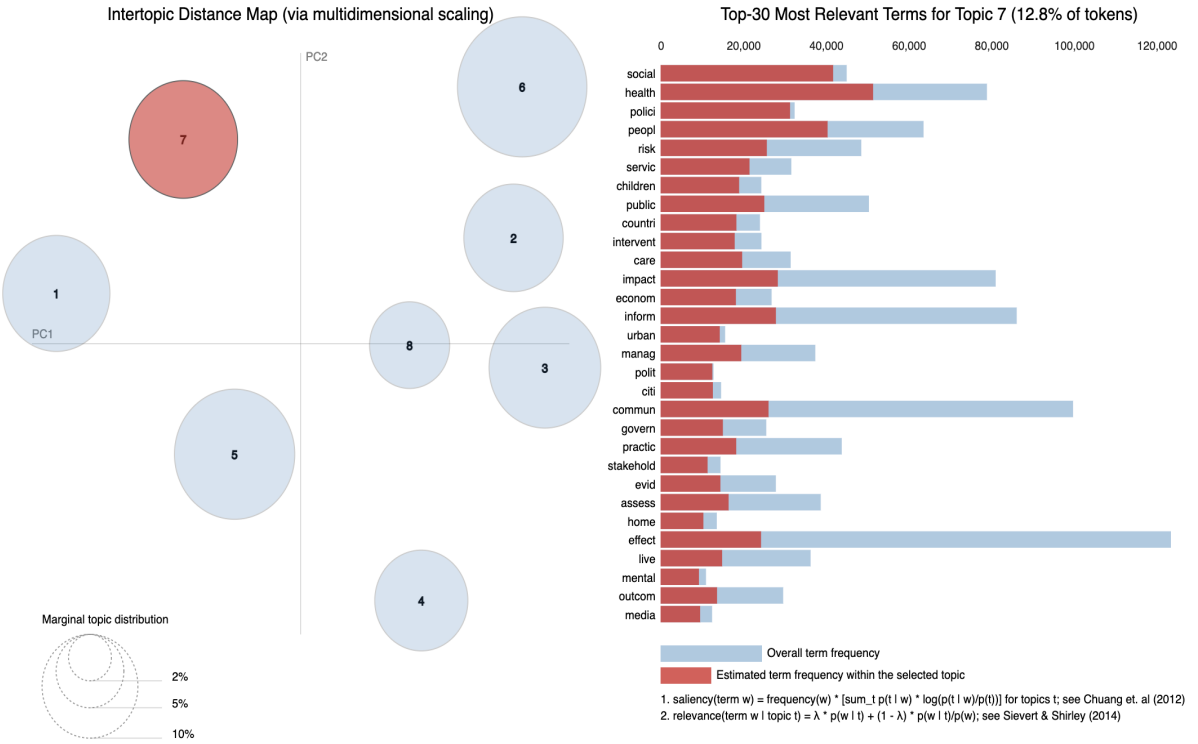

Figure 15: Visualisation of topic 7 by pyLDAvis

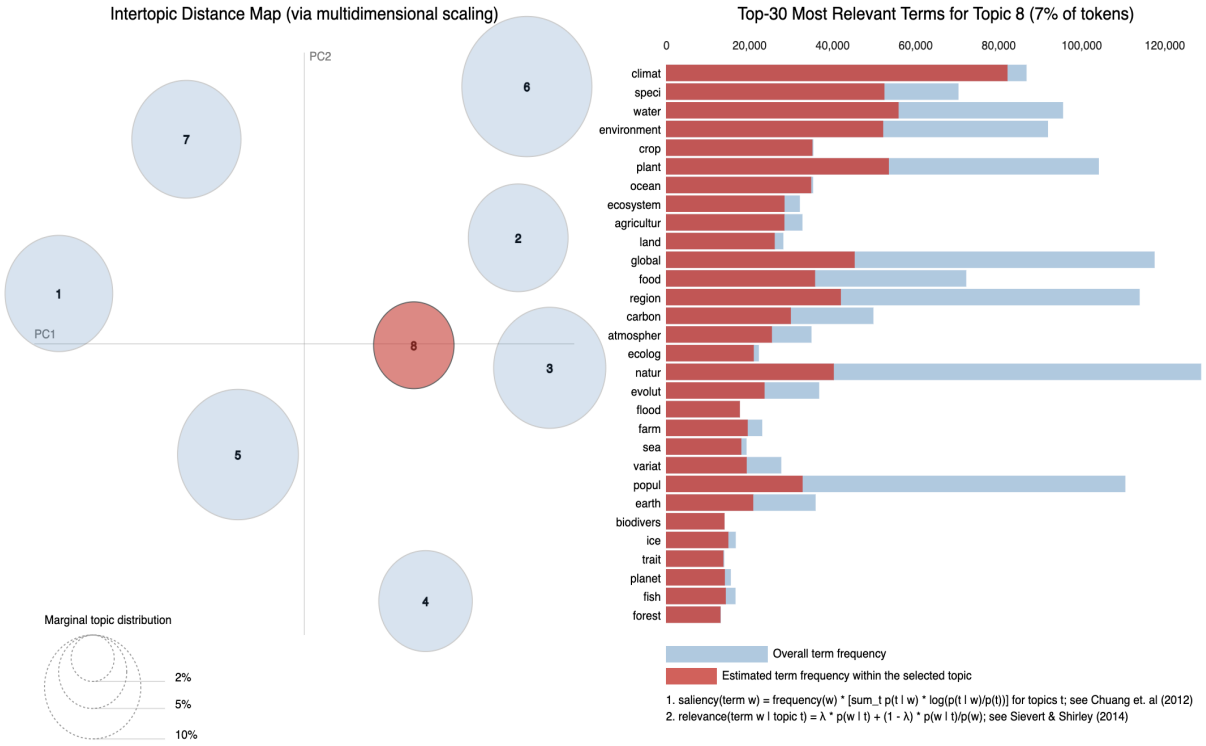

Figure 16: Visualisation of topic 8 by pyLDAvis

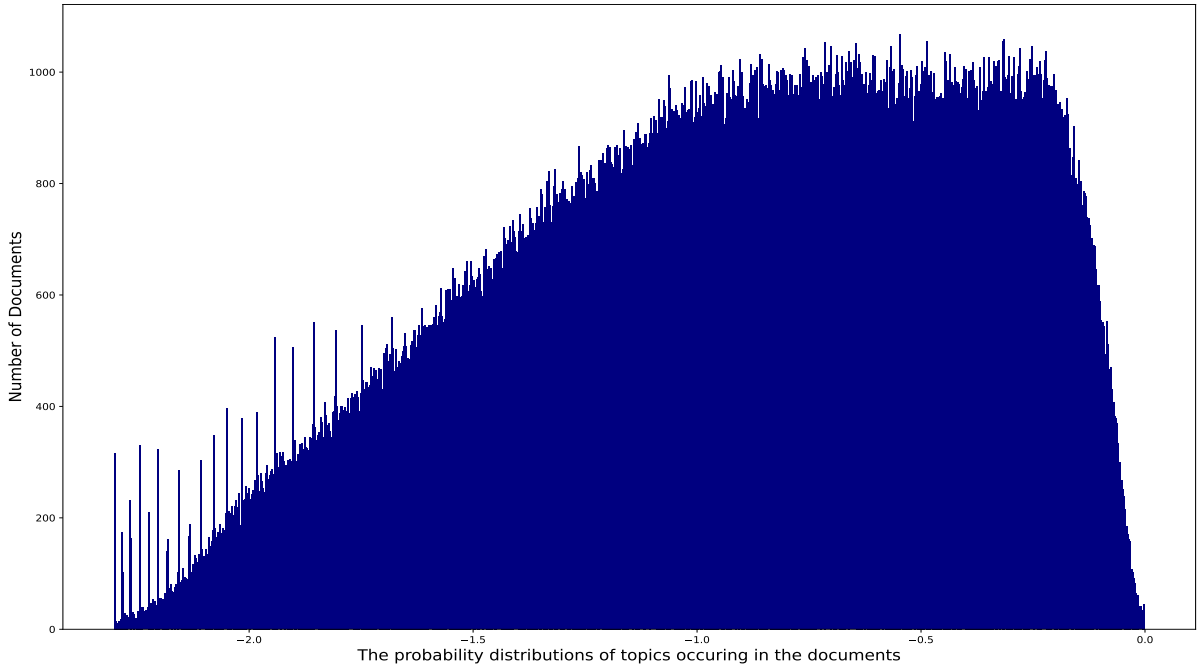

Figure 17: The probability distribution of topic probabilities that are more than 0.005 in all documents

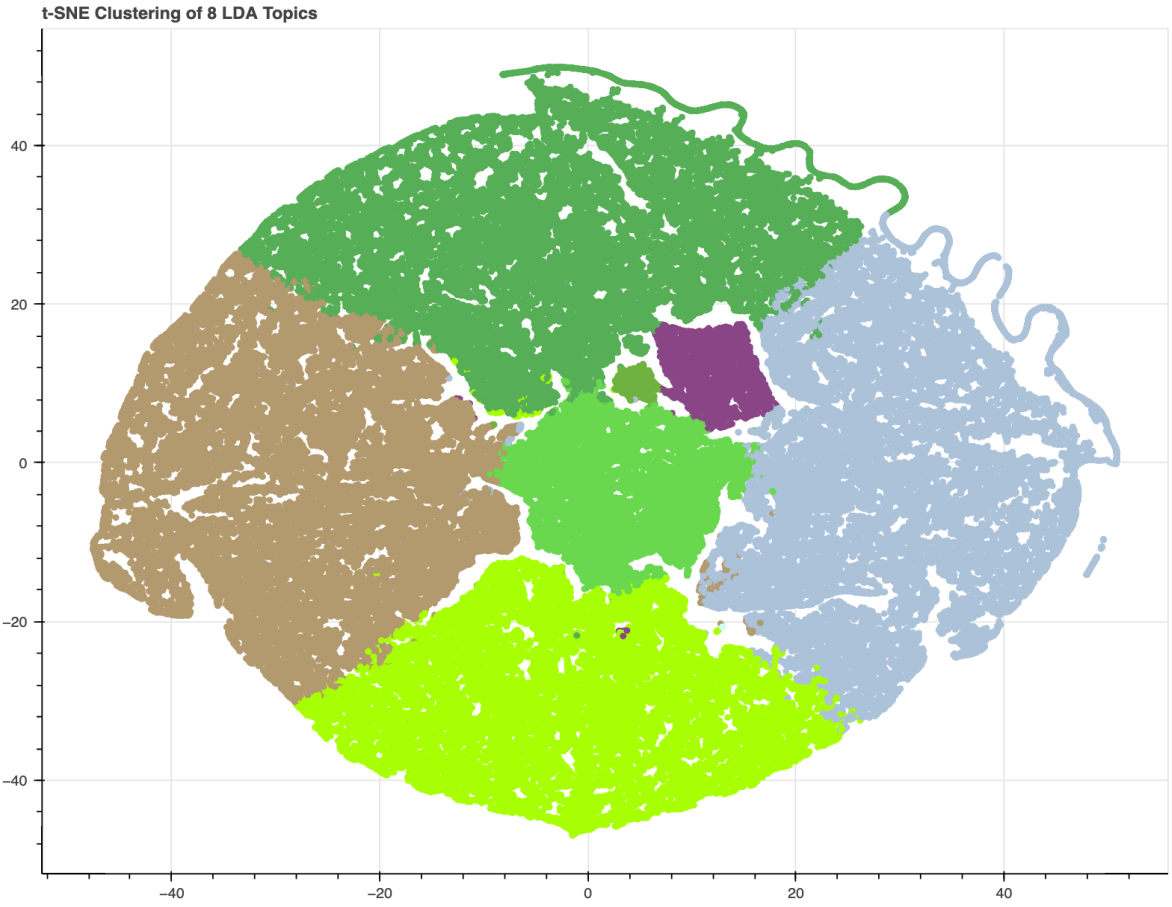

Figure 18: t-SNE clustering for visualisation of eight topics

Figure 19: Word count and importance of topic keywords

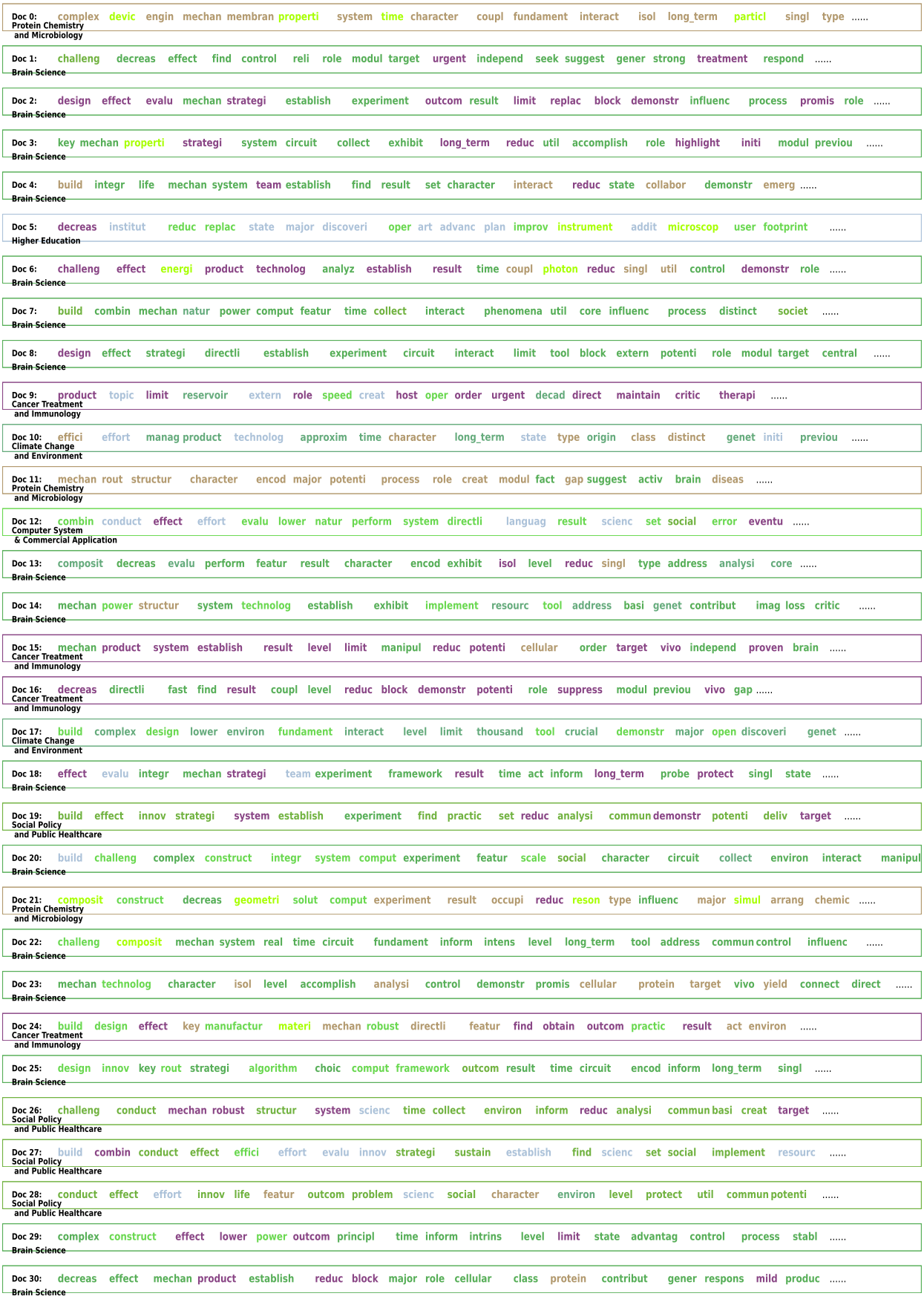

Figure 20: Sentences coloring of dominant topic in part of documents

References

[1] W. Magua, X. Zhu, A. Bhattacharya, A. Filut, A. Potvien, R. Leatherberry, Y.-G. Lee, M. Jens, D. Malikireddy, M. Carnes, and A. Kaatz, "Are female applicants disadvantaged in national institutes of health peer review? combining algorithmic text mining and qualitative methods to detect evaluative differences in r01 reviewers' critiques," *Journal of Women's Health*, vol. 26, 03 2017.

[2] D. van Dijk, O. Manor, and L. B. Carey, "Publication metrics and success on the academic job market," *Current Biology*, vol. 24, no. 11, pp. R516–R517, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982214004771

[3] D. Acuna, S. Allesina, and K. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, pp. 201–2, 09 2012.

[4] S. Allesina, "Measuring nepotism through shared last names: The case of italian academia," *PLOS ONE*, vol. 6, no. 8, pp. 1–6, 08 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0021160

[5] M. Hilbert, "Big data for development: A review of promises and challenges," *Development Policy Review*, vol. 34, no. 1, pp. 135–174, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/dpr.12142

[6] K. Kannan, P. Sekar, M. Sathik, and A. Assaf, "Stock market prediction and analysis using naïve bayes," 2016.

[7] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, 02 2014.

[8] M. Greenberg, B. Pardo, K. Hariharan, and E. Gerber, "Crowdfunding support tools: Predicting success & failure," 04 2013, pp. 1815–1820.

[9] T. Mitra and E. Gilbert, "The language that gets people to give: Phrases that predict success on kickstarter," 02 2014, pp. 49–61.

[10] B. Alberts, "Impact factor distortions," *Science*, vol. 340, no. 6134, pp. 787–787, 2013. [Online]. Available: https://science.sciencemag.org/content/340/6134/787

[11] *UK Research and Innovation GtR Database*, https://gtr.ukri.org/search/project?term=*, 2021, [Accessed: August 20, 2021].

[12] *National Science Foundation Funded Project Database*, https://www.nsf.gov/awardsearch/download.jsp, 2021, [Accessed: August 20, 2021].

[13] *National Institutes of Health RePORTER Database*, https://reporter.nih.gov/search/JbRgAtchAUKHOoW2X1vObQ/projects, 2021, [Accessed: August 20, 2021].

[14] *European Research Council Funded Project Database*, https://erc.europa.eu/projects-figures/project-database, 2021, [Accessed: August 20, 2021].

[15] F. Ahmed and A. Nürnberger, "Evaluation of n-gram conflation approaches for arabic text retrieval," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 7, pp. 1448–1465, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21063

[16] S. W. Thomas, B. Adams, A. Hassan, and D. Blostein, "Studying software evolution using topic models," *Sci. Comput. Program.*, vol. 80, pp. 457–479, 2014.

[17] R. E. Hintzen, M. Papadopoulou, R. Mounce, C. Banks-Leite, R. D. Holt, M. Mills, A. T. Knight, A. M. Leroi, and J. Rosindell, "Relationship between conservation biology and ecology shown through machine reading of 32,000 articles," *Conservation Biology*, vol. 34, no. 3, pp. 721–732, 2020. [Online]. Available: https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13435

[18] *Foreign exchange rates and currency conversion JSON API*, https://fixer.io/documentation, 2021, [Accessed: August 20, 2021].

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, Mar. 2003.

[20] Z. Qiu, B. Wu, B. Wang, C. Shi, and L. Yu, "Collapsed gibbs sampling for latent dirichlet allocation on spark," in *Proceedings of the 3rd International Conference on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications - Volume 36*, ser. BIGMINE'14. JMLR.org, 2014, p. 17–28.

[21] *Gensim: Topic Modelling for Humans*, https://radimrehurek.com/gensim/auto_examples/index.html#documentation, 2009, [Accessed: August 20, 2021].

[22] J. Chuang, M. E. Roberts, B. M. Stewart, R. Weiss, D. Tingley, J. Grimmer, and J. Heer, "TopicCheck: Interactive alignment for assessing topic model stability," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 175–184. [Online]. Available: https://aclanthology.org/N15-1018

[23] X. Wang, A. Fang, I. Ounis, and C. MacDonald, "Evaluating similarity metrics for latent twitter topics," in *ECIR*, 2019.

[24] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," 03 2013, pp. 13–22.

[25] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. [Online]. Available: https://aclanthology.org/E14-1056

[26] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. USA: Association for Computational Linguistics, 2011, p. 262–272.

[27] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," 2009.

[28] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[29] B. Fitelson, "A probabilistic theory of coherence," *Analysis*, vol. 63, no. 3, pp. 194–199, 2003. [Online]. Available: http://www.jstor.org/stable/3329309

[30] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. USA: Association for Computational Linguistics, 2010, p. 100–108.

[31] M. Papadopoulou, "The cultural evolution of evolution," 2017.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157.

[33] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[34] P. Li, "Robust logitboost and adaptive base class (abc) logitboost," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'10. Arlington, Virginia, USA: AUAI Press, 2010, p. 302–311.

[35] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview," Tech. Rep. MSR-TR-2010-82, June 2010. [Online]. Available: https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/

[36] *scikit-learn: Machine Learning in Python.*, https://scikit-learn.org/stable/about.html, 2007, version 0.24.2.

[37] Y. Wang and T. Wang, "Application of improved lightgbm model in blood glucose prediction," *Applied Sciences*, vol. 10, p. 3227, 05 2020.

[38] D. Kim and A. Oh, "Topic chains for understanding a news corpus," 02 2011, pp. 163–176.

[39] *Hungarian algorithm to solve the linear sum assignment problem*, https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html#assignment-problems, 2021, version 1.7.1.

[40] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109

[41] S. Miyamoto, R. Abe, Y. Endo, and J.-i. Takeshita, "Ward method of hierarchical clustering for non-euclidean similarity measures," in *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2015, pp. 60–63.

[42] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845

[43] C. Sievert and K. Shirley, "Ldavis: A method for visualizing and interpreting topics," 06 2014.

[44] M. A. Taddy, "On estimation and selection for topic models," 2011.

[45] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 01 1999.