Science & Technology
Facilities Council

Technical Report
RAL-TR-2017-004

A machine learning approach to the classification of technical abstracts in a two-level ontology

E Tattershall, E Yang

January 2017

Enquiries concerning this report should be addressed to:

RAL Library
STFC Rutherford Appleton Laboratory
Harwell Oxford
Didcot
OX11 0QX

Tel: +44(0)1235 445384
Fax: +44(0)1235 446403
email: libraryral@stfc.ac.uk

Science and Technology Facilities Council reports are available online at: http://epubs.stfc.ac.uk

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

A Machine Learning Approach to the Classification of Technical Abstracts in a Two-Level Ontology

E Tattershall E. Yang

January 2017

Abstract

This paper describes an approach to multi-label hierarchical document classification on an open-source corpus of 30,000 grant proposals. After text cleaning and feature extraction, an array of linear classifiers are trained and evaluated with a number of metrics, and found to classify unseen documents into 34 categories with a precision of 80%.

Acknowledgements

With thanks to L. Gilbert and F. Kayumba on the library staff at STFC, for their advice and feedback on categorisation and ontologies.

1 Introduction

Text classification is the practice of allocating documents to a predefined set of categories. Traditionally, it is done manually, requiring expert labour and a substantial amount of time per document. Since this can be impractical or prohibitively expensive for large corpora of documents, the development of robust methods of automatic text classification presents an attractive alternative.

Since this is an interesting problem at the junction between the fields of machine learning, information retrieval and natural language processing, a good deal of work has been done on the topic of computerised text classification. Examples include classification of research papers using the Association for Computing Machinery's taxonomy [7][8], news articles in the Reuters corpus [4][2], biological research papers [5] and medical research papers [12].

Our interest in text classification concerns an open-access corpus of 70,000 grant proposal abstracts. Half of the documents in the corpus are pre-labelled in a two-level ontology, while the other half are not. We intend to use the labelled half to train a classifier that can then be used to label the other half, so that we may use the labelled database to create visualisations of the data based on aggregated statistics. While this dataset has been explored using unsupervised topic modelling methods [11], it has not yet been used in supervised classification, and therefore presents an exciting opportunity to explore the use of automatic text classification techniques on a substantial, well-labelled dataset.

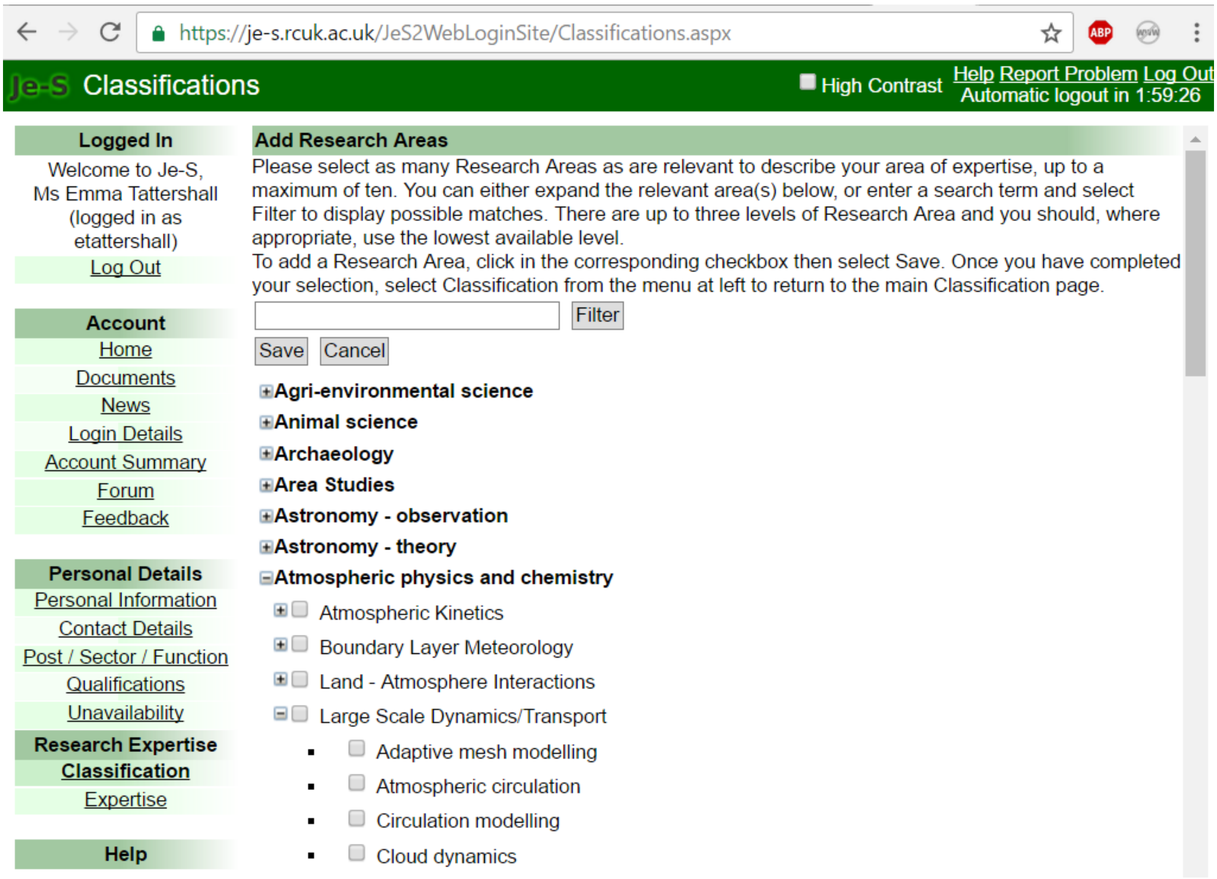
2 Dataset description

Our dataset consists of some 70,000 grant proposal abstracts hosted by RCUK's open data website Gateway to Research (GtR)[9]. Each of these grant proposals was submitted at some point in the period 1998-2017 and was funded by one of the UK's seven research councils (AHRC, BBSRC, EPSRC, ESRC, NERC, MRC and STFC) plus Innovate UK and the National Centre for the Replacement, Reduction and Refinement of Animals in research (NC3Rs).

The grants, as well as network data on individual researchers and organisations, can be downloaded via GtR's official APIs in XML and/or JSON format. Each downloaded grant contains a title, some abstract tect, information about funding (research council, start, end and value in pounds) and links to the records for associated researchers and organisations. Additionally, some grants contain more text in 'Technical Abstract' (present in 41% of documents) and 'Potential Impact' fields (present in 61% of documents), and crucially, subject area labels.

GtR supports two labelling schemes. The first is the Health Research Categories scheme (HC)[13], which is used to tag all 7000 MRC grants (10% of the grants on GtR). The second is the Joint Electronic Submission scheme (JE-S), which is used to tag 32,000 AHRC, BBSRC, EPSRC, ESRC, NERC and STFC grants. Since this second scheme is more applicable to our work at STFC, we have chosen to focus initially on this hierarchical scheme.

The full list of JE-S labels is maintained on the JE-S website [10]. When researchers submit a grant proposal, they are able to choose research area labels from a check box list on the website [See figure 1]. There are three levels to this hierarchical labelling system: *subjects* such as 'Atmospheric Physics and Chemistry', *topics* such as 'Large Scale Dynamics/Transport' and *keywords* such as 'Cloud Dynamics'. If researchers check one or more boxes when they submit their proposal, the top two levels of labels (subject and topic) are passed onto the GtR system. Users may select multiple subjects and topics (up to a maximum of 10), but the average number of subject labels is 1.9, while the average number of topic labels is 2.6. Very few grants (less than 2%) are tagged with more than 5 topic labels. Certain research councils are over-represented in this sample of 32,000 labelled grants, because their grants have a greater propensity to be labelled [See table 1]



Add Research Areas

Please select as many Research Areas as are relevant to describe your area of expertise, up to a maximum of ten. You can either expand the relevant area(s) below, or enter a search term and select Filter to display possible matches. There are up to three levels of Research Area and you should, where appropriate, use the lowest available level.
To add a Research Area, click in the corresponding checkbox then select Save. Once you have completed your selection, select Classification from the menu at left to return to the main Classification page.

⊞**Agri-environmental science**
⊞**Animal science**
⊞**Archaeology**
⊞**Area Studies**
⊞**Astronomy - observation**
⊞**Astronomy - theory**
⊟**Atmospheric physics and chemistry**
  ⊞□ Atmospheric Kinetics
  ⊞□ Boundary Layer Meteorology
  ⊞□ Land - Atmosphere Interactions
  ⊟□ Large Scale Dynamics/Transport
    - □ Adaptive mesh modelling
    - □ Atmospheric circulation
    - □ Circulation modelling
    - □ Cloud dynamics

Figure 1: The Joint Electronic Submission system's classification page

There are 80 distinct subject labels and 610 distinct topic labels used to tag GtR grants. 97 topics have more than one subject parent. A list of subjects, as well as their observed frequencies is displayed in table 2.

3 Approach to the task

3.1 The hierarchical nature of the problem

Much of the literature on classification tasks deals with flat classification problems, where there are no explicit relations between labels. However, there are still examples where researchers deal with hierarchical classification schemes similar to ours. Common approaches to the problem include:

a) Transformation into a flat classification problem: The hierarchical structure is ignored and leaf nodes are treated as independent labels, allowing conventional classification techniques to be used. While this method results in a simple algorithm, it forfeits information about the relationships between labels and can lead to an extremely multi-class problem.

b) Top down classification: A classifier is built for each decision node. New instances are fed down the tree of classifiers until they reach a leaf node. This method preserves the label hierarchy but suffers from the blocking problem – an incorrect classification at the top of the tree cannot be corrected at the lower levels of the tree, so errors are propagated down.

c) 'Big-Bang' classification: A single, complex classifier is trained on the hierarchical dataset. While this approach has been shown to be effective, the resulting classifier is very expensive to train and not flexible under changes in the label hierarchy.

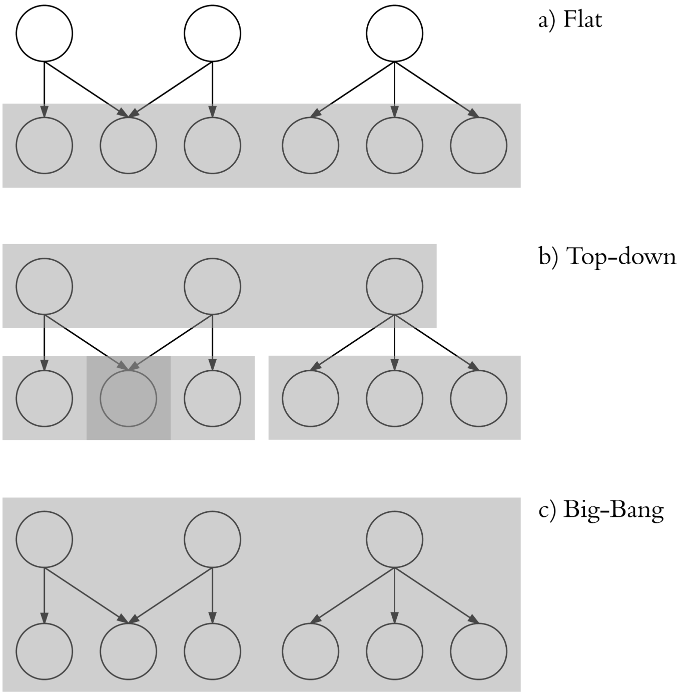


Figure 2: Common approaches to hierarchical classification. Each grey box represents one classifier.

Flat classification is unsuitable for this problem domain since it would result in a single classifier with several hundred classes! Top-down classification is appealingly intuitive and flexible, and has been shown to yield good results on similar datasets [7][6]. Therefore, we will proceed with building a single subject-discriminating classifier and an array of topic-discriminating classifiers.

3.2 Distribution of classes in the labelled and unlabelled datasets

Since we would like to eventually apply the classifier trained on the labelled half of the dataset to the half which is unlabelled, it needs to be applicable to this slightly different domain. In particular, the proportions of particular classes in the labelled and unlabelled halves of the dataset are likely to be somewhat different because some research councils are over/under-represented in each half [See table 1]. For example, just 29% of the grants funded by the Biotechnology and Biological Sciences Research Council are labelled, so we may expect to see significantly less of the subjects in the BBSRC's remit in the labelled half of the dataset. Therefore, we must be careful to avoid using any classification algorithms that learn label proportions in the training dataset, such as some variants of Naive Bayes.

Council	Number of grants	Number with JE-S labels	Proportion with labels
AHRC	5,414	4,295	79%
BBSRC	10,661	2,880	27%
EPSRC	14,466	11,894	82%
ESRC	5,571	5,061	91%
NERC	6,768	5,244	77%
STFC	3,572	2,170	61%

Table 1: Proportion of grants labelled, by research council

3.3 Class imbalances in the dataset

Even with a top-down approach, we still have an 80-class classifier at the subject level. While there are 750 examples per class on average, there exist very small classes, such as 'Dance', which has 42 examples and 'Facility development', which has 74, and large classes such as 'Information and communication technologies', which has over 3000. Additionally, since researchers often assign more than one subject label to their grants, we noticed that there are some subjects that very rarely appear alone. To illustrate this problem, see table 2 below; 'Total frequency' is the number of grants tagged with a given subject in the dataset, 'Lone frequency' is the number of grants in which that subject is the grant's only subject label and 'Mean labels' is the mean size of the set of subject labels that the given subject appears in.

Subject name	Total frequency	Lone-frequency	Mean labels
Agri-environmental science	943	107	2.9
Animal science	940	192	2.5
Archaeology	540	157	2.4
Area Studies	96	2	3.4
Astronomy - observation	906	555	1.6
Astronomy - theory	152	16	2.5
Atmospheric physics and chemistry	986	74	2.6
Atomic and molecular physics	234	60	2.1
Bioengineering	576	71	2.8
Biomolecules and biochemistry	1034	165	2.5
Catalysis and surfaces	956	177	2.3
Cell biology	691	47	2.7
Chemical measurement	544	117	2.3
Chemical reaction dynamics and mechanisms	286	49	2.3
Chemical synthesis	982	257	2.2
Civil engineering and built environment	741	310	2.2
Classics	143	60	2.0
Climate and climate change	2201	52	3.0
Complexity science	166	17	2.9
Cultural and museum studies	603	85	2.7
Dance	42	8	2.4
Demography and Human Geography	1206	116	3.2
Design	360	100	2.4
Development studies	759	64	3.0
Drama and theatre studies	286	79	2.5
Ecology, biodiversity and systematics	1336	97	2.9
Economics	1001	226	2.7
Education	700	108	2.7

Continuation of Table 2			
Subject name	Total frequency	Lone-frequency	Mean labels
Electrical engineering	182	51	2.2
Energy	1305	567	2.0
Environmental engineering	160	37	2.8
Environmental planning	366	41	3.2
Facility Development	74	22	2.5
Food science and nutrition	196	18	2.8
Genetics and development	1092	67	2.8
Geosciences	1951	420	2.6
History	1419	453	2.3
Information and communication technologies	3448	1972	1.7
Instrumentation, sensors and detectors	359	45	2.8
Languages and Literature	1103	458	2.0
Law and legal studies	550	142	2.5
Library and information studies	123	19	2.6
Linguistics	536	129	2.2
Management and business studies	991	296	2.3
Manufacturing	256	53	2.3
Marine environments	1567	62	3.0
Materials processing	532	47	2.6
Materials sciences	1857	561	2.1
Mathematical sciences	1693	990	1.7
Mechanical engineering	875	378	1.9
Media	441	98	2.6
Medical and health interface	1389	293	2.5
Microbial sciences	877	37	3.1
Music	333	176	1.9
Nuclear physics	210	123	1.8
Omic sciences and technologies	760	20	3.1
Optics, photonics and lasers	717	231	2.1
Particle astrophysics	324	61	2.3
Particle physics - experiment	820	456	1.7
Particle physics - theory	285	72	1.9
Philosophy	459	201	2.1
Planetary science	121	24	2.4
Plant and crop science	707	84	2.7
Plasma physics	180	91	1.9
Political science and international studies	956	217	2.6
Pollution, waste and resources	527	15	3.1
Process engineering	705	176	2.3
Psychology	1313	531	2.1
Science and Technology Studies	109	8	3.3
Social anthropology	452	44	3.0
Social policy	892	67	3.1
Social work	158	19	3.0
Sociology	1797	175	3.0
Solar and terrestrial physics	93	24	2.4
Superconductivity, magnetism and quantum fluids	488	176	1.9
Systems engineering	433	82	2.3
Terrestrial and freshwater environments	892	36	3.1
Theology, divinity and religion	247	81	2.2
Tools, technologies and methods	2615	318	2.7
Visual arts	726	242	2.4

Table 2: Subject labels in the JE-S ontology, and their frequencies in the GtR dataset

Initial experiments revealed that classifier predictions tended to be more accurate for classes that had a larger number of examples, appeared alone frequently and had a low number of mean labels. Classes such as Area studies, which is used to tag just 96 grants, appears in just *two* grants as the only label and is, on average, just one of 3+ labels used to tag a grant, proved impossible to distinguish. Additionally, there are other combinations of labels that are very easily confused by the classifier. For example, 'Particle physics - experiment' and 'Particle physics - theory' are very easily confused at this top level.

Therefore, we have chosen to combine some subjects (such as Particle physics - theory and Particle physics - experiment) and eliminate some altogether. We prefer to combine small classes to ensure a more balanced dataset. We have also eliminated or combined any topic labels with a dataset frequency of less than 40. The final set of 34 subjects is shown below in table 3, and the full ontology is included in the appendix.

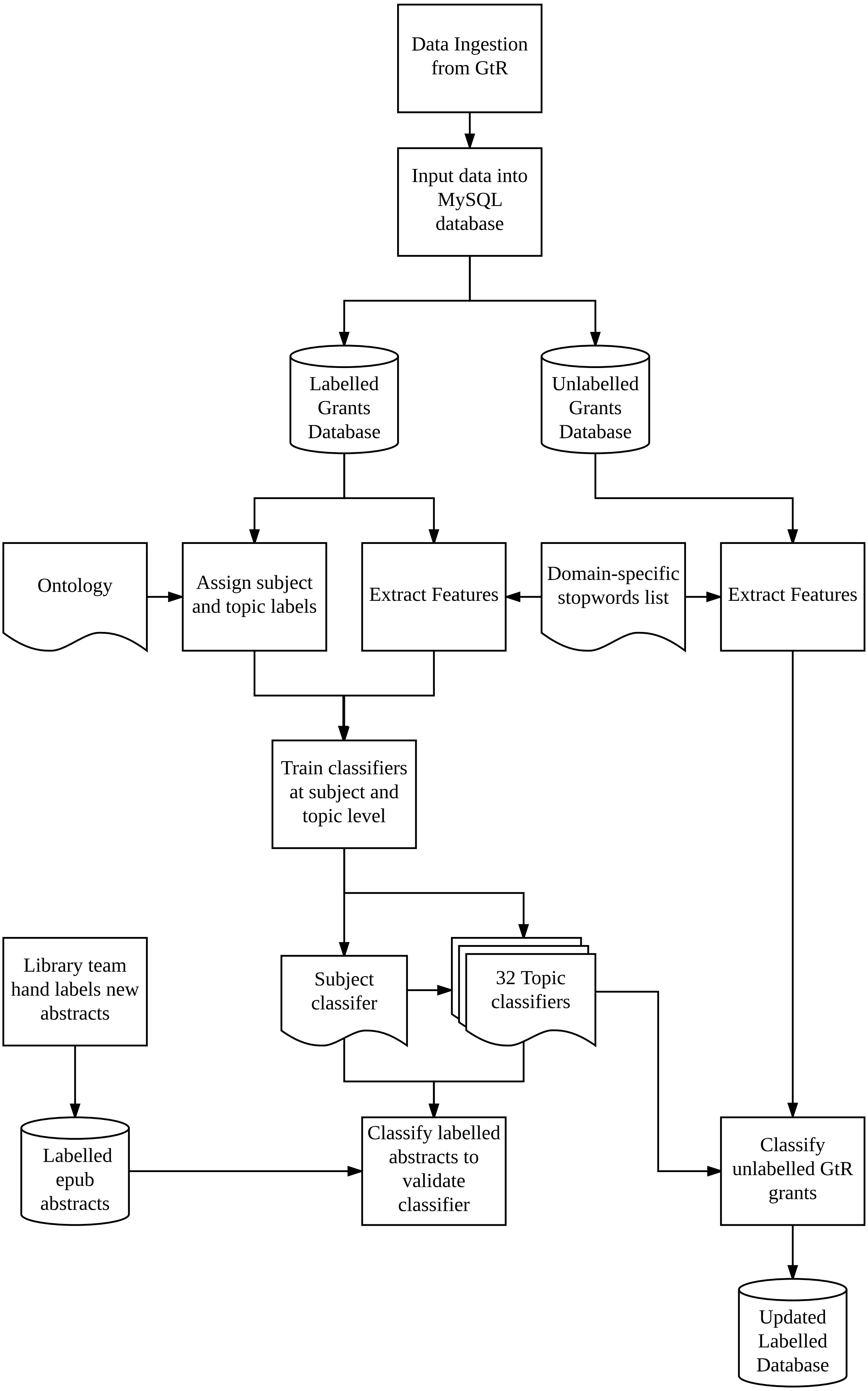
Subject name	Total frequency
Agricultural sciences	1560
Animal science	940
Archaeology	538
Astronomy	1061
Atmospheric physics and chemistry	986
Atoms, quanta and optics	948
Biochemical sciences	3276
Chemical sciences	2364
Civil engineering	791
Climate and climate change	2201
Ecology, biodiversity and systematics	1433
Energy	1390
Genetics and development	1085
Geosciences	2077
History and theology	1587
Information and communication technologies	4410
Languages, literature and creative arts	2567
Linguistics	536
Management and business studies	1123
Marine environments	1567
Materials sciences and manufacturing	2561
Mathematics	1877
Mechanical engineering	928
Medical and health interface	1741
Nuclear physics	210
Particle physics	1251
Philosophy	459
Plasma physics	180
Pollution, waste and resources	965
Process engineering	788
Psychology	1313
Social sciences	4930
Superconductivity, magnetism and quantum fluids	488
Terrestrial and freshwater environments	1004
Unclassified	595

Table 3: Subject labels used in our own ontology

While we still have some large subject classes, such as Sociology, a new class formed from the union of Sociology, Social Anthropology, Social Work, Demography and Human Geography, Political Science and Education, we have eliminated or merged the smallest classes, leading to a more balanced dataset. We also now have 595 instances that are not labelled because all of their subject labels have been eliminated, making up approximately 2% of the dataset.

Of the 34 subjects listed above, 32 are further split into topics. We have kept 222 distinct topics, 24 of which have more than one subject parent. The two subjects that we don't split further are 'Philosophy' and 'Psychology'. Both of these subjects originally had list of subject dominated by one general class (e.g. 'Psychology – general').

It should be noted that the process of combining and eliminating subjects and topics was somewhat subjective, and may be problematic in the areas where we lacked specialised domain knowledge.


Data Ingestion from GtR
Input data into MySQL database
Labelled Grants Database
Unlabelled Grants Database
Ontology
Assign subject and topic labels
Extract Features
Domain-specific stopwords list
Extract Features
Train classifiers at subject and topic level
Library team hand labels new abstracts
Subject classifer
32 Topic classifiers
Labelled epub abstracts
Classify labelled abstracts to validate classifier
Classify unlabelled GtR grants
Updated Labelled Database

4 Feature extraction

We decided to use only the plain text of the grant proposals to make classifications, and chose to omit the links to researchers and organisations (each of which are identified with a unique ID) to ensure that the resulting classifier is as broadly applicable as possible.

The length of the text portion of the grant proposals – the combination of the title, abstract text, technical abstract text and potential impact fields – ranges from 14 to 10,000 characters in the labelled portion of our dataset, with a median length of 3800, approximately 600 words. While extremely short abstracts are close to unclassifiable, they make up a very small proportion of the labelled grants; just 92 have a text length of less than 100 characters.

We chose to adopt a slightly extended bag-of-words approach for the feature extraction, using unigrams and bigrams only and placing a limit on the number of bigrams to include. The steps we took in the feature extraction process are as follows:

1. Basic text cleaning: we identified and removed HTML tags and special characters, newline and tab characters.
2. Acronym identification: acronyms with internal periods are vulnerable to being split during any subsequent punctuation-based tokenisation process. Therefore, we used regular expressions to identify such acronyms and remove internal punctuation in a sympathetic measure.
3. Tokenisation: This was based on all punctuation other than apostrophes, which were simply removed.
4. Stopword removal: We created our own extended list of domain-specific non-content words. This was based on a long open source list[1] that we pruned to fit our domain. We added non-content words that we found by clustering the raw words in a small sample of our corpus using the MALLET tool [REF], after we noticed that some of the clusters contained structural words only (e.g. fellowship, fellow, department, economy, successful, support, essential . . .). The resulting list contained 750 words.
5. Vectorisation: We converted the remaining words in our training set into a sparse matrix, with dimensions number-of-documents x vocabulary-size. We did this using scikit-learn's vectoriser tool and repeated the process for unigrams and bigrams, setting a limit of 50,000 features for the bigram vectoriser. We then concatenated the two matrices.
6. Tf-idf transformation: Tf-idf stands for term-frequency, inverse-document-frequency weighting. It is a common technique used to highlight rare, but important words by prioritising words that occur in few documents, but are repeated several times in the documents that they do occur in. We applied this weighting to our feature matrix.

While we did consider applying a stemming algorithm, since stemmers are notoriously unpredictable, we were concerned that it might have unexpected results on some of the technical terms used in the abstracts.

5 Evaluation method

5.1 Choice of metric

Since this is a multi-label problem, the methods used to evaluate a set of predictions are quite subjective and depend strongly on the situation in which the final classifier will be used. For instance, if the classifier labels an instance as 'Plasma Physics', but the true labels are 'Manufacturing' and 'Plasma Physics', how do we penalise the classifier for the missing label? In the opposite situation, in which the classifier predicts

[1] http://www.lextek.com/manuals/onix/stopwords2.html

an additional label, how do we penalise the addition? Researchers tend to use several different scoring methods in conjunction with one another – this is the approach that we have chosen to take. The metrics that we use are:

1. Exact match score (EM): A harsh metric that scores only correct sets of predictions – prediction=[1,2] and true labels=[1,2] scores 1, while prediction=[1] scores 0.
2. Hamming score: The number of correct predicted labels divided by the set of predicted and true labels. This metric attempts to penalise both extraneous labels and missing ones.
3. One-prediction score: A generous metric in which the classifier is constrained to make only one prediction, and scores 1 if that prediction matches just one of the true labels.
4. Global precision: The number of times that predicted labels are correct divided by the number of predictions.
5. Global recall: The number of times that predicted labels are correct divided by the number of true labels in the dataset.
6. Global F1 score: The harmonic mean of the two scores above, calculated as 2*P*R/(P+R). This score gives an understanding of the performance of the classifier independent of any precision-recall trade-off.
7. Individual label precision: For a given label, the number of times that it is correctly predicted divided by the total number of times it is predicted.
8. Individual label recall: For a given label, the number of times that it is correctly predicted divided by the total number of times it occurs in the dataset.

We expect that there will be ceiling on any accuracy score simply because different researchers will tag their work in different ways. In addition to assigning multiple labels to a grant, these labels are often annotated with percentages at the researcher's own discretion. It is not impossible to imagine that one researcher might give a grant a single label, while another might give the same grant one label marked 80% and four marked 5%. It is impossible to match both labelling behaviours and difficult to evaluate the extent of this problem because of its subjectivity. Therefore, we do not expect precision or recall scores above 95%.

5.2 Avoiding overfitting

We will attempt to avoid overfitting by first splitting the labelled data into a development set (90%) and a final testing set (10%). The choice of the classifier algorithm and hyperparameters will be chosen based on k-fold cross validation on this development set, in which we will be careful to avoid contamination between testing and training partitions during feature extraction (which is rerun on each of the k training folds). The final classifier configuration will then be tested on the final test set. Additionally, the library team at STFC has offered to hand-label some 200 scientific abstracts with the subject and topic categories we are using. We will use the classifier to label these and compare the results.

6 Classification

After experimenting with a Multinomial Naïve Bayes variant and Logistic Regression, we settled on using a Ridge Regression classifier. This is a linear classifier that attempts to draw separating hyperplanes between classes, while minimising its coefficients to prevent overfitting. Linear classifiers are a good fit for text classification problems like this one because the large number of features mean that the problem is likely to be linearly separable, even for the large number of classes that we use. They have been shown to require minimal tuning and no feature selection[3][1]. We chose to use scikit-learn's built in stochastic average gradient descent (sag) solver since it is capable of dealing with sparse input problems.

While scikit-learn's Ridge classifier does not natively support multi-label classification, it does return its decision function for each test instance. This is a vector in which each element can be interpreted as the distance of the instance from the separating hyperplane for each class. In binary classification, a positive value indicates a positive classification, but for our purposes, we may set our own threshold and manage the precision-recall trade-off ourselves. After some experimentation, we found that a threshold of -0.2 maximised the F1 score for a given classifier configuration.

We trained 33 classifiers: one subject-discriminating classifier and 32 topic-discriminating classifier; one for each subject in our ontology except for 'Philosophy' and 'Psychology'.

The hyperparameters of the classifiers were tuned using K-fold cross validation. We eventually settled on a relatively high regularisation constant - the parameter which controls the complexity of the model – to defend against overfitting. We were pleasantly surprised at the quality of the early results; the dataset was very good and the classifiers required relatively little tuning. We attributed this to the information-dense nature of the domain; scientific abstracts are intended to convey a great deal of information in a fairly short document.

Finally, we used the development set to determine class-by-class prediction confidence scores based on the decision function. For instance, for a given class a decision function score of -0.4 might have indicated a positive classification 30% of the time, while a score of +0.4 might have indicated a classification 95% of the time. We determined the prediction confidence for each class over decision function scores ranging from -1.0 to 1.0 and smoothed them with a Savitzky-Golay filter to allow for easy interpolation. For the evaluation of classifier performance on the unseen test set, we set the minimum confidence to 50% for each class. [See figure 3]

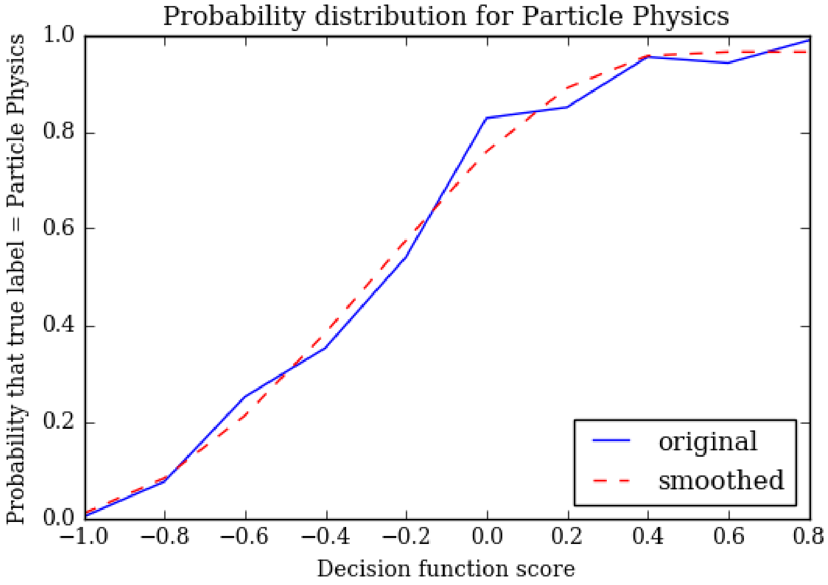


Figure 3: Decision function value and probability of correct classification for Particle Physics

7 Evaluation results and discussion

After choosing the hyperparameters, we trained our set of 33 classifiers on the entire development set (90% of the data) and tested it on the unseen testing set. The results are as follows:

Global results for each level:

Level	Exact Match	Hamming	One-prediction	Precision	Recall	F1
Subject	0.58	0.74	0.88	0.80	0.79	0.79
Topic	0.33	0.55	0.71	0.66	0.66	0.66

Table 4: Overall results on test partition

We were pleasantly surprised by the performance of the classifier at the subject level. The results at the topic level are inevitably a little worse, since the classification error is passed on. However, the performance is still adequate, with 2/3 of labels correct and 2/3 of labels matched.

Results for individual subjects:

Subject name	Subject classifier			Topic classifier average		
	Precision	Recall	F1	Precision	Recall	F1
Agricultural sciences	0.85	0.77	0.81	0.68	0.65	0.67
Animal science	0.74	0.71	0.72	0.53	0.57	0.55
Archaeology	0.84	0.80	0.82	0.53	0.62	0.57
Astronomy	0.94	0.94	0.94	0.69	0.81	0.74
Atmospheric physics and chemistry	0.85	0.86	0.86	0.75	0.84	0.79
Atoms, quanta and optics	0.81	0.74	0.77	0.67	0.67	0.67
Biochemical sciences	0.78	0.80	0.79	0.58	0.68	0.63
Chemical sciences	0.79	0.77	0.78	0.67	0.65	0.66
Civil engineering	0.68	0.80	0.73	0.58	0.66	0.62
Climate and climate change	0.86	0.83	0.84	0.81	0.78	0.79
Ecology, biodiversity and syst.	0.79	0.90	0.84	0.64	0.79	0.71
Energy	0.81	0.80	0.80	0.70	0.65	0.67
Genetics and development	0.7	0.75	0.73	0.56	0.59	0.58
Geosciences	0.90	0.88	0.89	0.82	0.80	0.81
History and theology	0.70	0.73	0.71	0.61	0.63	0.62
Information and comm. techs.	0.81	0.80	0.80	0.67	0.69	0.68
Languages, literature and arts	0.79	0.81	0.80	0.59	0.69	0.64
Linguistics	0.65	0.67	0.66	0.60	0.60	0.60
Management and business studies	0.67	0.66	0.67	0.57	0.55	0.56
Marine environments	0.87	0.83	0.85	0.83	0.75	0.79
Materials sciences and manufact.	0.80	0.81	0.80	0.63	0.72	0.67
Mathematics	0.83	0.77	0.80	0.69	0.66	0.68
Mechanical engineering	0.79	0.73	0.76	0.69	0.70	0.69
Medical and health interface	0.79	0.67	0.73	0.65	0.54	0.59
Nuclear physics	0.93	0.62	0.74	0.68	0.65	0.67
Particle physics	0.89	0.84	0.86	0.73	0.75	0.74
Philosophy	0.69	0.58	0.63	N/A	N/A	N/A
Plasma physics	0.75	0.63	0.69	0.65	0.55	0.59
Pollution, waste and resources	0.73	0.67	0.70	0.63	0.65	0.64
Process engineering	0.68	0.58	0.63	0.54	0.51	0.52
Psychology	0.79	0.74	0.77	N/A	N/A	N/A
Social sciences	0.79	0.85	0.82	0.56	0.59	0.58
Superconductivity, magnetism, quant.	0.78	0.58	0.67	0.65	0.64	0.65
Terrestrial and freshwater environs	0.85	0.73	0.78	0.80	0.72	0.76

Table 5: Subject-by subject results on the test partition. Note that due to space constraints, we do not include the individual results for each of the 200 topics: instead we state the average results at topic level for each parent subject

The table above shows that our classifiers are better at labelling some subjects than others. In particular, the results for 'Astronomy' are excellent, with the classifier scoring above 90% at subject level. The results for 'Climate and climate change', 'Ecology, biodiviersity and systematics', 'Geosciences', 'Marine environments' and 'Particle physics' are also very good. Other subjects perform less well; examples are 'Linguistics', 'Management' and 'Philosophy'. However, we do see good overall performance and find that individual performance is acceptable across the board.

8 Conclusion

We have built and described a highly flexible, tunable and specific classifier, trained on a large corpus of grants. On our test set, we observed a precision of 0.80 and a recall of 0.79 at the top level of our hierarchical scheme, and a precision and recall of 0.66 at the bottom level. Our good understanding of classifier performance on our development set has allowed us to implement a tunable measure of classifier confidence so that future users may manage their own precision-recall tradeoff to suit their application. On the whole, we found this to be a highly interesting project with an excellent dataset which yielded impressive results even with a simple classification scheme.

References

[1] Thorsten Joachims. "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features". In: *Proceedings of the 10th European Conference on Machine Learning.* ECML '98. London, UK, UK: Springer-Verlag, 1998, pp. 137–142. ISBN: 3-540-64417-2. URL: http://dl.acm.org/citation.cfm?id=645326.649721.

[2] Thorsten Joachims. "A Statistical Learning Learning Model of Text Classification for Support Vector Machines". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '01. New Orleans, Louisiana, USA: ACM, 2001, pp. 128–136. ISBN: 1-58113-331-6. DOI: 10.1145/383952.383974. URL: http://doi.acm.org/10.1145/383952.383974.

[3] Fabrizio Sebastiani. "Machine Learning in Automated Text Categorization". In: *ACM Comput. Surv.* 34.1 (Mar. 2002), pp. 1–47. ISSN: 0360-0300. DOI: 10.1145/505282.505283. URL: http://doi.acm.org/10.1145/505282.505283.

[4] Juho Rousu et al. "Learning Hierarchical Multi-category Text Classification Models". In: *Proceedings of the 22Nd International Conference on Machine Learning.* ICML '05. Bonn, Germany: ACM, 2005, pp. 744–751. ISBN: 1-59593-180-5. DOI: 10.1145/1102351.1102445. URL: http://doi.acm.org/10.1145/1102351.1102445.

[5] David Chen, Hans-Michael Müller, and Paul W. Sternberg. "Automatic document classification of biological literature". In: *BMC Bioinformatics* 7.1 (2006), p. 370. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-370. URL: http://dx.doi.org/10.1186/1471-2105-7-370.

[6] Eduardo P. Costa et al. "Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees". In: *Advances in Bioinformatics and Computational Biology: Second Brazilian Symposium on Bioinformatics, BSB 2007, Angra dos Reis, Brazil, August 29-31, 2007. Proceedings.* Ed. by Marie-France Sagot and Maria Emilia M. T. Walter. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 126–137. ISBN: 978-3-540-73731-5. DOI: 10.1007/978-3-540-73731-5_12. URL: http://dx.doi.org/10.1007/978-3-540-73731-5_12.

[7] António Paulo Santos and Fátima Rodrigues. "Multi-label Hierarchical Text Classification Using the Acm Taxonomy". In: 2009.

[8] Ekaterina Chernyak. "An Approach to the Problem of Annotation of Research Publications". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* WSDM '15. Shanghai, China: ACM, 2015, pp. 429–434. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2697032. URL: http://doi.acm.org/10.1145/2684822.2697032.

[9] *Gateway to Research.* URL: www.gtr.rcuk.ac.uk/.

[10] *Joint Electronic Submission System.* URL: https://je-s.rcuk.ac.uk/JeS2WebLoginSite/Login.aspx.

[11] James Gardiner Juan Mateos-Garcia. *Arloesiadur Data Pilot 1: Mapping research networks with open data.* URL: www.nesta.org.uk/blog/arloesiadur-data-pilot-1-mapping-research-networks-open-data.

[12] NHI. *NLM Medical Text Indexer (MTI).* URL: https://ii.nlm.nih.gov/MTI/.

[13] *UKCRC Health Research Classification System.* URL: http://www.hrcsonline.net/hc/.

9 Appendices

9.1 Full list of the subjects and topics used our classifier

Subject	Topics
Agricultural sciences	Agricultural systems Crop protection Crop science Earth and environmental Environmental Physiology Food processing Interaction with organisms Plant biology Soil science
Animal science	Animal and human physiology Animal behaviour Animal diseases Animal musculoskeletal system Animal organisms Animal welfare Endocrinology Immunology Livestock production Systems neuroscience
Archaeology	Archaeology Of Literate Societies Landscape and Environmental Archaeology Palaeobiology Prehistoric Archaeology Science-Based Archaeology
Astronomy	Astronomy and Space Science Technologies Extra-Galactic Astronomy and Cosmology Galactic and Interstellar Astronomy Planetary science Solar and terrestrial physics
Atmospheric physics and chemistry	Atmospheric Kinetics Boundary Layer Meteorology Land - Atmosphere Interactions Large Scale Dynamics/Transport Ocean - Atmosphere Interactions Radiative Processes and Effects Stratospheric Processes Tropospheric Processes Upper Atmosphere Processes and Geospace Water In The Atmosphere
Atoms, quanta and optics	Cold Atomic Species Lasers and Optics Light-Matter Interactions Optical Phenomena Optoelectronics Quantum Optics and Information Scattering and Spectroscopy
Biochemical sciences	Biochemistry Bioenergy Biomaterials Cells Environmental biotechnology Genomics Microbiology and microorganisms Tissue Engineering

Continuation of Table 6	
Subject	Topics
	Tissue engineering
Chemical sciences	Analytical Science Asymmetric Chemistry Biological and Medicinal Chemistry Catalysis and Applied Catalysis Chemical Structure Chemical Synthetic Methodology Co-ordination Chemistry Complex fluids and soft solids Electrochemical Science and Engineering Gas and Solution Phase Reactions Physical Organic Chemistry Surfaces and Interfaces Synthetic biology
Civil engineering	Civil Engineering Materials Coastal and Waterway Engineering Ground Engineering Mining and Drilling Structural Engineering Urban and Land Management Waste Engineering Water Engineering
Climate and climate change	Climate and Climate Change (General) Palaeoenvironments Regional Weather and Extreme Events
Ecology, biodiversity and systematics	Behavioural Ecology Community Ecology Conservation Ecology Environmental Physiology Population Ecology Systematics and Taxonomy
Energy	Bioenergy Carbon Capture and Storage Coal Technology Energy - Conventional Energy - Marine and Hydropower Energy - Nuclear Energy Efficiency Energy Storage Fuel Cell Technologies Solar Technology Sustainable Energy Networks Sustainable Energy Vectors Wind Power
Genetics and development	Animal developmental biology Epigenetics Gene action and regulation Population Genetics/Evolution
Geosciences	Earth Resources Earth Surface Processes Geohazards Glacial and Cryospheric Systems Hydrogeology Mantle And Core Processes Palaeoenvironments Properties Of Earth Materials Quaternary Science Sediment/Sedimentary Processes

Continuation of Table 6	
Subject	Topics
	Tectonic Processes Volcanic Processes
History and theology	History Theology, divinity and religion
Information and communication technologies	Artificial Intelligence Technologies Complexity Science Computer Graphics and Visualisation Control Engineering Design Engineering Digital Signal Processing Electronic Devices and Subsystems Fundamentals of Computing High Performance Computing Human-Computer Interactions ICT Networks and Distributed Systems Image and Vision Computing Microsystems Mobile Computing Music and Acoustic Technology New and Emerging Computing Paradigms Optical Communications Optoelectronics Power Systems Radio Frequency (RF) and Microwave Technology Robotics and Autonomy Software Engineering System on Chip Vision, Hearing and Other Senses - Applications in ICT
Languages, literature and creative arts	Classics Dance Drama and theatre studies Languages and Literature Media Music Visual arts
Linguistics	Computational Linguistics Non-Computational Linguistics
Management and business studies	Building Operation and Management Construction Operations and Management Management and business studies Manufacturing Enterprise Operations and Management Transportation Operations and Management
Marine environments	Biogeochemical Cycles Ecosystem Scale Processes Land - Ocean Interactions Ocean - Atmosphere Interactions Ocean Circulation
Materials sciences and manufacturing	Biomaterials Complex fluids and soft solids Manufacturing Materials Characterisation Materials Synthesis and Growth Materials processing Microsystems
Mathematics	Algebra and Geometry Complexity Science Continuum Mechanics Logic and Combinatorics

Continuation of Table 6	
Subject	Topics
	Mathematical Analysis Mathematical Aspects of Operational Research Mathematical Physics Non-linear Systems Mathematics Numerical Analysis Statistics and Applied Probability
Mechanical engineering	Acoustics Aerodynamics Combustion Engineering Dynamics and Tribology Instrumentation Engineering and Development Materials testing and engineering Robotics and Autonomy
Medical and health interface	Biomaterials Biomechanics and Rehabilitation Biomedical neuroscience Diet and health Drug Formulation and Delivery Environment And Health Medical Imaging Medical Instrumentation, Devices and Equipment Medical science and disease Mental Health Tissue Engineering
Nuclear physics	Nuclear Astrophysics Nuclear Structure Relativistic Heavy Ions
Particle physics	Accelerator Research and Development B Physics/Flavour Physics Beyond The Standard Model Cosmology Direct Dark Matter Detection Gravitational Waves Mathematical Physics Neutrino Physics Relativistic Heavy Ions The Standard Model
Philosophy	
Plasma physics	Plasmas - Laser and Fusion Plasmas - Technological
Pollution, waste and resources	Assessment of Contaminated Land and Groundwater Earth Resources Ecotoxicology Pollution Waste Engineering Water Quality
Process engineering	Bioprocess Engineering Design of Process systems Fluid Dynamics Food processing Heat and Mass Transfer Mining and Drilling Multiphase Flow Particle Technology Reactor Engineering Separation Processes
Psychology	
Social sciences	Demography and Human Geography

Continuation of Table 6	
Subject	Topics
	Economics Education Law and legal studies Political science and international studies Social Statistics, Computation and Methods Social anthropology Social policy Social work Sociology
Superconductivity, magnetism and quantum fluids	Condensed Matter Physics Magnetism/Magnetic Phenomena Quantum Fluids and Solids
Terrestrial and freshwater environments	Biogeochemical Cycles Earth Surface Processes Ecosystem Scale Processes Land - Atmosphere Interactions Land - Ocean Interactions Soil science Water Quality

Table 6: Subjects and topics in the ontology we used in the classification

9.2 Best features for each class

Our Ridge classifier works by training a matrix of coefficients with dimensions (number of classes x number of features). For each class, the most important features are those with the largest magnitude coefficients in its row in the matrix. If the coefficient of a given feature is positive in the row for a given class, its presence in the document increases the likelihood that the document belongs to that class. On the other hand, if the feature has a negative coefficient, it decreases the likelihood of the document belonging to that class.

Lists of the most important features for each class are shown below. Note that the most powerful negative features tend to disambiguate between similar classes or clear up misunderstandings; e.g. for 'Agricultural sciences', 'plant' is a strong positive feature while 'power plants' is a strong negative.

Subject	Top 10 positive features	Top 10 negative features
Agricultural sciences	wheat, pests, agricultural, food security, growers, plants, plant, crops, arabidopsis, crop	land surface, tropical forests, microorganisms, vaccine, carbon stored, diatoms, power plants, bioenergy, plant diversity, river
Animal science	understanding brain, cortex, animal welfare, neuroscience, neurons, veterinary, poultry, animals, animal, livestock	human brain, brain injury, plants, sexual selection, growth factor, specific proteins, disease causing, ensembl, effects host, proteins
Archaeology	ancient, excavations, neolithic, excavation, fossils, prehistoric, fossil record, archaeologists, archaeology, archaeological	latin inscriptions, history heritage, papyri, abrupt climate, ma, troy, latin, europe north, limbed vertebrates, earths climate
Astronomy	clover, telescopes, stars, astronomical, astronomers, space weather, galaxies, astronomy, astrophysics, telescope	nuclear reactions, supersymmetric, ppgp, matter particles, upgraded, agreed related, birth stars, silica suspension, cover travel, epsrc
Atmospheric physics and chemistry	radiative, stratosphere, complex terrain, air, land surface, troposphere, ozone, aerosol, atmosphere, atmospheric	dioxide methane, ocean, climate change, pco2, combustion, drought, sea level, antarctica, tree mortality, carbon dioxide

Continuation of Table 7		
Subject	Top 10 positive features	Top 10 negative features
Atoms, quanta and optics	light, wavelength, trap, laser sources, ultracold, lasers, quantum, laser, photonics, optical	quantum algorithms, atomic scale, infrared laser, single spin, terahertz frequencies, magnetic, quantum condensates, luminescent, plasma interactions, division multiplexing
Biochemical sciences	sequencing, microorganisms, cell, protein, genome, bacteria, enzymes, biomass, proteins, microbial	coding rnas, barley, wastewater treatment, dna methylation, enzyme responsive, control gene, disease resistance, crop species, drug molecules, dna sequence
Chemical sciences	synthesis, chemical, soft matter, catalytic, molecule, molecular, synthetic, catalysts, molecules, chemistry	materials synthesis, electron microscopy, fuel cell, oxides, membranes, biosensors, substrate, biocatalysis, electrical conductivity, polymer materials
Civil engineering	urban, resource recovery, concrete, pipe, civil engineering, sustainable urban, earthquake engineering, wastewater, geotechnical, wastewater treatment	integrity, hydrological, marine energy, extreme events, flexible risers, validation, investors, model problem, fuel, software engineering
Climate and climate change	earths climate, observations, warming, aerosol, climatic, osmosis, complex terrain, ice, ocean, climate	past ice, terrestrial ecosystems, marine environment, functional diversity, budgets, soils, beneath ice, forest degradation, microbial, acetone
Ecology, biodiversity and systematics	ecology, ecosystems, offspring, ecosystem, habitats, evolutionary, conservation, biodiversity, ecological, species	diatoms, sex chromosomes, marine ecosystem, chromosomes, genes species, genetic variation, soils, hybrid zones, ecosystem responses, sequencing
Energy	pv, hydrogen storage, nuclear, fuel cell, fuel, solar cells, supergen, photovoltaic, electricity, energy	fuel economy, aerospace, lithium oxygen, emissions co2, electronics, catalysis, nuclear physics, demand uk, graphene, power electronics
Genetics and development	parasites, epigenetics, genetic diversity, epigenetic, developmental, genetics, genome, evolutionary, genetic, genes	gene silencing, genes proteins, senescence, host pathogen, archaea, vaccination, diversity panel, rat, biological evolution, throughput
Geosciences	rocks, minerals, seismic, earths, earth, hydrological, volcanic, geological, mantle, sediment	climate sensitivity, geological structures, food webs, earthquake engineering, air, tropical, oceanography, salinity, air sea, swash zone
History and theology	religious, war, christian, buddhist, medieval, empire, world war, historians, historical, history	literary, theatre, art historians, fiction, art history, writers, english heritage, architectural, art historical, sculpture
Information and communication technologies	digital, software engineering, software, reconfigurable, complexity, privacy, automatically, robotics, computer, algorithms	numerical, computer modelling, qubits, computing power, dirac, random, ultra precision, throughput, ska, plasma
Languages, literature and creative arts	musical, arts, art, artists, cinema, film, creative, music, theatre, literary	creative industries, historians, writing history, traditional cultural, late antique, century britain, state art, religious, medieval culture, digital resources
Linguistics	languages, words, natural language, linguists, speech, grammatical, corpus, linguistics, language, linguistic	written language, spoken language, english language, language skills, programming language, literary, dialectal, nonwords, programming languages, modelling language

Continuation of Table 7		
Subject	Top 10 positive features	Top 10 negative features
Management and business studies	employees, innovation, firm, managerial, buildings, corporate, managers, business, organisational, firms	agricultural, uk energy, banking, messages, expected, implications policy, historic buildings, transnational corporations, service providers, models relate
Marine environments	oceans, carbon stored, osmosis, climate, sea, seawater, ecosystem, ecosystems, marine, ocean	past climate, ipcc, dioxide levels, tropics, asian monsoon, earth scientists, marine energy, coastal waters, mass loss, iodp exp
Materials sciences and manufacturing	nanostructured, ceramic, alloys, polymers, ferroelectric, graphene, manufacturing, biomaterials, polymer, materials	energy density, photonics, optical communications, spintronic, light induced, pv, innovative solutions, plasmonic, power electronics, materials electronic
Mathematics	algebraic, problems, theory, random, statisticians, stochastic, statistical, mathematicians, mathematics, mathematical	logical, sublinear algorithms, qcd, empirical, sublinear, correlated, csps, lyapunov, cfd, mathematics education
Mechanical engineering	vibration, engine, fatigue, aerodynamic, robot, nde, robots, robotics, acoustic, combustion	civil engineering, control engineering, wave problems, robotic control, corrosion resistance, unsteady flows, electrical power, electric aircraft, sliding mode, metal forming
Medical and health interface	patient, health policy, brain, bioactive, human health, lung, biomaterials, bone, tissue, clinical	stem cell, specific cell, deficits, cellular, medical professionals, depression, neural basis, related diseases, help doctors, illness
Nuclear physics	follow grant, atomic nucleus, gsi, h710003, hadrons, nuclear reactions, nuclear, gravitational radiation, nuclei, nuclear physics	astrophysics reactions, matter antimatter, neutron gamma, xmmnewton, nustar uk, nustar fair, elementary particles, ukqcd, physics astrophysics, particle accelerator
Particle physics	lhc, collider, t2k, integrable, ppgp, string theory, cta, particle physics, hllhcuk, wakeham	silicon detector, antihydrogen, pure mathematics, institute astronomy, universe studying, patt, radio telescopes, plasmas, laser driven, geometric objects
Philosophy	philosopher, political philosophy, epistemology, political thought, argue, moral, ethics, philosophical, philosophers, philosophy	literary theory, literary, film theory, writing, methodological, fourth century, church, cultural life, french thought, visual culture
Plasma physics	warm dense, tokamak, dense matter, fusion energy, microplasmas, laser driven, plasma physics, fusion, plasmas, plasma	plasma membrane, plasma assisted, astrophysical plasmas, solar surface, mhd turbulence, hipims, laser beam, solar wind, pulsed laser, sun solar
Pollution, waste and resources	resource recovery, fish, ore, shale gas, contaminated, deposits, nerc, aquatic, pollutants, pollution	soil organic, flood, variability, food webs, heavy metals, earths interior, rock types, environmental change, atmospheric aerosols, bed
Process engineering	heat, bubble, continuous flow, reactors, batch, purification, reactor, refrigeration, multiphase, chemical engineering	flow induced, turbine, chemical reactions, biocatalysis, product market, permeability, gas separation, hydrogen fuel, cfd modelling, computational modelling
Psychology	perceptual, social psychology, series experiments, individual differences, memory, psychosocial, psychologists, cognitive, psychology, psychological	tongue, native language, mental states, psychiatry, cognitive function, social class, subjects, language teaching, deaf people, children young

Continuation of Table 7		
Subject	Top 10 positive features	Top 10 negative features
Social sciences	learning, ethnographic, empirical, law, economics, legal, interviews, esrc, social, policy	religious identities, social economic, historians, book, organisational, digital economy, products services, privacy, styles, task
Superconductivity, magnetism and quantum fluids	spintronic, superfluid, quantum, magnetic, excitations, spintronics, superconductors, condensed, electrons, condensed matter	quantum computation, ultracold, optical lattice, ferromagnet, coherent transport, entanglement quantum, liquid nitrogen, ferromagnetism, control electron, colossal magnetoresistance
Terrestrial and freshwater environments	fertiliser, soils, aquatic, rhizosphere, catchment, nerc, rivers, river, weathering, soil	marine, ocean, respiration, agricultural landscapes, chalk, catalyst grant, carbon stocks, water resource, land carbon, accumulation

Table 7: Top 10 positive and negative features for each subject