Neil Byers
INLS 625 – Information Analytics
November 24, 2019

Semester Project: Report

Introduction

This project seeks to use text mining techniques to analyze the differences in research output between NASA scientists and Chinese scientists between 2011 and 2019. In 2011, the US Congress banned NASA from working bilaterally with any scientists associated with the Chinese government or Chinese corporations except when specifically authorized to do so.[1] Such a ban, some members of Congress argued, would help preserve American national security interests.[2] I felt curious about the practical effects such a ban might have had on the research output of NASA as compared to Chinese scientists with whom they had been all but prohibited from collaborating. If any salient differences in the research output of these two groups resulted from the ban, they would be worthwhile to identify.

The purpose of conducting this analysis is twofold. First, it is in the interest of members of Congress to be aware of the implications of their ban. How has it affected the research output of NASA scientists? Are there areas in which American interests might be furthered through greater Chinese collaboration? Or are we succeeding in staying ahead of a competitor? Secondly, such an analysis will be of use to the scientists themselves. Oftentimes scientists become so focused on their specific area of research that it becomes difficult to ascertain how that research fits into field-wide trends. What overarching trends define the research output of NASA scientists? How are those of Chinese scientists different? Can NASA scientists identify fertile areas for exploration through such methods?

I do not claim to provide answers to these questions, but I hope to conduct an exploratory study that might serve as a starting point for better understanding them. Specifically, this project seeks to address the following research question: What differences in the research focuses of NASA scientists and Chinese scientists can we identify using text mining and bibliometric analysis? This inquiry splits broadly into levels. First, I implement text mining techniques to discern whether or not salient differences exist between papers associated with NASA scientists as opposed to those associated with Chinese scientists. Secondly, I use feature correlation and multiple correspondence analysis to get a better idea of where such differences lie.

Data: Sources and Format

All of the data used over the course of this project were downloaded from the Web of Science bibliometric database. The data consist of three sets gathered using the following search techniques:

1. Web of Science research area category: “Astronomy and Astrophysics”

---

[1] https://www.govinfo.gov/content/pkg/PLAW-112publ55/html/PLAW-112publ55.htm
[2] https://web.archive.org/web/20130915190451/http://culberson.house.gov/bolden-in-beijing/

2. Year Range: 2011-2019
3. Subsets:
    a. All records with authors that have affiliations with Chinese institutions, excluding those with authors that have an affiliation with NASA (21192 instances)
    b. All records with authors that have an affiliation with NASA, excluding those with authors that have affiliations with Chinese institutions (13236 instances)
    c. All records that have authors listed as being affiliated with NASA *and* authors listed as having Chinese affiliations (823 instances)

Each of these three datasets were downloaded in .csv format for use with WEKA and LightSIDE as well as in Bibtex format for use with the R Bibliometrix package. The attributes in each file include standard bibliographic information such as title, authors, funding institution(s), journal, publication date, conference associations, etc. I used three separate datasets because Web of Science's search tools make it far easier to split groups of articles between NASA and Chinese scientists than would be the case were I to try and split them myself.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	PT	AU	BA	BE	GP	AF	BF	CA	TI	SO	SE	BS	LA	DT	CT	CY	CL	SP	HO	DE	ID	AB	C1
2	J	Wang, A; Ling, ZC; Yan, YC; McEwen, AS; Mellon				Wang, Alian; Ling, Zongcheng; Yan,			Subsurface	ICARUS			English	Article						Mars; Hydr	AQUEOUS-S	We report I	[Wang
3	J	Bar, RE; Trakhtenbrot, B; Oh, K; Koss, MJ; Wong				Bar, Rudolf E.; Trakhtenbrot, Benny			BAT AGN Sp	MONTHLY NOTICES OF THE ROYAL A			English	Article						galaxies: ac	ACTIVE GAL	We present	[Bar, F
4	J	Dharmawardena, TE; Kemper, F; Srinivasan, S; S				Dharmawardena, Thavisha E.; Kem			The nearby	MONTHLY NOTICES OF THE ROYAL A			English	Article						stars: AGB and post-AG		We present	[Dhar
5	J	Millan, RM; von Steiger, R; Ariel, M; Bartalev, S;				Millan, Robyn M.; von Steiger, Rud			Small satell	ADVANCES IN SPACE RESEARCH			English	Review						Small satell	MISSION; E	This is a CO	[Milla
6	J	Bernardini, E; de Martino, D; Mukai, K; Falanga				Bernardini, E.; de Martino, D.; Muk			2PBC J0658	MONTHLY NOTICES OF THE ROYAL A			English	Article						novae, cata	CATACLYSM	The hardX-	[Berna
7	J	Anderson, GE; Miller-Jones, JCA; Middleton, M				Anderson, G. E.; Miller-Jones, J. C. A			Discovery c	MONTHLY NOTICES OF THE ROYAL A			English	Article						stars: black	MASS BLAC	We report t	[Ande
8	J	Troja, E; Castro-Tirado, AJ; Gonzalez, JB; Hu, Y;				Troja, E.; Castro-Tirado, A. J.; Becer			The aftergl	MONTHLY NOTICES OF THE ROYAL A			English	Article						gravitation	GAMMA-RA	GRB 16082	[Troja
9	J	Zhou, G; Huang, CX; Bakos, GA; Hartman, JD; La				Zhou, G.; Huang, C. X.; Bakos, G. A.;			Two New H	ASTRONOMICAL JOURNAL			English	Article						planetary s	GIANT PLAN	Wide field	[Zhou
10	J	Abbott, BP; Abbott, R; Abbott, TD; Abraham, S;				Abbott, B. P.; Abbott, F		LIGO Sci Co	Search for i	PHYSICAL REVIEW D			English	Article							STARS; CLU	Gravitation	[Abbo
11	J	MacMillan, DS; Fey, A; Gipson, JM; Gordon, D;				MacMillan, D. S.; Fey, A.; Gipson, J.			Galactocen	ASTRONOMY & ASTROPHYSICS			English	Article						astrometry	SPIRAL STR	Aims. The I	[MacM
12	J	Galbany, L; Ashall, C; Hoflich, P; Gonzalez-Gaita				Galbany, L.; Ashall, C.; Hoflich, P.; (			Evidence fo	ASTRONOMY & ASTROPHYSICS			English	Article						supernovae	POTSDAM M	Aims. We p	[Galba
13	J	Soam, A; Liu, T; Andersson, BG; Lee, CW; Liu, JH				Soam, Archana; Liu, Tie; Andersson			Magnetic F	ASTROPHYSICAL JOURNAL			English	Article						ISM: magne	STAR-FORM	We present	[Soam
14	J	Trakhtenbrot, B; Arcavi, I; MacLeod, CL; Ricci, C				Trakhtenbrot, Benny; Arcavi, Iair; M			1ES 1927+6	ASTROPHYSICAL JOURNAL			English	Article						galaxies: ac	TIDAL DISR	We study t	[Trakh
15	J	Cambianica, P; Cremonese, G; Naletto, G; Lucc				Cambianica, P.; Cremonese, G.; Nal			Quantitativ	ASTRONOMY & ASTROPHYSICS			English	Article						comets: ge	SIZE-FREQU	Aims. We p	[Camb
16	J	Fellers, C; Fornasier, S; Ferrari, S; Hasselmann, P				Fellers, C.; Fornasier, S.; Ferrari, S.;			Rosetta/OS	ASTRONOMY & ASTROPHYSICS			English	Article						space vehic	EXPOSED W	Context. Fr	[Feller
17	J	Fornasier, S; Feller, C; Hasselmann, PH; Barucci				Fornasier, S.; Feller, C.; Hasselmann			Surface evo	ASTRONOMY & ASTROPHYSICS			English	Article						comets: inc	EXPOSED W	Context. Th	[Forna
18	J	Fornasier, S; Hoang, VH; Hasselmann, PH; Feller				Fornasier, S.; Hoang, V. H.; Hasselm			Linking sur	ASTRONOMY & ASTROPHYSICS			English	Article						comets: inc	EXPOSED W	Aims. The R	[Forna
19	J	Hasselmann, PH; Barucci, MA; Fornasier, S; Boc				Hasselmann, P. H.; Barucci, M. A.; F			Pronounce	ASTRONOMY & ASTROPHYSICS			English	Article						comets: inc	WATER ICE;	A smooth d	[Hasse
20	J	Masoumzadeh, N; Kolokolova, L; Tubiana, C; El				Masoumzadeh, N.; Kolokolova, L.;			Phase-curv	ASTRONOMY & ASTROPHYSICS			English	Article						comets: ge	SIZE DISTRI	Aims. The R	[Maso
21	J	Tognon, G; Ferrari, S; Penasa, L; La Forgia, F; Ma				Tognon, G.; Ferrari, S.; Penasa, L.; L			Spectroph	ASTRONOMY & ASTROPHYSICS			English	Article						comets: ge	WATER ICE;	Context. Be	[Togn
22	J	Tubiana, C; Rinaldi, G; Guttler, C; Snodgrass, C;				Tubiana, C.; Rinaldi, G.; Guettler, C			Diurnal var	ASTRONOMY & ASTROPHYSICS			English	Article						comets: ge	WATER ICE;	Context. O	[Tubia
23	J	Soam, A; Lee, CW; Andersson, BG; Maheswar, G				Soam, Archana; Lee, Chang Won; A			First Sub-pa	ASTROPHYSICAL JOURNAL			English	Article							ISOLATED S	L1521F is f	[Soam
24	J	Abbott, BP; Abbott, R; Abbott, TD; Abraham, S;				Abbott, B. P.; Abbott, F		LIGO Sci Co	Binary Blac	ASTROPHYSICAL JOURNAL LETTERS			English	Article						black holes	COMPACT-(	We present	[Abbo
25	J	Abbott, BP; Abbott, R; Abbott, TD; Abraham, S;				Abbott, B. P.; Abbott, F		LIGO Sci Co	Directional	PHYSICAL REVIEW D			English	Article							RADIATION	We perform	[Abbo
26	J	Abbott, BP; Abbott, R; Abbott, TD; Abraham, S;				Abbott, B. P.; Abbott, F		LIGO Sci Co	Searches fo	ASTROPHYSICAL JOURNAL			English	Correction									[Abbo
27	J	Li, SS; Zang, W; Udalski, A; Shvartzvald, Y; Hube				Li, S. -S.; Zang, W.; Udalski, A.; Shva			OGLE-2017	MONTHLY NOTICES OF THE ROYAL A			English	Article						asteroseism	SURFACE BI	We present	[Li, S.
28	J	Long, F; Herczeg, GJ; Harsono, D; Pinilla, P; Taz				Long, Feng; Herczeg, Gregory J.; Ha			Compact D	ASTROPHYSICAL JOURNAL			English	Article						circumstell	PROTOPLAN	We present	[Long,
29	J	Han, C; Bennett, DP; Udalski, A; Gould, A; Bond				Han, Cheongho; Benne		KMTNet Co	OGLE-2018	ASTRONOMICAL JOURNAL			English	Article						gravitation	JUPITER/SA	We report a	[Han,
30	J	Han, C; Yee, JC; Udalski, A; Bond, IA; Bozza, V; C				Han, Cheongho; Yee, Je		KMTNet Co	Spectrosco	ASTRONOMICAL JOURNAL			English	Article						gravitation	GRAVITATIO	We report t	[Han,
31	J	Tafoya, D; Orosz, G; Vlemmings, WHT; Sahai, R;				Tafoya, D.; Orosz, G.; Vlemmings, W			Spatio-kine	ASTRONOMY & ASTROPHYSICS			English	Article						stars: AGB a	PREPLANET	Context. W	[Tafoy
32	J	Kriss, GA; De Rosa, G; Ely, J; Peterson, BM; Kaas				Kriss, G. A.; De Rosa, G.; Ely, J.; Pete			Space Teles	ASTROPHYSICAL JOURNAL			English	Article						galaxies: ac	BROAD-LIN	We model t	[Kriss,
33	J	Paliya, VS; Koss, M; Trakhtenbrot, B; Ricci, C; O				Paliya, Vaidehi S.; Koss, M.; Trakhte			BAT AGN Sp	ASTROPHYSICAL JOURNAL			English	Article						galaxies: ac	BLACK-HOL	The recentl	[Paliya

Raw Web of Science .csv

Data: Risks and Limitations

Using Web of Science data to approach the problem at hand substantially limits my ability to draw conclusions from my analysis. The data itself comes with a number of caveats that I had to keep in mind as I took lessons from my efforts:

1. These datasets only include papers published by academic sources, such as articles, books, and conference proceedings. As such, any classified activity by NASA scientists is left out. Perhaps more salient is the presumed omission of work by the Chinese National Space Administration. Because this organization is so closely tied to the Chinese military, it is doubtful that much (if not all) of official Chinese space research is not included in my data. Without data from the Chinese government, I am severely limited in my ability to analyze the effects a ban which prohibits NASA from working with said government.

2. My search was restricted using Web of Science's 'Astronomy & Astrophysics' research area category. This narrow area certainly does not account for all space-related research that might be undertaken by NASA or Chinese scientists. Related papers might be labeled under separate categories, though identifying all of these works would be impossible.
3. Because these data come from academic sources, work by the private sector is probably underrepresented by a significant degree. This is unfortunate because the private sector has made great strides in space research, especially in rockets and robotics, over the past decades. As with concern #1, however, I see no way to accommodate for this deficiency other than to acknowledge it.
4. I assume that many articles authored by Chinese authors were translated into English from Chinese, meaning that different linguistic patterns may surface to distinguish said articles from those written by NASA scientists in English. Such differences may limit my ability to discern whether or not differences found by algorithms result from linguistic or content-related differences. I plan to limit the effect of this phenomenon by analyzing the feature tables I produce to ensure that linguistic phenomena do not emerge as the most differentiating features in the set.
5. The two datasets are not mutually exclusive. Despite the congressional ban, a significant number of papers emerged that listed both NASA authors and Chines authors. This probably results from activities pursued under specific authorization from Congress as well as collaborations that will not affect national security interests. As a result, my classification task is no longer binary. I account for this consideration by making two copies of my feature representation, one with a 'NASA' T/F binary target variable and another with a 'CHINA' T/F binary target variable. I then run algorithms over both to compare how easily one can predict NASA articles with how easily one can predict CHINA articles. This accounts for articles labeled as having both NASA and China affiliations.

These limitations affect my ability to draw strong conclusions from my analysis. I cannot claim, for example, to have a complete view of the research output of scientists affected by the 2011 congressional ban. That being said, my analysis still provides an oblique perspective on the issue at hand. By looking at the academic output of University-affiliated Chinese scientists as opposed to official, government-related research, we can still form an image of the country's priorities.

Data: Cleaning and Preparation

To prepare for my analysis, I had to perform a number of operations. First, I assigned "CHINA," "NASA," or "BOTH" labels to all instances in the 3 respective data files. I then eliminated unwanted columns and was left with the following 4 attributes:

- Title
- Author Keywords
- Web of Science assigned keywords
- Abstracts

I felt that these 4 columns would be most informative for ascertaining the research interests and activities of the two groups of scientists. I then eliminated all instances that did not at least have a title (312 instances). I felt that the other 3 columns were helpful, but not absolutely necessary. Finally, I compiled all of my instances into a single .csv file and assigned labels relative to two separate target variables: ‘CHINA’ (T/F) and ‘NASA’ (T/F). By doing so I could treat my problem as two separate binary classification tasks and account for articles associated with *both* NASA and China.

	A	B	C	D	E	F	G	H
1	TargetVar	CHINA	NASA	TI	DE	ID	AB	
2	CHINA	TRUE	FALSE	Geological cl	Chang'e-4/Cl	POLE-AITKEN	On January 3, 2019, 's Cha	
3	CHINA	TRUE	FALSE	Laboratory sy	Hydrated Al-:	SOLID-SOLUT	Orbital remote sensing ha	
4	CHINA	TRUE	FALSE	Impact crate	Impact crate	INNER SOLA	Impact craters are the pre	
5	CHINA	TRUE	FALSE	The Itokawa	Asteroid rego	X-RAY-DIFFR	Asteroid regolith simulant	
6	CHINA	TRUE	FALSE	Is atmospher	Organisation	SPECTRAL EN	In order to quantify the de	
7	CHINA	TRUE	FALSE	Designing ob	X-ray pulsar based navigat		The accuracy in pulsar-bas	
8	CHINA	TRUE	FALSE	The capabilit	Tianlai radio array; Space		The bistatic radar system	
9	CHINA	TRUE	FALSE	Smart-RTK: I	Android smart devices; Kir		Global Navigation Satellite	
10	CHINA	TRUE	FALSE	Dynamics m	Space debris	CAPTURE	Tether-net is a new active	
11	CHINA	TRUE	FALSE	Quality asses	Timing group	DIFFERENTIA	The quality of broadcast g	
12	CHINA	TRUE	FALSE	The applicati	Coastline inf	WATER INDE	The coastline is the dividir	
13	CHINA	TRUE	FALSE	Invariance of	Coring drill; I	PREDICTION	In this paper, we study the	
14	CHINA	TRUE	FALSE	A solar elect	Solar electro	PROTON; FLI	A new solar electron even	
15	CHINA	TRUE	FALSE	V1082-Sgr: A	stars: evoluti	X-RAY SOUR	V1082 Sgr is a cataclysmic	
16	CHINA	TRUE	FALSE	Gaia paralla:	parallaxes; g	RR LYRAE; CA	We have established a mi:	
17	CHINA	TRUE	FALSE	Investigation	shock waves,	CURRENT SH	On 2017 September 10, a	

Cleaned .csv with binary classes added

Having given shape to my raw data, the next step was to transform said data into text features that would be digestible by machine learning algorithms. To do so I used Carnegie Melon University’s LightSIDE tool.[3] Using this tool I was able to transform my data into a feature array with the following elements:

- Unigrams – This made for the most simple feature representation possible. The simplicity helped to limit the processing requirements involved and also would help limit overfitting.
- No stopwords – This is a standard step for text analysis, as these words do little to identify class associations.
- Stemming – I wanted to normalize my unigrams for broader comparisons across instances, to limit overfitting, and again to cut down on processing requirements.

[3] http://www.cs.cmu.edu/~cprose/LightSIDE.html

- Remove Punctuation – Because academic abstracts follow subdued stylistic norms across many fields, I assumed that punctuation would not help to distinguish class associations.
- Track Feature Hit Location – this is a default setting in LightSIDE that keeps track of where words occur in their respective text fields.

Having formed my feature array, I sought to eliminate features that would make it too 'easy' for the algorithms to associate instances with either NASA or China. I felt that leaving the following features in my set would limit my ability to associate a model's success with the *research focus* of articles as opposed to non-content-related words like 'Nasa' and 'China.' In other words, identifying an article as a "NASA" article because the algorithm finds the word "Nasa" in the abstract tells me little about the research focus of that article. Removed features included:

- nasa
- nasa/goddard
- goddard
- nasa/gsfc
- nasa/ipac
- esa/nasa
- china
- chine
- china-vo

In its usable form my data consisted of two feature arrays, one each for the CHINA and NASA target variables, each with a shape of 35251x20506.

Methodology: Models

In order to get an idea of whether or not discernable differences existed between the 'NASA' articles and the 'CHINA' articles, I built classification models using four different algorithms. I then tested each model (with the exception of one) using 10-fold cross validation. I noted the accuracy and Kappa scores for each model and compared them to one another. The four algorithms I used were Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, and Random Tree.[4] I chose these four because they were the only algorithms I could get to finish given my computational resources. The first three I ran using LightSIDE's built-in functionality while the fourth I ran using WEKA.

I elected to use Accuracy and Kappa score as evaluation statistics for a number of reasons. For one, LightSIDE only displays these statistics (as opposed to F-Score, Mean Absolute Error, Precision, Recall etc.). I tried to get these statistics through WEKA, but my large dataset made doing so impossible with my existing processing power. My lack of precision/recall statistics for three of my models proved to be an issue, for it might have been useful to analyze false positives and false negatives for each model. Having these numbers could have potentially helped me to determine if one category was mistaken for the other across various models. This

[4] I was unable to evaluate the latter model using cross validation because the memory restrictions of WEKA/Java caused the cross validation to fail mid-execution. I thus elected to evaluate the Random Tree model using a single 90/10 random split.

proved to be one more example of how prohibitive working with large datasets and different toolsets can be.

My results for the four models were as follows. For each algorithm I built two models, one each for the 'NASA' and 'CHINA' target variable sets.

Logistic Regression: CHINA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8496
Kappa	0.6775

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	10393	2843
TRUE	2458	19557

Logistic Regression: NASA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8597
Kappa	0.7065

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	18860	2332
TRUE	2613	11446

SVM: CHINA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8276
Kappa	0.6317

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	10130	3106
TRUE	2970	19045

SVM: NASA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8374
Kappa	0.6601

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	18415	2777
TRUE	2956	11103

Naïve Bayes: CHINA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7998
Kappa	0.5855

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	10699	2537
TRUE	4521	17494

Naïve Bayes: NASA

Model Evaluation Metrics:

Metric	Value
Accuracy	0.806
Kappa	0.6042

Model Confusion Matrix:

Act \ Pred	FALSE	TRUE
FALSE	16837	4355
TRUE	2485	11574

Random Tree: CHINA

```
=== Summary ===

Correctly Classified Instances        2309               65.5035 %
Incorrectly Classified Instances      1216               34.4965 %
Kappa statistic                          0.2517
Mean absolute error                      0.345
Root mean squared error                  0.5873
Relative absolute error                 73.5405 %
Root relative squared error            121.2626 %
Total Number of Instances             3525

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
                 0.497     0.25       0.545     0.497     0.52
                 0.75      0.503      0.712     0.75      0.731
Weighted Avg.    0.655     0.408      0.65      0.655     0.652

=== Confusion Matrix ===

    a    b   <-- classified as
  659  666 |    a = FALSE
  550 1650 |    b = TRUE
```

Random Tree: NASA

```
=== Summary ===

Correctly Classified Instances        2294               65.078  %
Incorrectly Classified Instances      1231               34.922  %
Kappa statistic                          0.2617
Mean absolute error                      0.3492
Root mean squared error                  0.5909
Relative absolute error                 72.8941 %
Root relative squared error            120.8141 %
Total Number of Instances             3525

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
                 0.733     0.474      0.702     0.733     0.717
                 0.526     0.267      0.564     0.526     0.544
Weighted Avg.    0.651     0.392      0.647     0.651     0.649

=== Confusion Matrix ===

    a    b   <-- classified as
 1559  569 |    a = FALSE
  662  735 |    b = TRUE
```

Model Comparison

	Logistic Regression	Support Vector Machine	Naive Bayes	Random Tree
China Accuracy, Kappa	**.8496, .6775**	.8276, .6317	.7998, .5855	.6550, .2517
NASA Accuracy, Kappa	**.8597, .7065**	.8374, .6601	.8060, .6042	.6508, .2617

The purpose of this step was to determine whether or not demonstrable differences existed between NASA research and Chinese research. All models with the exception of the Random Tree provided accuracy ratings of ~80% or higher, with Kappa scores of ~.55 or higher. These are promising results because they demonstrate that the models can classify articles as belonging to NASA or Chinese scientists more reliably than would be the case should we randomly assign them to either class. It is perhaps notable that all algorithms were slightly better at predicting NASA articles than they were at predicting whether or not articles belonged to the 'CHINA' class. This could indicate that NASA articles possess slightly more definitive characteristics than their Chinese counterparts (perhaps because so many of the NASA articles relate to NASA-specific space missions and hardware). The Random Tree method is less robust than the other three, which has been demonstrated by the above numbers. Furthermore, because we used cross-validation to evaluate the first three models, the ratings listed above are probably deflated compared to what they might achieve with an external test set.

Methodology: MCA and Feature Correlation

Because I found that models could distinguish between CHINA and NASA articles with a relatively high degree of accuracy, I used two separate methods to attempt to discern what made them distinguishable. The first of which was simply to use LightSIDE to pick out the most positively correlative text features for each class. Presumably this would illuminate which words and concepts were most closely identified with either the NASA or CHINA class. From this I hoped to discern which class topics or ideas aligned with which class.

Top 20 Most Positively Correlative Features by Class

CHINA		NASA	
Feature	Correlation	Feature	Correlation
decai	0.16286251	planet	0.24499843
b.v.	0.16029457	instrument	0.23592511
quark	0.1427423	mission	0.21630271
qcd	0.13550626	planetari	0.18930026
meson	0.12345233	exoplanet	0.18821654
lhc	0.11863433	imag	0.18579798
bar	0.11591002	,Äôs	0.18415568
theori	0.11460165	infrar	0.18217835
scalar	0.11356941	inc.	0.18217292
symmetri	0.11251452	scienc	0.18143423
hadron	0.11229218	detect	0.1792801
boson	0.11180779	telescop	0.17318933
fb	0.11129605	spectroscopi	0.16633784
author	0.11029131	atmospher	0.16571945
collis	0.10840115	these	0.16181325
equat	0.10427032	character	0.15884352
root	0.10391884	similar	0.15857727
calcul	0.10154606	observ	0.15529454
pi	0.10079075	present	0.15079553
quantum	0.10018012	dust	0.15060263

These features seem to indicate some different areas of focus for each body of research. Judging by this information, the Chinese articles seem to be more focused on fundamental and particle physics (ex. "boson" and "quark") while the NASA articles seemed to be more focused on observational astronomy and planetary science ("exoplanet" and "mission").

In order to investigate these differences from a different perspective, I decided to turn to the Bibliometrix package in R. Following the examples laid out in Darvish 2018,[5] I ran Multiple Correspondence Analysis (MCA) over the Bibtex files for each group. Doing so allowed me to easily produce concept maps for each body of research which could then be compared and contrasted. I ran this analysis over the NASA-only or China-only (excluding the 'BOTH' set) sets because I wanted to identify differences between the two exclusive sets.

Commands entered in RStudio:

Load package:

```
>Library(Bibliometrix)
```

Load data file:

```
>ChinaData <- readFiles(Users/neilbyers..../CHINAnotNASA.bib)
```

Convert to data frame:

```
>ChinaFrame <- convert2df(ChinaData, dbsource = "wos", format = "bibtex")
```

Run MCA analysis:

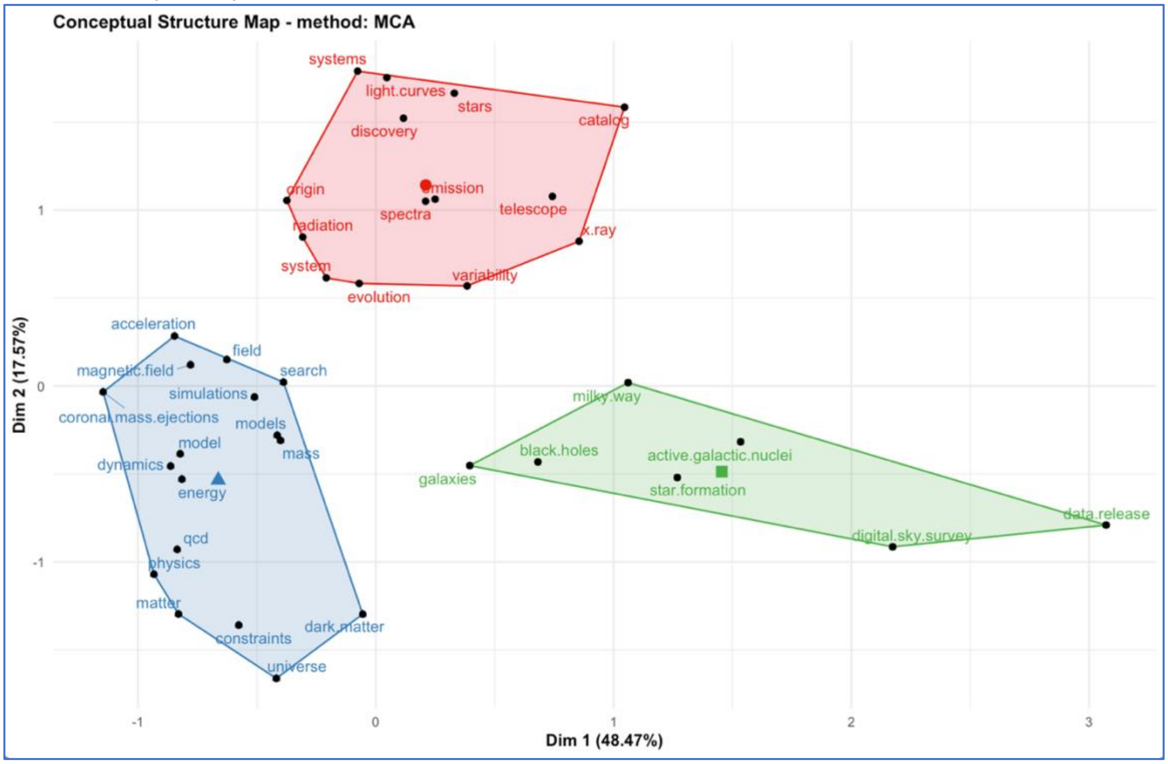
```
> ChinaConceptual <- conceptualStructure(ChinaFrame, field="ID", method="MCA",
stemming=FALSE, minDegree=250, clust=3, k.max=5)
```

The above command generates three clusters, which is the default. I used 250 as the minimum degree for the China set and 175 for the NASA set to compensate for the different sample sizes of each (~21000 articles vs ~13000 articles). As such, a term required fewer hits to show up in the NASA plots than was required for a term to show up in the China plots. The points are plotted using Euclidean distance and are clustered according to the number of clusters set in the parameters (3).

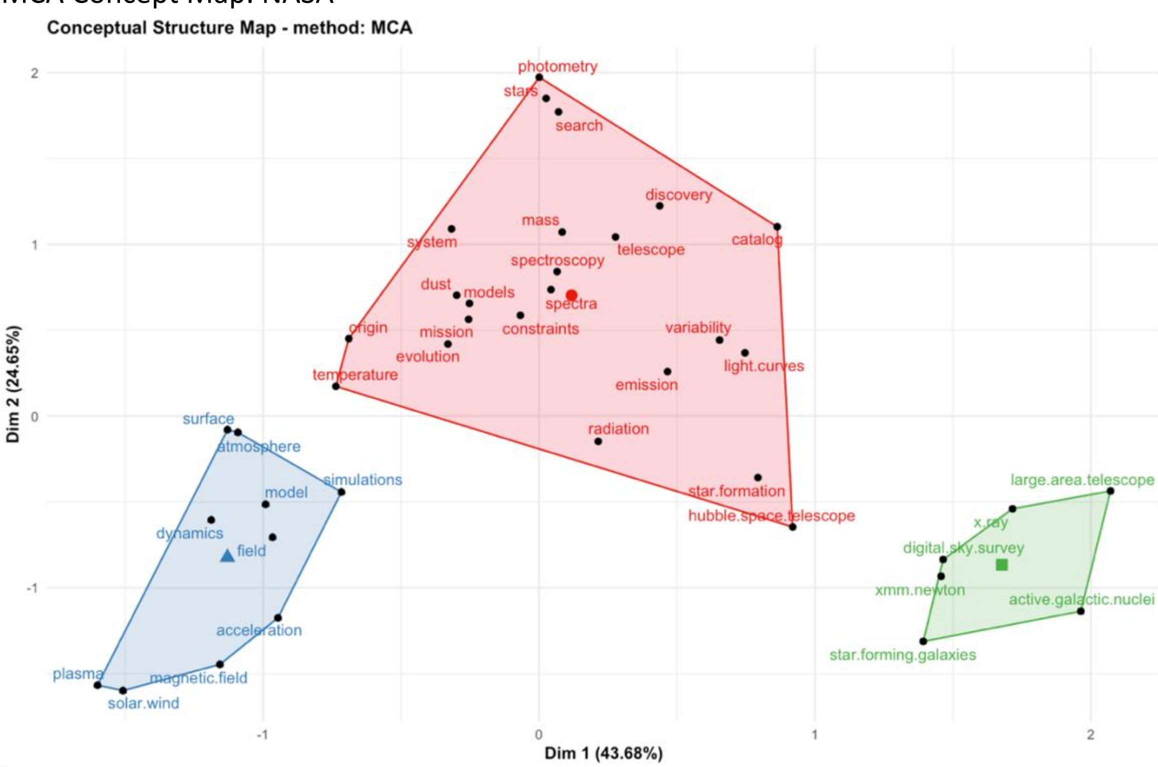
---

[5] Darvish, H. (2018). "Bibliometric analysis using Bibliometrix an R Package." Retrieved From https://www.researchgate.net/publication/329058973_Bibliometric_analysis_using_Bibliometrix_an_R_Package

MCA Concept Map: CHINA



MCA Concept Map: NASA



These plots demonstrate similar divisions to those indicated by the correlated features listed above. Interestingly, the green and red clusters in both charts seem to align with one another. The green clusters both contain the ‘active.galactic.nuclei’ and ‘digital.sky.survey’ terms while the red clusters seem to center around telescope observation and spectroscopy. Notably, however, the red (telescope) cluster in the NASA map seems to be more extensive and diverse than the red cluster in the China map. This aligns with the differences seen in the correlative features. Though the blue cluster seems to indicate a focus on particle and fundamental physics in the China map (“physics,” “mass,” “dynamics”), the blue cluster in the NASA map does not follow any easily discernible themes. Again, this aligns with the differences seen in the correlative features.

In general, both of these analyses seem to support the idea that the papers in the CHINA set are more focused on theory and fundamentals while the papers in the NASA set are more concerned with astronomical observation and planetary science missions.

Conclusions

In order to gauge the effects of the US Congress’s 2011 law forbidding NASA from cooperating with Chinese scientists, I hoped to use bibliometric data to determine whether or not demonstrable differences could be found and identified between the research output of these two groups. My hope was that by finding and identifying such differences, I could gain some idea as to the tangible effects of the 2011 ban on the work of these two groups of scientists.

The analysis outlined above demonstrated major content differences between the sets of articles I used. Several machine learning algorithms were able to classify articles as either belonging to the “NASA” set or the “CHINA” set with high degrees of accuracy. Doing so seemed to indicate that discernable differences existed, so I compared positively correlative features and MCA concept maps between the two sets to attempt to identify the roots of those differences. Ultimately, these efforts seem to indicate a mission-based observational and exploratory focus in the NASA set and a more theoretical, physics-based focus in the China set.

Unfortunately, it is impossible to use my results to make any conclusions about the 2011 congressional ban. This is due to several factors:

1. The Web of Science data set only includes articles, conference proceedings, and other academic formats. While NASA scientists routinely publish articles, the Chinese National Space Administration does not even show up as an organization in Web of Science’s database. This is presumably because the Chinese government hopes to keep the research of its scientists “in house.” The effect of this is to essentially render my comparison as one between apples and oranges. My datasets include research primarily from the American space agency and not from independent scientists at American universities or other institutions. On the Chinese side, they *only* represent such independent or university-based scientists and expressly *exclude* the Chinese space agency. As such, I cannot make any conclusions about how effective this ban has been in limiting Chinese adoption of American methods, nor can I make the conclusion that NASA scientists are losing out by not cooperating with their Chinese counterparts
2. My dataset was built using Web of Science’s ‘Astronomy & Astrophysics’ research area category, meaning that only articles labeled as such by the database would be included.

This surely leaves out many articles that touch on space-research and thus should have been included in my analysis.

When combined these two limitations indicate that the differences I observed resulted more from my data collection than any real-world phenomena. Without the activity of China's space agency and all of its mission-related activities, 'Astronomy & Astrophysics' as a research area would probably include little else besides fundamental & theoretical physics. This is exactly what I found on the Chinese side of my data. Unfortunately, I did not have the resources to build a more complete or representative dataset.

Despite these limitations, this project served several functions. For one, it taught me much about the selection and preparation of bibliometric data for use with machine learning algorithms. Furthermore, it may provide a road map for those hoping to perform a more conclusive analysis should more representative data become available. Finally, though my results were limited in their usefulness for comparison between NASA and Chinese scientists, both groups would benefit from the analysis of the output for the group to which they belong. As mentioned above, scientists can become very narrowly focused as a result of specialization and specific research focus. It is always instructive and useful to be able to see one's work as a part of a whole, a goal for which this type of project is well-suited. Using Web of Science data in this way seems to be a much more productive endeavor than using it to compare two competing space administrations.