

Using Machine Learning to Predict Citation Metrics

Bercin Cenik & Tina Huang

Qualifications for a job in academic science



- Funding status
- Commitment to research
- Commitment to teaching
- Educational background

What faculty hiring committees want

Charles B Wright & Nathan L Vanderford

PhD trainees aspiring to become faculty need to know the credentials search committees value most in an applicant.

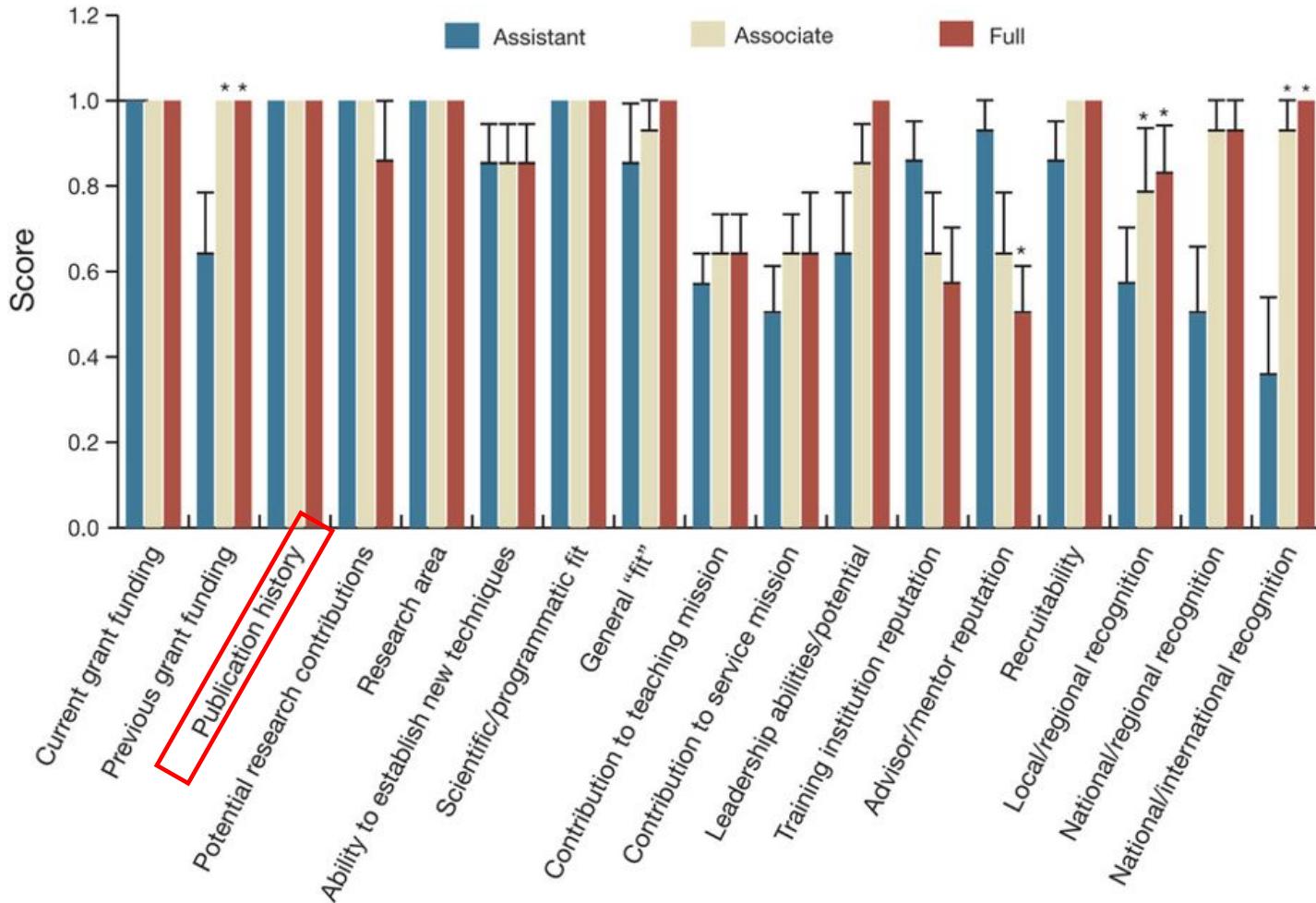


Table 1 Faculty credentials fall into four categories

	Assistant	Associate	Full
Core competencies			
Current grant funding	1.00	1.00	1.00
Publication history	1.00	1.00	1.00
Research area	1.00	1.00	1.00
Scientific/programmatic fit	1.00	1.00	1.00
Potential research contributions	1.00	1.00	0.86
Ability to establish new techniques	0.86	0.86	0.86
General fit	0.86	0.93	1.00
Recruitability	0.86	1.00	1.00
Initial necessities			
Advisor/mentor reputation	0.93	0.64	0.50
Training institution reputation	0.86	0.64	0.57
Necessities for advancement			
National/international recognition	0.36	0.93	1.00
Regional/national recognition	0.50	0.93	0.93
Previous grant funding	0.64	1.00	1.00
Leadership abilities/potential	0.64	0.86	1.00
Unnecessary credentials			
Contribution to service mission	0.50	0.64	0.64
Contribution to teaching mission	0.57	0.64	0.64
Local/regional recognition	0.57	0.79	0.83

Credential scores followed one of four distinct trends: core competencies, initial necessities, necessities for advancement, and unnecessary credentials.

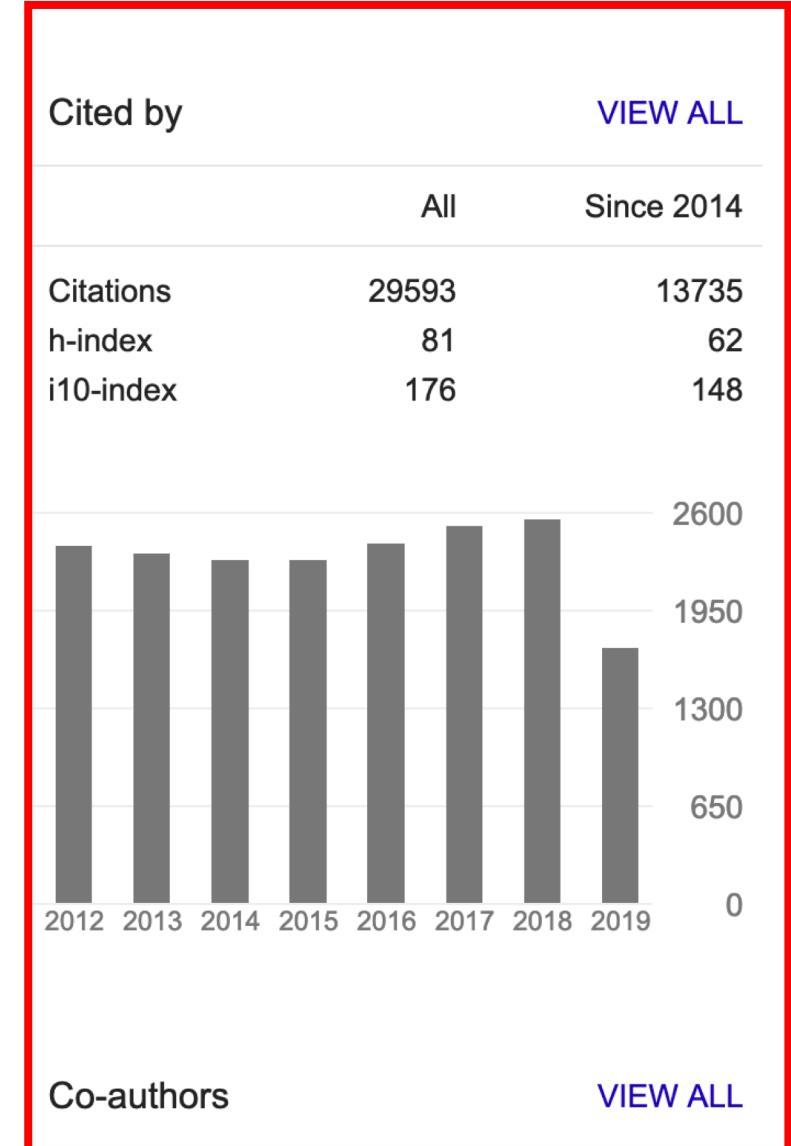
Scientific citation is providing detailed reference in a scientific publication, typically a [paper](#) or book, to previous published (or occasionally private) communications which have a bearing on the subject of the new publication. The purpose of citations in original work is to allow readers of the paper to refer to cited work to assist them in judging the new work, source background information vital for future development, and acknowledge the contributions of earlier workers. Citations in, say, a review paper bring together many sources, often recent, in one place.

To a considerable extent the quality of work, in the absence of other criteria, is judged on the number of citations received, adjusting for the volume of work in the relevant topic.^[citation needed] While this is not necessarily a reliable measure, counting citations is trivially easy; judging the merit of complex work can be very difficult.^[citation needed]

Citation impact quantifies the citation usage of scholarly works
Citation counts for an author, an individual article, or a publication

Reference section in scientific paper.

Websites that generate citation metrics



How many times will my article be cited?

...if it is published in <insert favorite journal here>



Project Goals

“So, how many times will my paper be cited?”

- Acquire citation metrics from web of science
 - 1997 – 2017
 - Nature Reviews Genetics
 - Title, authors, publication year, citation count (total and individual years)
- Visualization of citation metrics
- Use machine learning to predict citation counts for individual articles
 - Title of the article
 - Publication year
 - Name of senior author

[Search](#)[Tools ▾](#) [Searches and alerts ▾](#) [Search History](#) [Marked List](#)

Results: 3,121
(from Web of Science Core Collection)

You searched for: PUBLICATION NAME: (nature reviews genetics)
[...More](#)

[Create Alert](#)

Refine Results

 Search within results for... 

Filter results by:

-  Highly Cited in Field (158)
-  Open Access (373)
-  Associated Data (3)

[Refine](#)

Publication Years

- 2017 (118)
- 2016 (144)
- 2015 (127)

Sort by: [Date ↴](#) Times Cited Usage Count Relevance More ▾

◀ 1 of 313 ▶

Select Page  Export... [Add to Marked List](#)

1. [Principles of gene regulation across tissues](#)

By: Burgess, Darren J.
NATURE REVIEWS GENETICS Volume: 18 Issue: 12 Pages: 701-701 Published: DEC 2017

[Find it @ NU](#) [Full Text from Publisher](#)

 [Analyze Results](#)

 [Create Citation Report](#)

Times Cited: 1
(from Web of Science Core Collection)

Usage Count ▾

2. [Demographic history, selection and functional diversity of the canine genome](#)

By: Ostrander, Elaine A.; Wayne, Robert K.; Freedman, Adam H.; et al.
NATURE REVIEWS GENETICS Volume: 18 Issue: 12 Pages: 705-720 Published: DEC 2017

[Find it @ NU](#) [Full Text from Publisher](#) [View Abstract ▾](#)

Times Cited: 19
(from Web of Science Core Collection)

Usage Count ▾

3. [Population genetics of sexual conflict in the genomic era](#)

By: Mank, Judith E.
NATURE REVIEWS GENETICS Volume: 18 Issue: 12 Pages: 721-730 Published: DEC 2017

[Find it @ NU](#) [Full Text from Publisher](#) [View Abstract ▾](#)

Times Cited: 22
(from Web of Science Core Collection)

Usage Count ▾

4. [Convergence between biological, behavioural and genetic determinants of obesity](#)

By: Ghosh, Sujoy; Bouchard, Claude
NATURE REVIEWS GENETICS Volume: 18 Issue: 12 Pages: 731-748 Published: DEC 2017

[Find it @ NU](#) [Full Text from Publisher](#) [View Abstract ▾](#)

Times Cited: 18
(from Web of Science Core Collection)

Usage Count ▾

Results can be downloaded as an .xls file
For our analyses, this was reformatted as a csv and cleaned up with pandas
Visualizations were generated with matplotlib

Summary (and fun facts)

- 3121 articles published between 1997 and 2017
- The highest cited paper was cited >5000 times
- On average, a NRG paper is cited 2 times
- Half of the dataset had no citations



Goal

To predict whether a paper has above- or below-median total citation count based on its:

Title
Authors

```
count      3121.000000
mean       115.244473
std        284.517875
min        0.000000
25%       0.000000
50%       2.000000
75%      130.000000
max      5424.000000
Name: Total Citations, dtype: float64
```

	Title	Authors	Publication Year	Total Citations	Average per Year	1997 1998 1999 2000 2001 ... 2008 2009 2010 2011 2012										
						1997	1998	1999	2000	2001	...	2008	2009	2010	2011	2012
0	RNA-Seq: a revolutionary tool for transcriptomics	Wang, Zhong; Gerstein, Mark; Snyder, Michael	2009	5424	493.09	0	0	0	0	0	...	0	69	223	341	469
1	Network biology: Understanding the cell's func...	Barabasi, AL; Oltvai, ZN	2004	4202	262.63	0	0	0	0	0	...	248	326	285	316	359
2	The fundamental role of epigenetic events in c...	Jones, PA; Baylin, SB	2002	3855	214.17	0	0	0	0	0	...	262	263	284	292	270
3	APPLICATIONS OF NEXT-GENERATION SEQUENCING Seq...	Metzker, Michael L.	2010	3395	339.50	0	0	0	0	0	...	0	0	148	343	444
4	Micrornas: Small RNAs with a big role in gene ...	He, L; Hannon, GJ	2004	3343	208.94	0	0	0	0	0	...	119	144	177	206	265

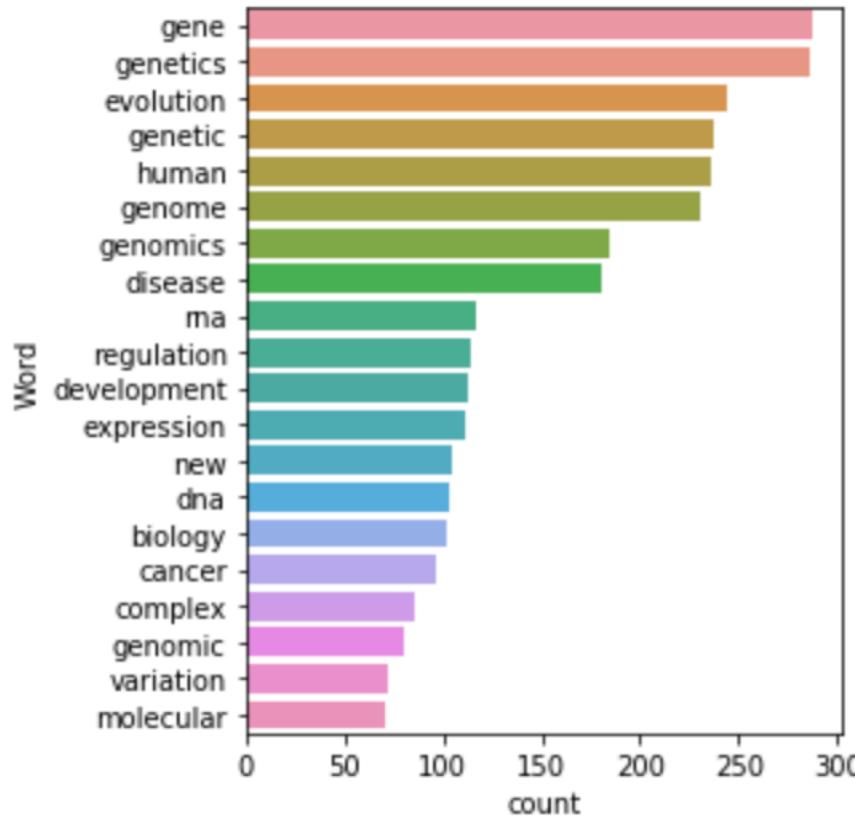
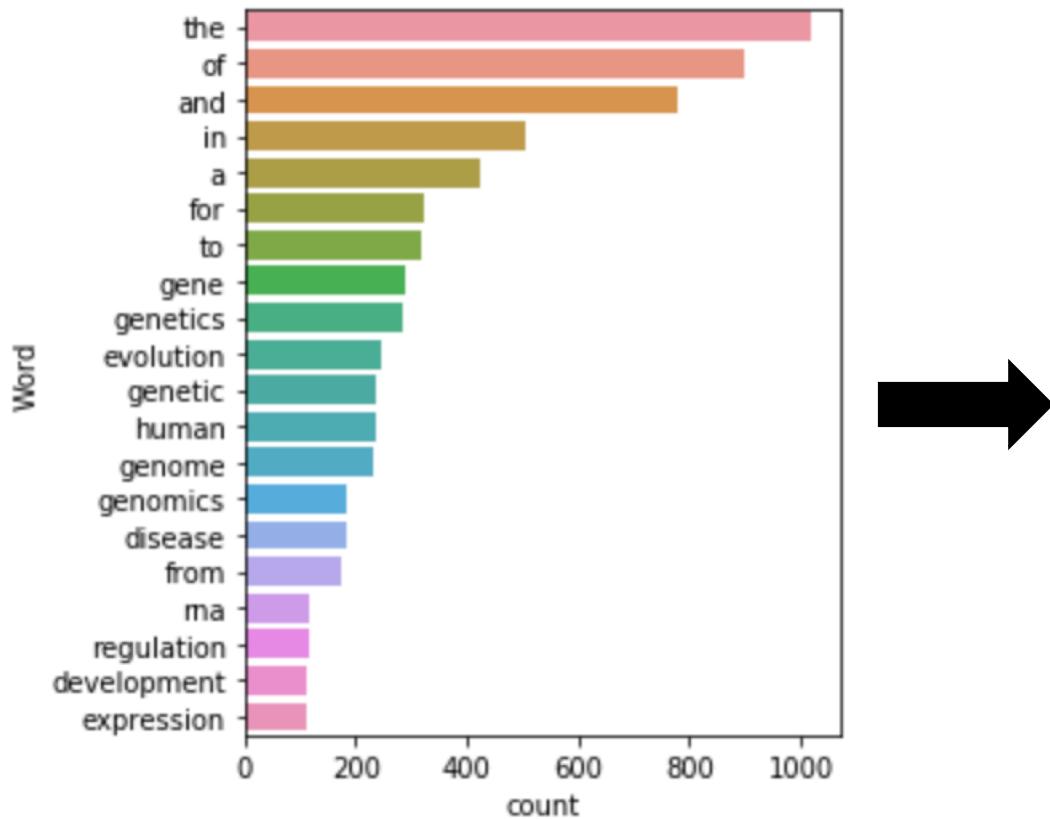
	finalcleaned_title	Average per Year	Total Citations	Category
0	rna seq revolutionary tool transcriptomics	493.09	5424	Top_half
1	network biology understanding cells functional...	262.63	4202	Top_half
2	fundamental role epigenetic events cancer	214.17	3855	Top_half
3	applications next generation sequencing sequen...	339.50	3395	Top_half
4	micrornas small rnas big role gene regulation	208.94	3343	Top_half

Tokenization

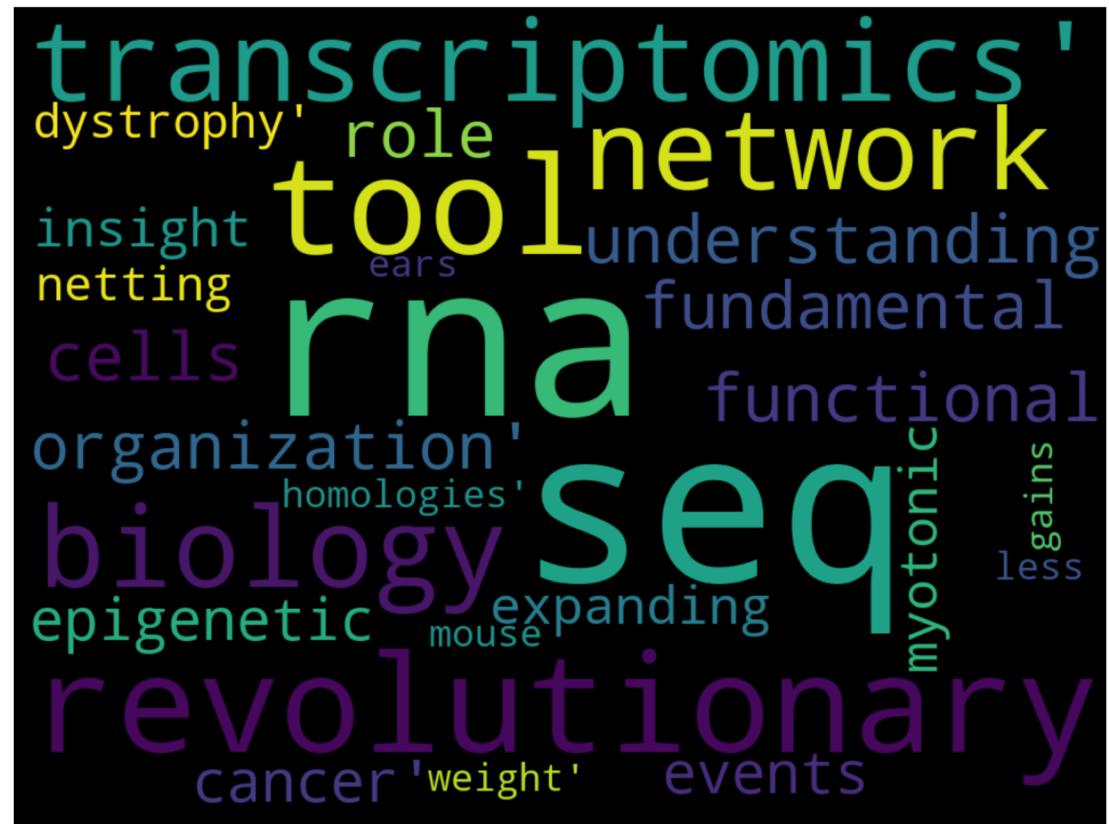


Title	cleaned_title	Authors	cleaned_authors
RNA-Seq: a revolutionary tool for transcriptomics	rna seq a revolutionary tool for transcriptomics	Wang, Zhong; Gerstein, Mark; Snyder, Michael	wang zhong gerstein mark snyder michael
Network biology: Understanding the cell's func...	network biology understanding the cells functi...	Barabasi, AL; Oltvai, ZN	barabasi al oltvai zn
The fundamental role of epigenetic events in c...	the fundamental role of epigenetic events in c...	Jones, PA; Baylin, SB	jones pa baylin sb
APPLICATIONS OF NEXT-GENERATION SEQUENCING Seq...	applications of next generation sequencing seq...	Metzker, Michael L.	metzker michael l
Micrornas: Small RNAs with a big role in gene ...	micrornas small rnas with a big role in gene r...	He, L; Hannon, GJ	he l hannon gj

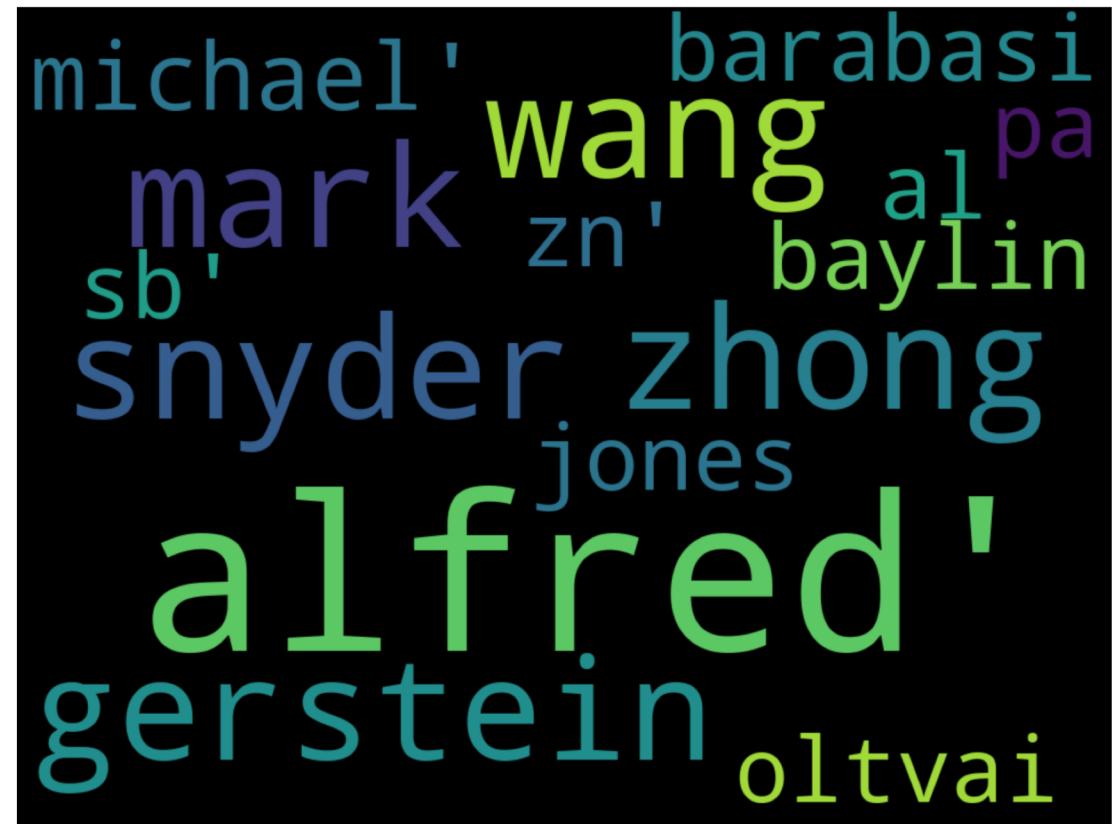
Stopwords filtering



Most frequent words



Title



Author

TF-IDF vs. Count Vectorizer

- TF-IDF: weights down common words, gives more importance to rare words
- Count Vectorizer: count the number of words

```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2 # initialise the vectoriser
3 tvec = TfidfVectorizer()
4 # fit the training data on the model
5 tvec.fit(X_train)
6
7 #transform training data into sparse matrix
8 X_train_tvec = tvec.transform(X_train)
9
10 # cross val score/ predict
11 tvec_score = cross_val_score(lr, X_train_tvec, y_train, cv=3)
```

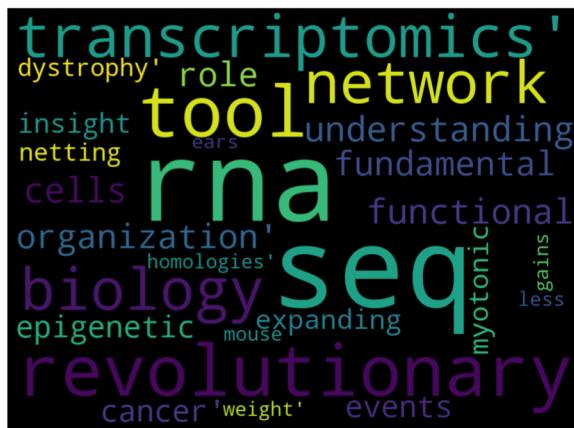
```
1 from sklearn.feature_extraction.text import CountVectorizer
2 # initialise the vectoriser
3 cvec = CountVectorizer()
4 # fit the training data on the model
5 cvec.fit(X_train)
6
7 #transform training data into sparse matrix
8 X_train_cvec = cvec.transform(X_train)
9
10 # cross val score/ predict
11 cvec_score = cross_val_score(lr, X_train_cvec, y_train, cv=3 )
12 X_train_cvec
13
```

ML: Predict citation based on title using logistic regression

```
TF-IDF Vectorizer Score: 0.714198009803197  
Count Vectorizer Score: 0.7234898027045001
```

params	scores
0 Count	0.723490
1 TF-IDF	0.714198

TF-IDF is used due to frequent appearance of words such as “RNA” and “Seq” in a genomics journal



Logistic Regression with TF-IDF

```
1 tfidf_vectorizer = TfidfVectorizer(max_df=0.8, max_features=1000)  
2 xtrain, xval, ytrain, yval = train_test_split(traindf['finalcleaned_title'], traindf['Categ  
3  
4 # create TF-IDF features  
5 xtrain_tfidf = tfidf_vectorizer.fit_transform(xtrain)  
6 xval_tfidf = tfidf_vectorizer.transform(xval)  
7  
8 from sklearn.linear_model import LogisticRegression  
9 classifier = LogisticRegression()  
10 classifier.fit(xtrain_tfidf, ytrain)  
11 print(f"Training Data Score: {classifier.score(xtrain_tfidf, ytrain)}")  
12 print(f"Testing Data Score: {classifier.score(xval_tfidf, yval)}")  
  
Training Data Score: 0.8441506410256411  
Testing Data Score: 0.7472
```

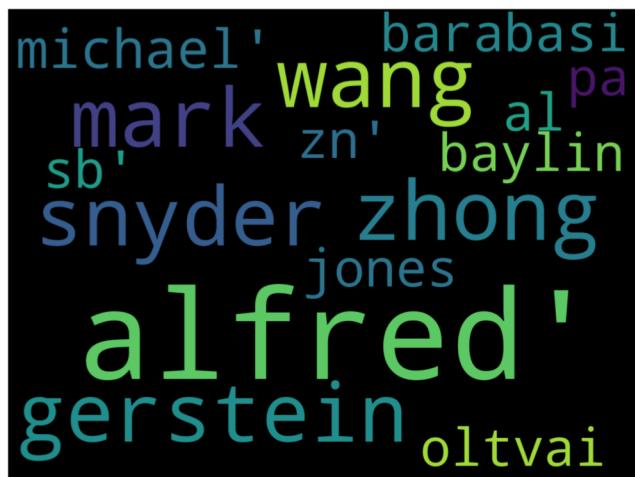
	Predictions	Actual
0	Top_half	Top_half
1	Top_half	Top_half
2	Bottom_half	Bottom_half
3	Top_half	Bottom_half
4	Bottom_half	Bottom_half
5	Bottom_half	Bottom_half
6	Bottom_half	Bottom_half
7	Top_half	Top_half
8	Top_half	Top_half
9	Top_half	Top_half

ML: Predict citation based on **author** using logistic regression

```
TF-IDF Vectorizer Score: 0.7767109411561862
Count Vectorizer Score: 0.8167510283504519
```

params	scores
0 Count	0.816751
1 TF-IDF	0.776711

- Count vectorizer is used because we don't want to discard anyone's name...



Logistic Regression with Count Vectorizer

```
1 from sklearn.feature_extraction.text import CountVectorizer
2
3 xtrain, xval, ytrain, yval = train_test_split(traindf['finalcleaned_authors'], traindf['Cat']
4
5 # initialise the vectoriser
6 cvec = CountVectorizer()
7
8 # fit the training data on the model, transform training data into sparse matrix
9 xtrain_cvec = cvec.fit_transform(xtrain)
10
11 from sklearn.linear_model import LogisticRegression
12 classifier = LogisticRegression()
13 classifier.fit(xtrain_cvec, ytrain)
14
15
16 cvec_score = cross_val_score(lr, xtrain_cvec, ytrain, cv=10)
17 #print(cvec_score)
18 df_cvec = pd.DataFrame(xtrain_cvec.toarray(), columns=cvec.get_feature_names())
19 print(df_cvec.head())
20
21 cvec_score = cross_val_score(lr, X_train_cvec, y_train, cv=10)
22 print('Score:', tvec_score.mean())
23
```

	aa	aaron	aarron	abbasi	abd	abdallah	abecasis	abel	abiola	abrahams	\
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0

	...	zollner	zon	zondervan	zoran	zou	zoya	zuben	zuber	zuccala	\
0	...	0	0	0	0	0	0	0	0	0	0
1	...	0	0	0	0	0	0	0	0	0	0
2	...	0	0	0	0	0	0	0	0	0	0
3	...	0	0	0	0	0	0	0	0	0	0
4	...	0	0	0	0	0	0	0	0	0	0

	zuniga
0	0
1	0
2	0
3	0
4	0

[5 rows x 3420 columns]

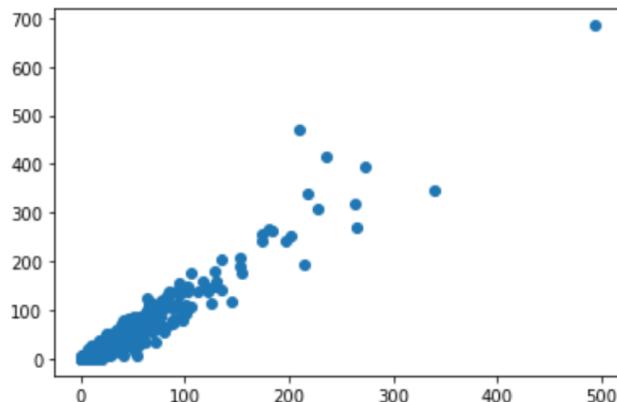
Score: 0.7767109411561862

ML: Predict total citation based on previous trends using linear regression

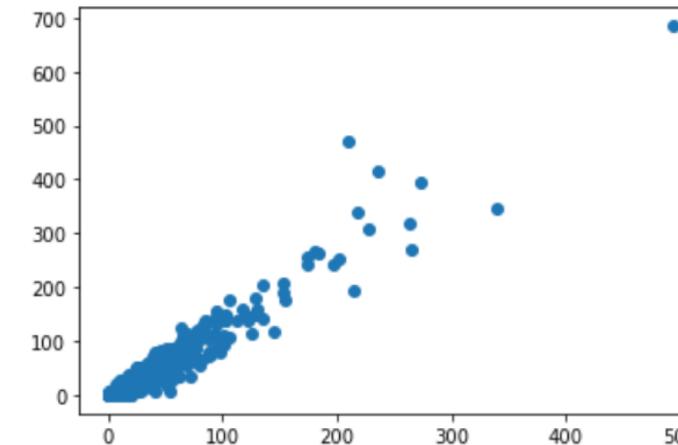
Average citation per year and last year's citations

```
1 peryear = naturegenetics["Average per Year"]
2 lastyear = naturegenetics["2017"]
3 plt.scatter(peryear, lastyear)
```

<matplotlib.collections.PathCollection at 0x11a77dd68>



Reshape
→



```
1 # Assign the data to X and y
2 # Note: Sklearn requires a two-dimensional array of values
3 # so we use reshape to create this
4
5 X = peryear.values.reshape(-1, 1)
6 y = lastyear.values.reshape(-1, 1)
7
8 print("Shape: ", X.shape, y.shape)
```

Shape: (3121, 1) (3121, 1)

ML: Predict total citation based on previous trends using linear regression

```
1 # Use sklearn's `train_test_split` to split the data into training and testing
2
3 from sklearn.model_selection import train_test_split
4
5 # YOUR CODE HERE
6 X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
7
```

```
1 # Create the model
2
3 # YOUR CODE HERE
4 from sklearn.datasets import make_regression
5 from sklearn.linear_model import LinearRegression
6 model = LinearRegression()
7
```

```
1 # Fit the model to the training data.
2
3 # YOUR CODE HERE
4 model.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
1 # Calculate the mean_squared_error and the r-squared value
2 # for the testing data
3
4 from sklearn.metrics import mean_squared_error, r2_score
5
6 # YOUR CODE HERE
7 # Use our model to predict a value
8 predicted = model.predict(X_test)
9
10 # Score the prediction with mse and r2
11 mse = mean_squared_error(y_test, predicted)
12 r2 = r2_score(y_test, predicted)
13
14 print(f"Mean Squared Error (MSE): {mse}")
15 print(f"R-squared (R2 ): {r2}")
```

```
Mean Squared Error (MSE): 66.3033064342903
R-squared (R2 ): 0.9544251088797852
```

```
1 # Call the `score` method on the model to show the r2 score
2
3 # YOUR CODE HERE
4 model.score(X_test, y_test)
```

```
0.954425108879785
```

```
1 totalyear = naturegenetics["Average per Year"]
2 lastyear2 = naturegenetics["2017"]
3 plt.scatter(totalyear, lastyear2)
```

```
<matplotlib.collections.PathCollection at 0x11ee43da0>
```

