ROOT DIRECTORY

SINGLE-SOURCE TRUTH FOR WIKIPEDIA CATEGORIES

SIMPLE QUESTIONS
HARD ANSWERS

How many articles on Wikipedia are related to mathematics?

31,444 Portal estimate

23,051 Wikidata estimate

34,425 PetScan estimate:

WAY TOO MANY CATEGORY TOOLS

1. ‘Portal’ pages
2. ‘Outline’ pages
3. ‘Indices’ pages
4. ‘Areas of’ pages
5. ‘Contents’ pages
6. ‘Category’ pages
7. ‘Glossaries’ pages
8. ‘Overviews’ pages
9. ‘Contents/Lists’ page
10. ‘Category: Lists’ page
11. ‘Lists of lists of lists’ page

ENTER MACHINE LEARNING

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	omplexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org /wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression		0.92	0.74	0.2

OBJECTIVE:

TRUTH TABLE

TITLE	Wiki Article	Simple Wiki Article	Probability Score	Complexity Score	Quality Score
Logistic Regression	https://en.wikipedia.org/wiki/Logistic_regression	https://simple.wikipedia.org/wiki/Logistic_Regression	0.92	0.74	0.2

MODELING

TF-IDF

MULTINOMIAL NAIVE BAYES

LOGISTIC REGRESSION

FEATURE IMPORTANCE

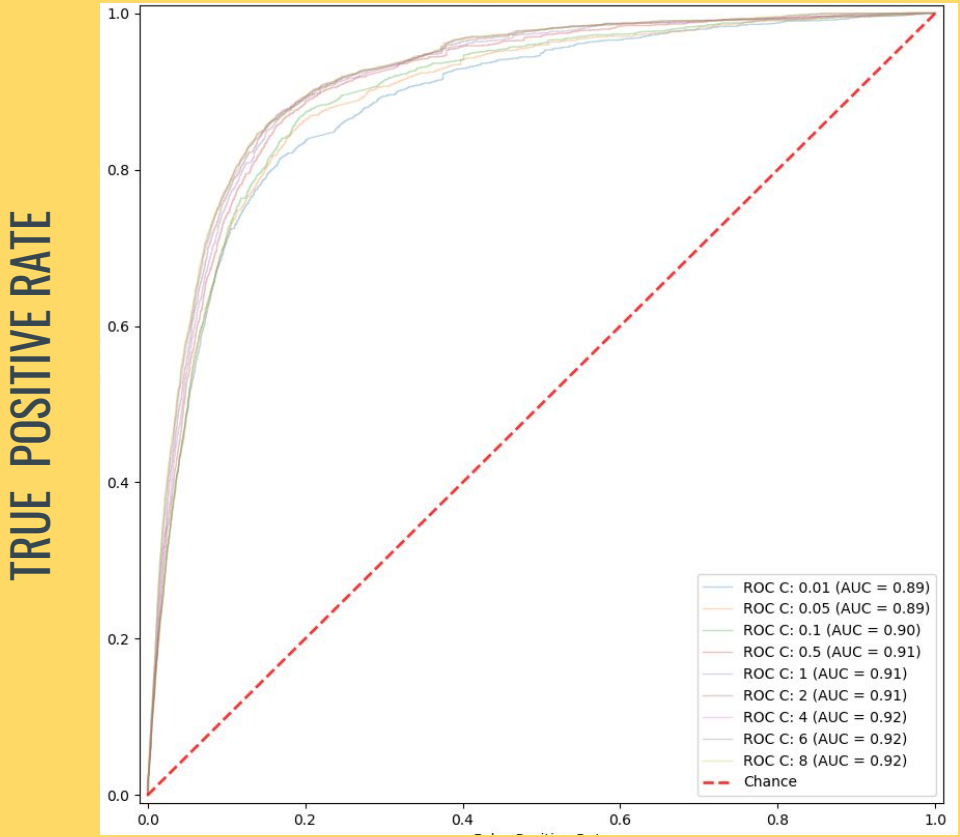
VERY MATHY WORDS

MATHEMATICIAN
THEOREM
NUMBER
ALGORITHM

VERY NOT MATHY WORDS

ENGINEER
PHYSICS
ART
CHEMISTRY

ROC AUC


TRUE POSITIVE RATE
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC C: 0.01 (AUC = 0.89)
ROC C: 0.05 (AUC = 0.89)
ROC C: 0.1 (AUC = 0.90)
ROC C: 0.5 (AUC = 0.91)
ROC C: 1 (AUC = 0.91)
ROC C: 2 (AUC = 0.91)
ROC C: 4 (AUC = 0.92)
ROC C: 6 (AUC = 0.92)
ROC C: 8 (AUC = 0.92)
Chance
FALSE POSITIVE RATE

CONFUSION MATRIX @ 0.20

FALSE POSITIVES

	title	predicted	actual	threshold_pred
2760	Criss-cross algorithm	0.952502	0.0	True
2761	Vector logic	0.953342	0.0	True
2762	Category:Biological theorems	0.954540	0.0	True
2763	Chaitin's constant	0.955203	0.0	True
2764	Computational complexity	0.958522	0.0	True
2765	P versus NP problem	0.959869	0.0	True
2766	Non-constructive algorithm existence proofs	0.960831	0.0	True
2767	Generic-case complexity	0.962550	0.0	True
2768	Time complexity	0.962793	0.0	True
2769	Computing the permanent	0.964047	0.0	True
2770	Integer factorization	0.965000	0.0	True
2771	Computable number	0.971522	0.0	True
2772	Theory of computation	0.971899	0.0	True
2773	Discrete logarithm	0.972355	0.0	True
2774	Linear programming	0.972618	0.0	True
2775	Existential theory of the reals	0.978830	0.0	True
2776	Computable function	0.980849	0.0	True
2777	Computational complexity theory	0.982427	0.0	True
2778	Numerical analysis	0.983520	0.0	True

FALSE POSITIVES

	titl	predicted	actual	reshold_pred
2760	Criss-cross algorithr	0.952502	0.0	True
2761	Vector logi	0.953342	0.0	True
2762	Category:Biological theorem	0.954540	0.0	True
2763	Chaitin's constar	0.955203	0.0	True
2764	Computational complexit	0.958522	0.0	True
2765	P versus NP probler	0.959869	0.0	True
2766	Non-constructive algorithm existence proof	0.960831	0.0	True
2767	Generic-case complexit	0.962550	0.0	True
2768	Time complexit	0.962793	0.0	True
2769	Computing the permaner	0.964047	0.0	True
2770	Integer factorizatio	0.965000	0.0	True
2771	Computable numbe	0.971522	0.0	True
2772	Theory of computatio	0.971899	0.0	True
2773	Discrete logarithr	0.972355	0.0	True
2774	Linear programmin	0.972618	0.0	True
2775	Existential theory of the real	0.978830	0.0	True
2776	Computable functio	0.980849	0.0	True
2777	Computational complexity theor	0.982427	0.0	True
2778	Numerical analysi	0.983520	0.0	True

SO, HOW MANY MATH ARTICLES ARE ON WIKIPEDIA?

JAKUB SVEC

www.github.com/jakubsvec001

www.linkedin.com/jakubsvec001

jakubsvec001@gmail.com