
Wikipedia

— Tuned Predictions on Big Data —

Scott Dobbins (presenting)
Rachel Kogan (project partner)

Natural Language Processing

Term Frequency Inverse Document Frequency (TF-IDF)

Times <word> appears in THIS article

Number of articles <word> appears in

WIKIPEDIA

The Free Encyclopedia

English

5 419 000+ articles

中文

945 000+ 條目

Español

1 338 000+ artículos

Deutsch

2 068 000+ Artikel

Français

1 876 000+ articles

Português

970 000+ artigos



日本語

1 063 000+ 記事

Русский

1 398 000+ статей

Italiano

1 361 000+ voci

Polski

1 226 000+ haseł

EN ▾🔍

🌐 Read Wikipedia in your language ▾

Need for bots

17.5 million pages

9.5 million non-redirect pages

6.5 million article pages

5.3 million non-list article pages

8,500 articles tagged with POV issues

1,200 articles tagged with weasel words


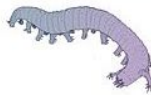

65,000 articles tagged for (non-)notability

Use

By dolphins

In 1997, use of sponges as a **tool** was c
when searching for food in the sandy se
known case of tool use in **marine mamn**
get a life losers

(List of Invertebrates)

Arachnids		102,248
Velvet worms		165
Horseshoe crabs		4
Paul Ryan		1
Others jellyfish, echinoderms, sponges, other worms etc.		68,658
Totals:		1,305,075

Facts and Opinions



More difficult examples:

- *Abortion is wrong* – **opinion, not a fact.**
- *The pro-life movement holds that abortion is wrong, or occasionally that it is only justified in certain special cases* – **fact, not an opinion.**

Bias/Neutrality

Britney Spears: Live and More!

From Wikipedia, the free encyclopedia



The **neutrality** of this article is **disputed**. Relevant discussion may be found on the [talk page](#). Please do not remove this message until [conditions to do so are met](#). *(December 2012)* *(Learn how and when to remove this template message)*

Consequences of Nazism

From Wikipedia, the free encyclopedia



This article has multiple issues. Please help [improve it](#) or discuss these issues on the [talk page](#). *(Learn how and when to [hide] remove these template messages)*

- The **neutrality** of this article is **disputed**. *(October 2012)*
- This article **needs additional citations for verification**. *(June 2007)*

Agenda

- Natural Language Processing
- Wikipedia
- Getting the Data
- Handling the Data
- Building Models
- Model Results
- Conclusions & Future Directions

Wikipedia

- Wants you to have its data
 - Full site and full site history are both readily downloadable
 - Edit history >> 1TB! (~64TB?)

enwiki-latest-pages-articles4.xml-p200511p35268...	02-Jun-2017	13:27	255158912
enwiki-latest-pages-articles4.xml-p200511p35268...	03-Jun-2017	13:11	829
enwiki-latest-pages-articles5.xml-p352690p56531...	02-Jun-2017	13:30	279718988
enwiki-latest-pages-articles5.xml-p352690p56531...	03-Jun-2017	13:11	829
enwiki-latest-pages-articles6.xml-p565314p89291...	02-Jun-2017	13:33	302640068
enwiki-latest-pages-articles6.xml-p565314p89291...	03-Jun-2017	13:11	829
enwiki-latest-pages-articles7.xml-p892914p12686...	02-Jun-2017	13:35	310507780
enwiki-latest-pages-articles7.xml-p892914p12686...	03-Jun-2017	13:11	832
enwiki-latest-pages-articles8.xml-p1268693p1791...	02-Jun-2017	13:37	340879834
enwiki-latest-pages-articles8.xml-p1268693p1791...	03-Jun-2017	13:11	835
enwiki-latest-pages-articles9.xml-p1791081p2336...	02-Jun-2017	13:38	330875269
enwiki-latest-pages-articles9.xml-p1791081p2336...	03-Jun-2017	13:11	835
enwiki-latest-pages-logging.xml.gz	03-Jun-2017	15:48	3033242204
enwiki-latest-pages-logging.xml.gz-rss.xml	03-Jun-2017	15:48	775
enwiki-latest-pages-meta-current.xml.bz2	03-Jun-2017	05:59	27343796753
enwiki-latest-pages-meta-current.xml.bz2-rss.xml	03-Jun-2017	13:30	793
enwiki-latest-pages-meta-current1.xml-p10p30303...	02-Jun-2017	21:47	183935635
enwiki-latest-pages-meta-current1.xml-p10p30303...	03-Jun-2017	13:19	826
enwiki-latest-pages-meta-current10.xml-p2336425...	02-Jun-2017	22:10	528230272
enwiki-latest-pages-meta-current10.xml-p2336425...	03-Jun-2017	13:20	850
enwiki-latest-pages-meta-current11.xml-p3046514...	02-Jun-2017	22:16	610390189
enwiki-latest-pages-meta-current11.xml-p3046514...	03-Jun-2017	13:20	850
enwiki-latest-pages-meta-current12.xml-p3926863...	02-Jun-2017	22:23	679410026
enwiki-latest-pages-meta-current12.xml-p3926863...	03-Jun-2017	13:21	850
enwiki-latest-pages-meta-current13.xml-p5040438...	02-Jun-2017	22:24	665792735
enwiki-latest-pages-meta-current13.xml-p5040438...	03-Jun-2017	13:21	850
enwiki-latest-pages-meta-current14.xml-p6197598...	02-Jun-2017	22:35	747165528
enwiki-latest-pages-meta-current14.xml-p6197598...	03-Jun-2017	13:21	850
enwiki-latest-pages-meta-current14.xml-p6197598...	02-Jun-2017	21:49	23533810
enwiki-latest-pages-meta-current14.xml-p6197598...	03-Jun-2017	13:21	850
enwiki-latest-pages-meta-current15.xml-p7744803...	02-Jun-2017	22:34	721089938
enwiki-latest-pages-meta-current15.xml-p7744803...	03-Jun-2017	13:21	850
enwiki-latest-pages-meta-current15.xml-p9244803...	02-Jun-2017	21:57	125685930
enwiki-latest-pages-meta-current15.xml-p9244803...	03-Jun-2017	13:22	850
enwiki-latest-pages-meta-current16.xml-p1101805...	02-Jun-2017	22:05	230088051
enwiki-latest-pages-meta-current16.xml-p1101805...	03-Jun-2017	13:22	856
enwiki-latest-pages-meta-current16.xml-p9518050...	02-Jun-2017	22:34	652694425
enwiki-latest-pages-meta-current16.xml-p9518050...	03-Jun-2017	13:22	853
enwiki-latest-pages-meta-current17.xml-p1153926...	02-Jun-2017	22:35	682306034
enwiki-latest-pages-meta-current17.xml-p1153926...	03-Jun-2017	13:22	856

AWS & Linux

Data sufficiently large that EC2 clusters were needed to unzip it piece by piece

Storage in S3

Analysis in Databricks (EC2)



pySpark

- A lot of filtering:
 - Filter out redirect pages
 - Filter out list pages
 - Filter for specific tag of interest (POV)
- Separate train and test early to avoid leakage
- Use parquet to transport data from pySpark to Scala Spark
- Use pandas and pickles to get data into normal Python

Regex

- Separate data by XML page tag
- Clean XML to plain bag of words

```
((\{\}\{?![\\w]{0,8}Summary)([\\n\\r\\t\\ | ~  
-z]*)(\{\}\{?![\\w]{0,8}Summary)([\\n\\r\\t\\  
| ~ -z]*)(\\}\}))([\\n\\r\\t\\ | ~ -z]*))*([\\}\}))
```

```
In [38]: POV_clean.loc[1,'text_clean']  
Out[38]: u' ayahuasca cooking loreto region ayahuasca iowaska entheogenic brew made  
banisteriopsis caapi vine ingredients brew used traditional spiritual medicine ceremonies  
among indigenous peoples amazon basin can mixed leaves psychotria viridis chacruna leaf  
chagropanga mimosa tenuiflora rootbark dimethyltryptamine dmt containing plant species  
reported psychoactive effects can felt consuming ayahuasca vine alone hallucinogen dmt will  
digested stomach remain inactive without inclusion monoamine oxidase inhibitor maoli  
banisteriopsis caapi therefore combination maoli-containing plant dmt-containing substance  
necessary full hallucinogenic effects resulting brew known number different names see  
ayahuasca known many names throughout northern south america brazil ayahuasca hispanicized  
style spelling word quechua languages spoken andean states ecuador bolivia peru colombia  
speakers quechua languages aymara language may prefer spelling ayawaska word refers liana  
banisteriopsis caapi brew prepared quechua languages aya means spirit soul corpse dead body  
waska means rope woody vine liana word ayahuasca variously translated liana soul liana dead  
spirit liana brazil brew liana informally called either caapi latter portuguese word liana  
woody climbing vine vegetal brazil organised spiritual tradition people drink ayahuasca brew  
prepared exclusively caapi viridis adherents vegetal call brew hoasca vegetal achuar people  
shuar people ecuador peru call natem whereas sharanahua peoples peru call shori century  
christian missionaries spain portugal first encountered indigenous south americans using  
ayahuasca earliest reports described work devil century active chemical constituent caapi  
named telepathine found identical chemical already isolated peganum harmala given name  
harmaline beat writer william burroughs read paper richard evans schultes subject traveling  
south america early sought hopes relieve cure opiate addiction see yage letters ayahuasca  
became widely known mckenna brothers published experience amazon true hallucinations dennis  
mckenna later studied pharmacology botany chemistry ayahuasca oo-koo-he became subject master  
thesis richard evans schultes allowed claudio naranjo make special journey canoe amazon river  
study yage south american indians brought back samples drug published first scientific  
description effects active alkaloids brazil number modern religious movements based use  
ayahuasca emerged famous santo daime vegetal udv usually animistic context may shamanistic  
often santo daime udv integrated christianity santo daime vegetal now members churches  
throughout world similarly europe started see new religious groups develop relation increased  
ayahuasca use westerners teamed shamans amazon rainforest regions forming ayahuasca healing  
retreats claim able cure mental physical illness allow communication spirit world recent years  
brew popularized wade davis one river english novelist martin goodman carlos castaneda chilean  
novelist isabel allende writer kira satak author jeremy narby cosmic serpent author jay  
griffiths wild elemental journey radio personality robin quivers also key plot season episode  
spellbound list law order special victims unit sections banisteriopsis caapi vine macerated
```

Stop words

about | above | after | again | against | all | and | any | are | arent | beca
use | been | before | being | below | between | both | but | cant | canno
t | could | couldnt | did | didnt | does | doesnt | doing | dont | down | du
ring | each | few | for | from | further | had | hadnt | has | hasnt | have |
havent | having | hed | hell | hes | her | here | heres | hers | herself | hi
m | himself | his | how | hows | ill | ive | into | isnt | its | its | itself | lets | m
ore | most | mustnt | myself | nor | not | off | once | only | other | ought
| our | ours | ourselves | out | over | own | same | shant | she | shed | sh
ell | shes | should | shouldnt | so | some | such | than | that | thats | the
| their | theirs | them | themselves | then | there | theres | these | they
| theyd | theyll | theyre | theyve | this | those | through | too | under | u
ntil | very | was | wasnt | wed | well | were | weve | were | werent | what
| whats | when | whens | where | wheres | which | while | who | whos |
whom | why | whys | with | wont | would | wouldnt | you | youd | youll |
youre | youve | your | yours | yourself | yourselves

Lemmatization

```
((\b(un|in|anti|il|dis|mis))|([w]+(ish|ishly|ely|ly|ally|al|ous|i  
ous|able|ible|ables|ibles|ably|ibly|ment|ments|eness|eness  
es|ness|nesses|ation|ations|ional|iful|fulness|ful|ion|ions)\b  
)|((((?<=[b])[b]?)|(?<=[d])[d]?)|(?<=[f])[f]?)|(?<=[g])[g]?)|(?<=[l])[l]  
?)|(?<=[m])[m]?)|(?<=[n])[n]?)|(?<=[p])[p]?)|(?<=[r])[r]?)|(?<=[s]  
)[s]?)|(?<=[t])[t]?)|(?<=[v])[v]?)|(?<=[z])[z]?)|(?<=[bcdfghjklmnpr  
stvwxyz]))(ied|ed|er|ers|ies|es|en|ing|ings)\b)|((?<=t)(re|res)\b  
)|(?<=[bcdfghjklmnprstvwxyz])(e)\b)|(?<=[bcdgklmnpqrtvw])(s)\b  
|(?<=[bcdfghjklmnprstvwxyz])(ied|ed|ator|ators|ies|es|en|ing  
|ate|y|ze|se|ren)\b))
```

Bag of Words

```
In [36]: POV_clean.loc[0,'text_clean']
```

```
Out[36]: u' ammonius saccas century greek philosopher alexandria often referred one founders  
neoplatonism mainly known teacher plotinus taught eleven years undoubtedly biggest influence  
plotinus development neoplatonism although little known philosophical views later christian  
writers stated ammonius christian now generally assumed different ammonius alexandria wrote  
biblical texts much known life ammonius saccas cognomen sakkas interpreted indicate porter  
youth seems misreading sakkas sakkophoros porter grammatically incorrect however erich seeberg  
argued cognomen refers india ruling clan gautama buddha also belonged related iranian saka  
scythians indo-scythians known antiquity cognomen sakkas therefore referred india marker  
ethnic identity according interpretation supported fact ammius marcellinus refers saccas  
ammonius thus sacian ammonius makes reading denoting sakkos impossible interpretation name  
subsequently contested corroborate porphyry report plotinus ammonius foremost student acquired  
high esteem indian philosophy eager desire travel india ammonius interpretation saccas denotes  
ethnic northern indian origin rather alluding gautama buddha supports possibility ammonius may  
raised christian reverted paganism reported eusebius drawing porphyry contra christianos case  
ammonius may second-generation indian remained contact philosophy ancestral country intensity  
commerce goods ideas alexandria india makes wholly possible option link india however  
consistent plotinus passion india also helps explain often noted substantial agreements shared  
ideas vedanta neoplatonism increasingly attributed direct indian influence details life come  
fragments left porphyry writings famous pupil ammonius saccas plotinus studied ammonius eleven  
years according porphyry age plotinus went alexandria study philosophy blockquote twenty-  
eighth year plotinus felt impulse study philosophy recommended teachers alexandria highest  
reputation came away lectures depressed full sadness told trouble one friends friend  
understanding desire heart sent ammonius far tried went heard said friend man looking day  
stayed continually ammonius acquired complete training philosophy became eager make  
acquaintance persian philosophical discipline prevailing among indians blockquote according  
porphyry parents ammonius christians upon learning greek philosophy ammonius rejected parents  
religion paganism conversion contested christian writers jerome eusebius state ammonius  
remained christian throughout lifetime blockquote porphyry plainly utters falsehood will  
opposer christians says ammonius fell life piety heathen customs ammonius held divine  
philosophy unshaken unadulterated end life works yet extant show celebrated among many  
writings left blockquote however told longinus ammonius wrote nothing ammonius principal  
influence plotinus unlikely ammonius christian one way explain much confusion concerning  
ammonius assume two people called ammonius ammonius saccas taught plotinus ammonius christian  
wrote biblical texts another explanation might one ammonius origen found neo-platonist views  
teacher essential beliefs essential nature christianity chose suppress ammonius choice  
paganism christianity insistence eusebius origen pupil jerome recognized fathers christian
```


Xgboost & Overfitting

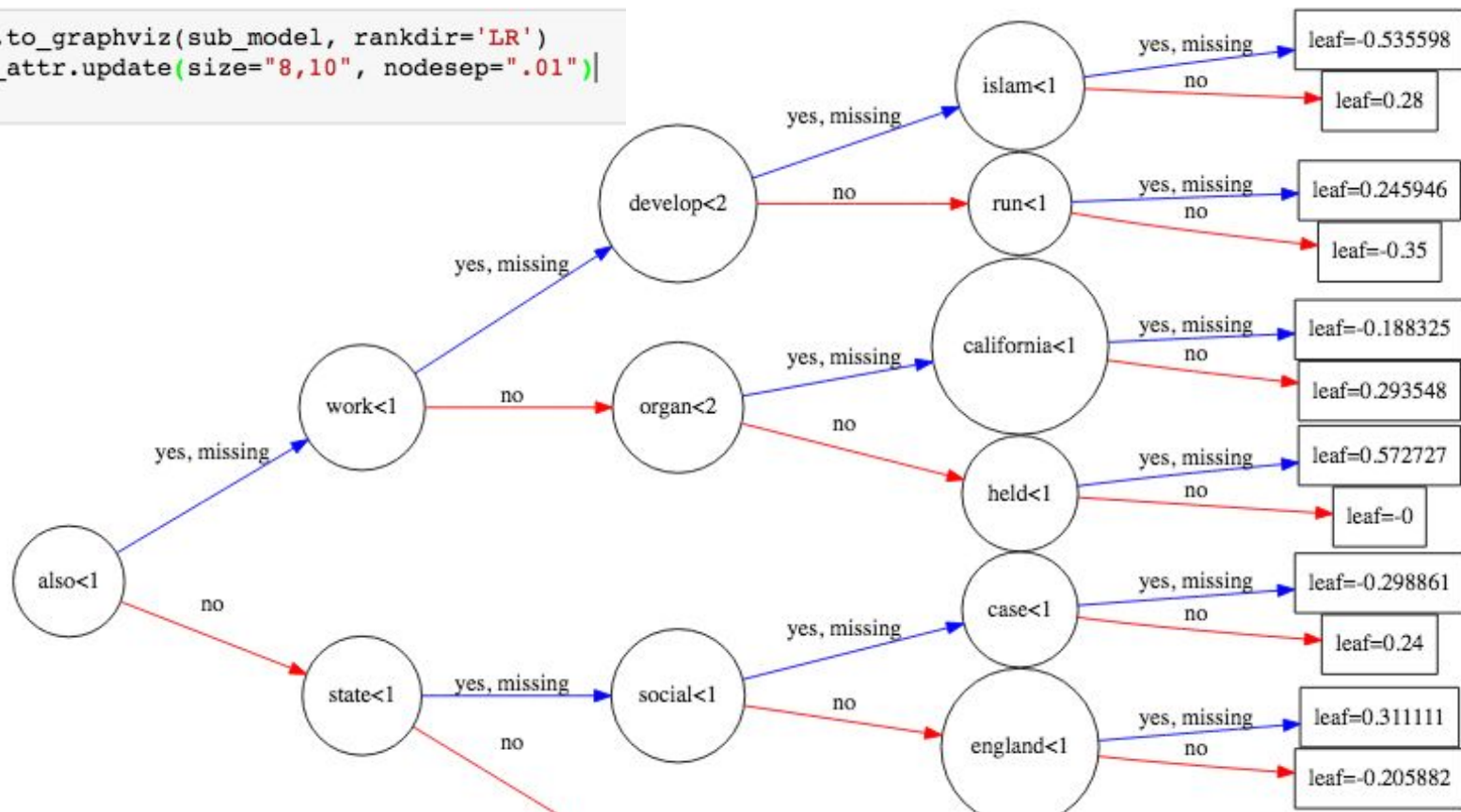
```
[0]      train-error:0.250343    val-error:0.307351
Multiple eval metrics have been passed: 'val-error' will be used for early stopping.
```

```
Will train until val-error hasn't improved in 50 rounds.
```

```
[20]      train-error:0.147328    val-error:0.211027
[40]      train-error:0.095706    val-error:0.186312
[60]      train-error:0.067839    val-error:0.159696
[80]      train-error:0.049338    val-error:0.149556
[100]     train-error:0.034491    val-error:0.143853
[120]     train-error:0.023984    val-error:0.13308
[140]     train-error:0.01873     val-error:0.132446
[160]     train-error:0.014619    val-error:0.126109
[180]     train-error:0.011421    val-error:0.119772
[200]     train-error:0.009593    val-error:0.119138
[220]     train-error:0.007081    val-error:0.116603
[240]     train-error:0.006167    val-error:0.115336
[260]     train-error:0.005482    val-error:0.114702
[280]     train-error:0.003426    val-error:0.112167
[300]     train-error:0.002969    val-error:0.112801
[320]     train-error:0.002056    val-error:0.1109
[340]     train-error:0.000685    val-error:0.108999
[360]     train-error:0.000685    val-error:0.112801
Stopping. Best iteration:
[327]     train-error:0.001599    val-error:0.106464
```

Tree Ensembles

```
In [45]: G = xgb.to_graphviz(sub_model, rankdir='LR')
G.graph_attr.update(size="8,10", nodesep=".01")
G
```

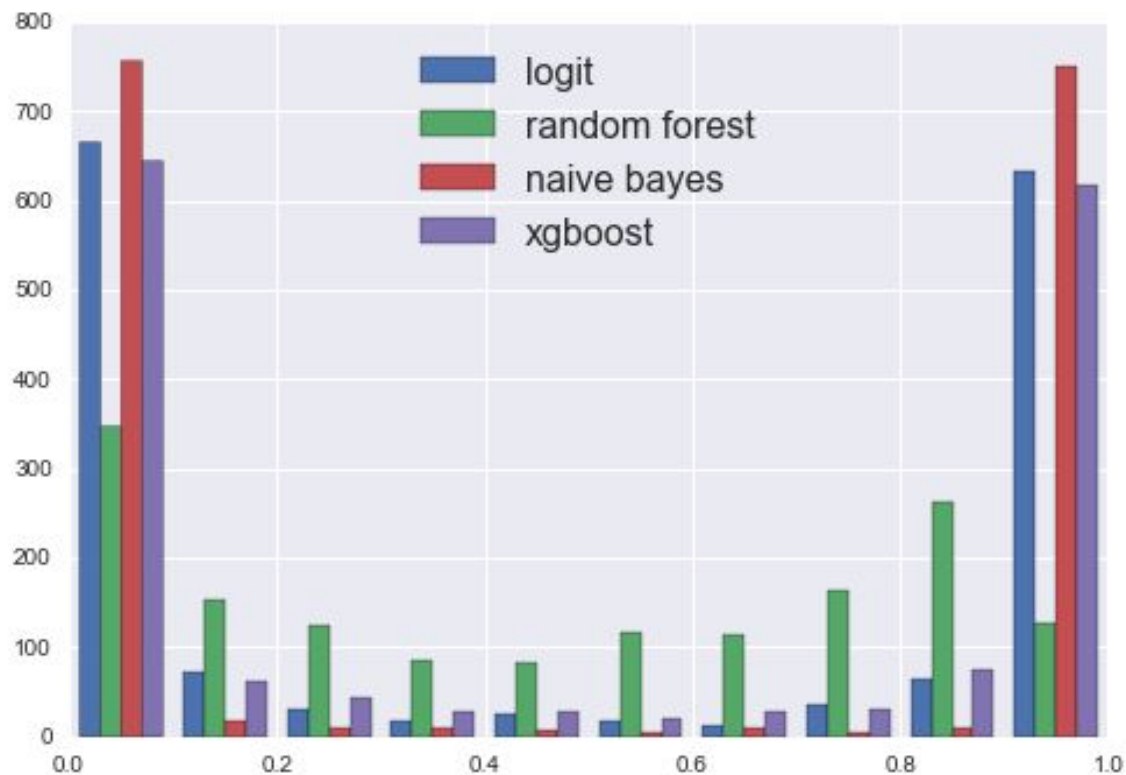


Model Comparison

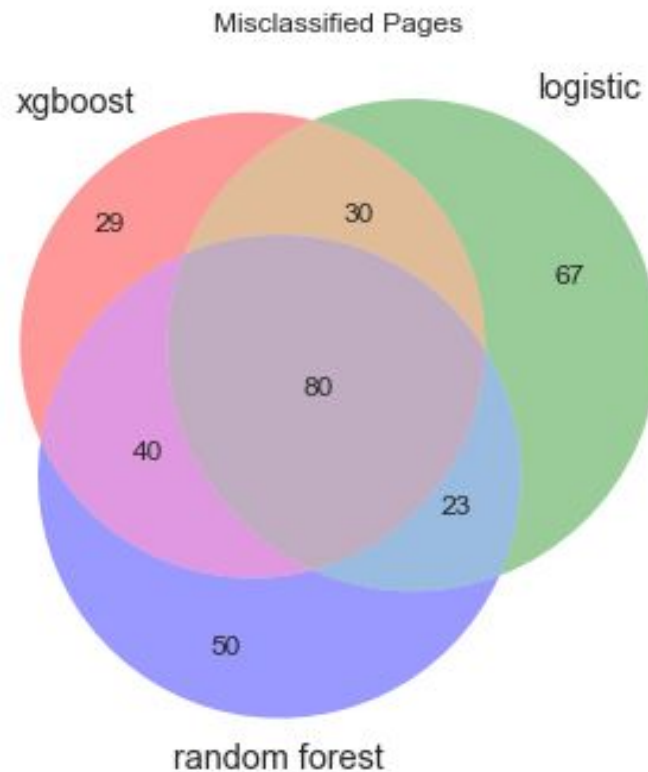
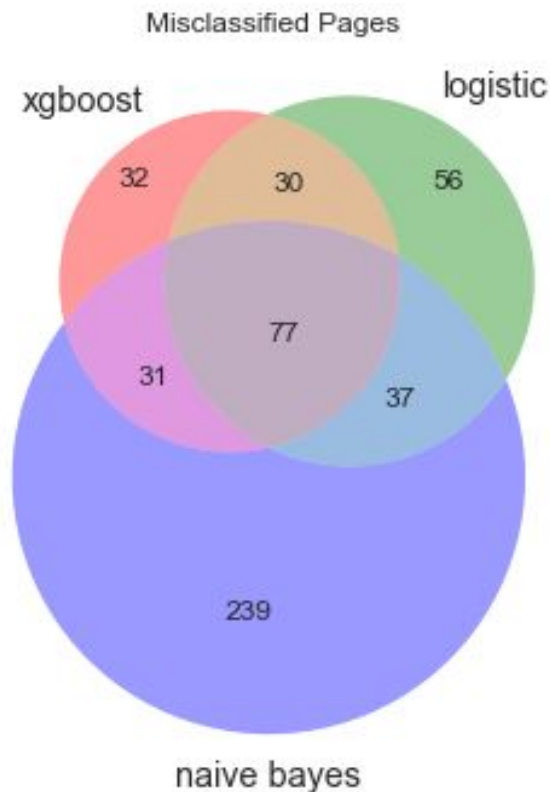
	Accuracy	Sensitivity	Specificity	ROC-AUC	Avg Logloss
logistic	0.873	0.889	0.860	0.900	0.785
forest	0.876	0.900	0.855	0.959	0.308
linear_svc	0.861	0.897	0.830	NaN	NaN
xgboost	0.887	0.908	0.868	0.943	0.341
naive_bayes	0.757	0.772	0.744	0.808	5.149

- Xgboost performs the best, followed by unboosted forests, and then logistic regression
- Specificity \approx Sensitivity: approximately equal performance between classes

Prediction Probabilities by Model



Misclassifications by Model



Future Directions

- Vandalism! Which Bayesian priors are most useful?
- Other tags (notability, weasel words, peacocking, advertisement, etc.)
- Graph theory on interconnectedness of various subjects
- Finish cosine similarity
- Citogenesis



A South American coati. In 2008, a 17-year-old student added an invented nickname to the Wikipedia article [coati](#) as a private joke, saying coatis were also known as "Brazilian aardvarks". The false information lasted for six years in Wikipedia and came to be propagated by hundreds of websites, several newspapers (one of which was later cited as a source in Wikipedia) and even books published by university presses.^{[14][15]}

Thank You!