Bibliometrics:
Predicting Publication Success

Team 4: Tamanna Baig and Jon Oakley

April 18th 2019

Clemson University

Introduction

Motivation

> “*Knowledge is power*”
>
> – Sir Francis Bacon

Motivation

- 2.5 million scientific papers published each year[1]
- Research trends
- Research *funding*

[1] The STM Report; Fourth Edition, March 2015

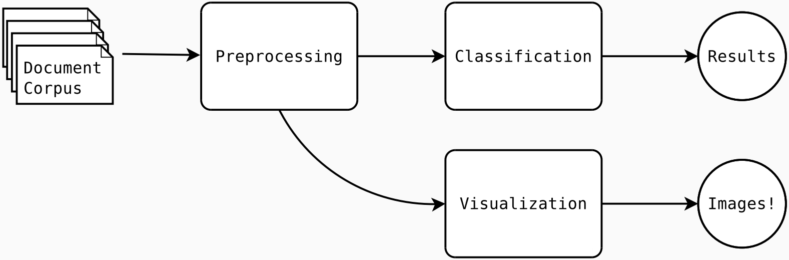
Background: Bibliometrics

Background: Bibliometrics

Defining the Problem

- Given: a corpus of academic publications
- Goal: predict which papers will be "successful"
- Success Metric: citation count
- Success Definition: citation count $>$ median citation count for that cluster
- Evaluation Metric: accuracy

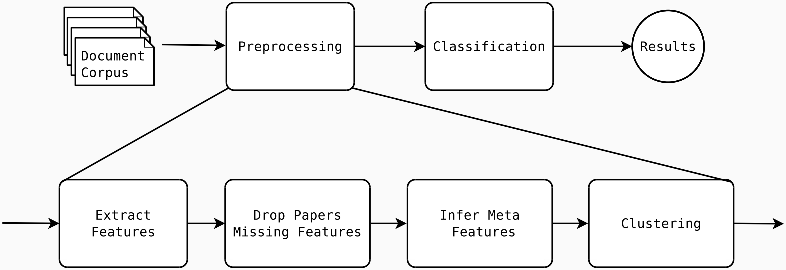
Overview



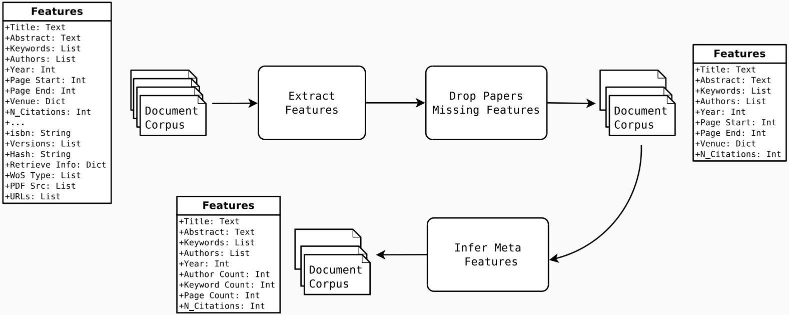
Document Corpus
Preprocessing
Classification
Results
Visualization
Images!

Preprocessing

Workflow


Document
Corpus
Preprocessing
Classification
Results
Extract
Features
Drop Papers
Missing Features
Infer Meta
Features
Clustering

Extraction, Thinning, and Meta Features



Features
+Title: Text
+Abstract: Text
+Keywords: List
+Authors: List
+Year: Int
+Page Start: Int
+Page End: Int
+Venue: Dict
+N_Citations: Int
+...
+isbn: String
+Versions: List
+Hash: String
+Retrieve Info: Dict
+WoS Type: List
+PDF Src: List
+URLs: List

Document Corpus

Extract Features

Drop Papers Missing Features

Document Corpus

Features
+Title: Text
+Abstract: Text
+Keywords: List
+Authors: List
+Year: Int
+Page Start: Int
+Page End: Int
+Venue: Dict
+N_Citations: Int

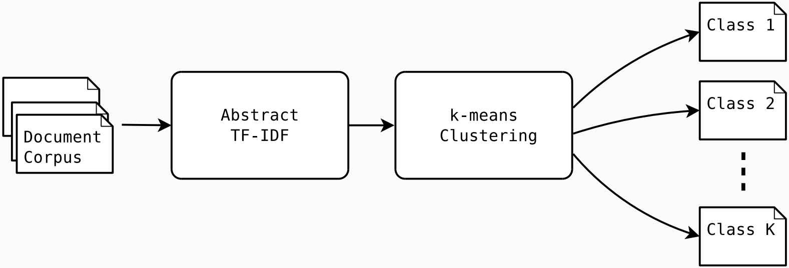
Infer Meta Features

Document Corpus

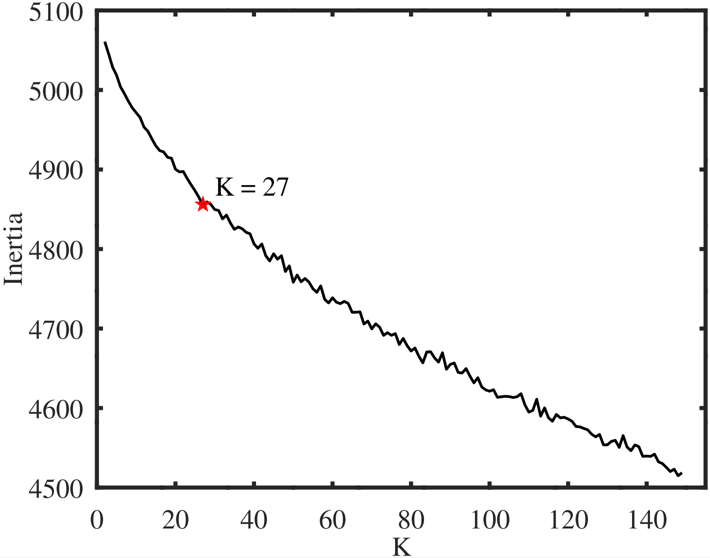
Features
+Title: Text
+Abstract: Text
+Keywords: List
+Authors: List
+Year: Int
+Author Count: Int
+Keyword Count: Int
+Page Count: Int
+N_Citations: Int

- Initial Corpus: 22,588
- Removed:
  - ☐ Papers after 2013
  - ☐ Outliers
  - ☐ Papers missing features
- Final Corpus: 4,914

Clustering


Document Corpus
Abstract TF-IDF
k-means Clustering
Class 1
Class 2
Class K

Clustering: Choosing K



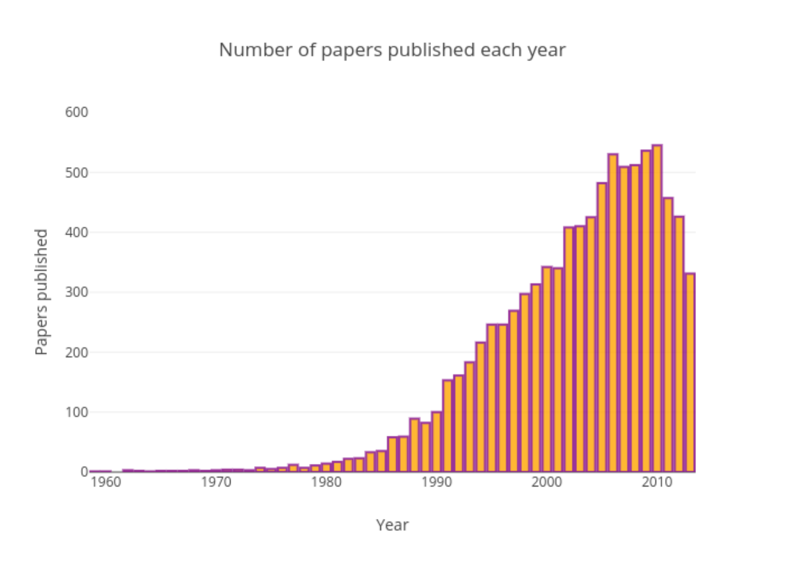
Clustering: Good Clusters

Introduction
Preprocessing
Visualization
Classification
Results
Conclusion

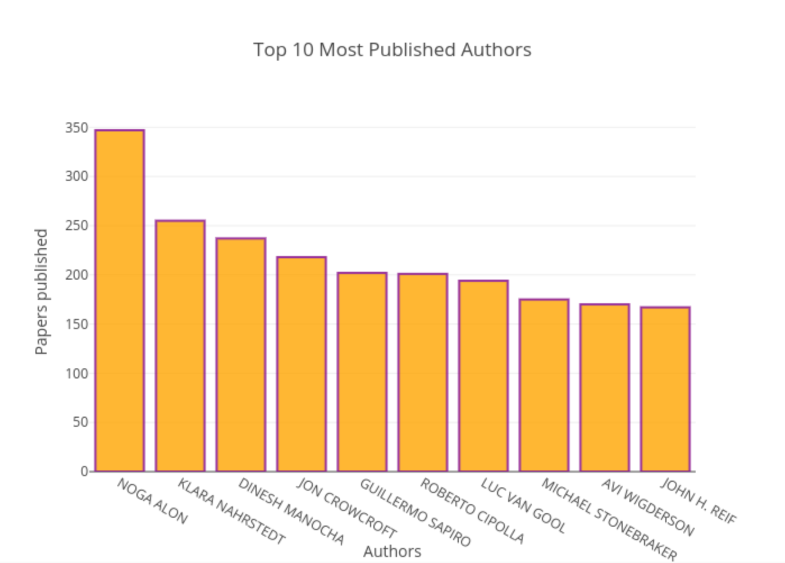
Cluster	Count	Keywords
~~0~~	~~15~~	~~alice, bob, girls, programming, communication~~
4	106	video, videos, 3D, quality, streams, users
15	170	query, XML, search, data, databases
17	869	design, people, user, information, research
20	93	internet, TCP, network, protocol, congestion
22	60	privacy, private, data, information, awareness

Visualization

Yearly Publications



Author Publications



Popular Publications

Introduction

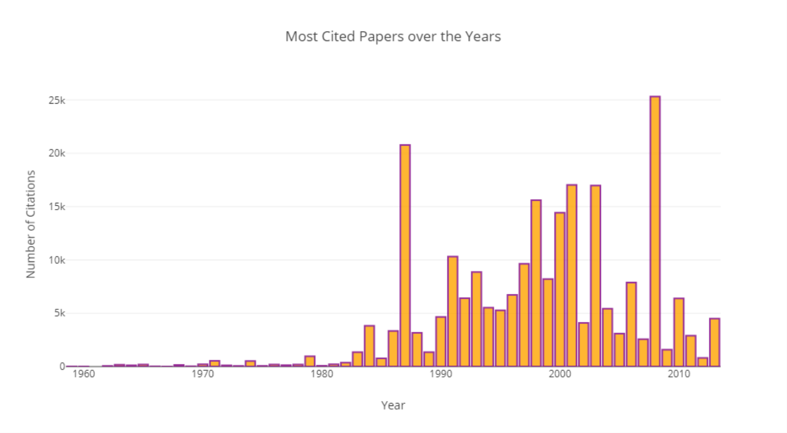
Preprocessing

Visualization

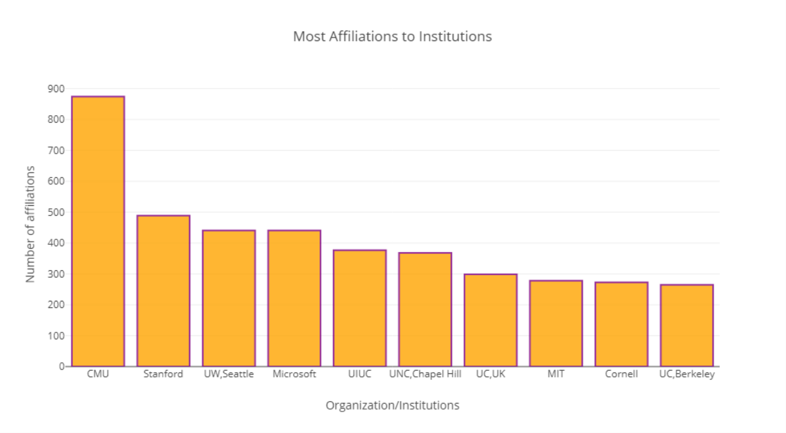
Classification

Results

Conclusion



Institution Publications



Classification

Workflow

Introduction

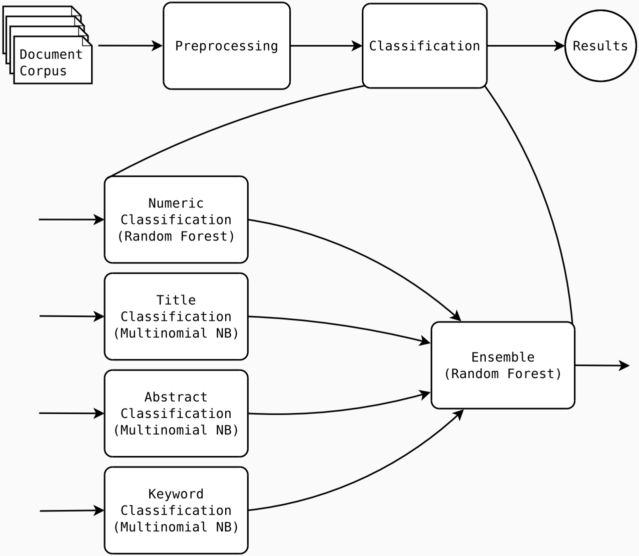
Preprocessing

Visualization

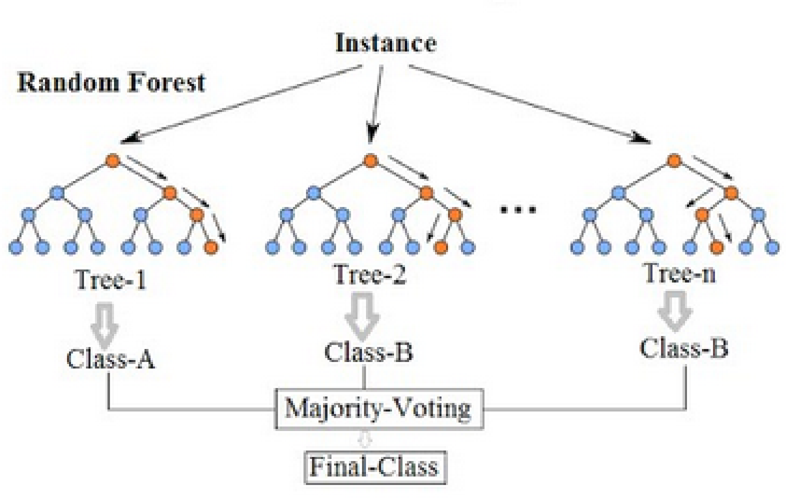
Classification

Results

Conclusion


Document Corpus
Preprocessing
Classification
Results
Numeric Classification (Random Forest)
Title Classification (Multinomial NB)
Abstract Classification (Multinomial NB)
Keyword Classification (Multinomial NB)
Ensemble (Random Forest)

Random Forest



https://www.kdnuggets.com/2017/10/random-forests-explained.html

Multinomial Naive Bayes

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c)P(c)$$

$$P(d|c) = P(f_1, f_2, ..., f_n|c) = P(f_1|c)P(f_2|c)...P(f_n|c)$$

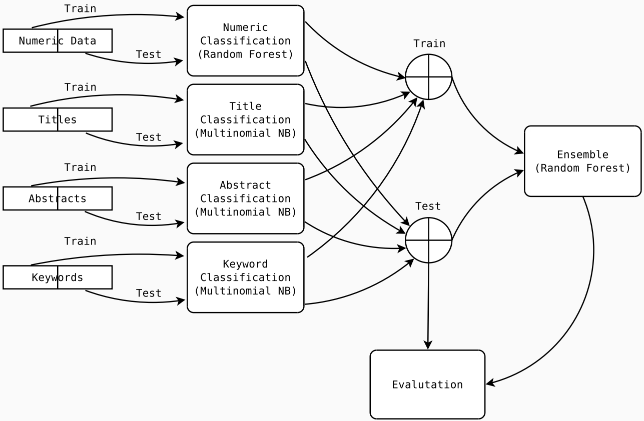
$$P(c) = \frac{N_c}{N_d}$$

- c: Class
- d: Document
- f_i: Feature
- N_c: Number of words in class c
- N_d: Number of words in document d

Classifier Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

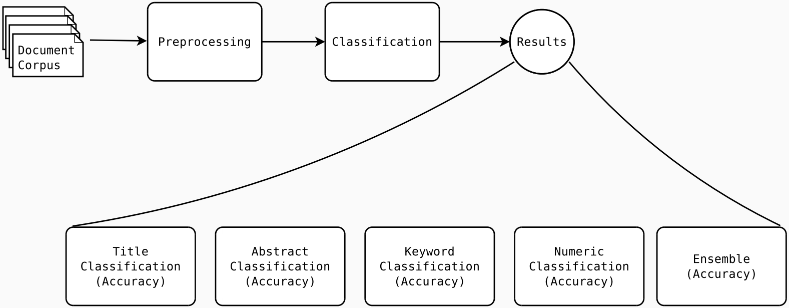
Training Method


Numeric Data
Titles
Abstracts
Keywords
Train
Test
Train
Test
Train
Test
Train
Test
Numeric Classification (Random Forest)
Title Classification (Multinomial NB)
Abstract Classification (Multinomial NB)
Keyword Classification (Multinomial NB)
Train
Test
Ensemble (Random Forest)
Evalutation

- test_train_split: 50%
- GridSearchCV
- CV = 5
- RF Params: n_estimators and max_features

Results

Workflow


Document Corpus
Preprocessing
Classification
Results
Title Classification (Accuracy)
Abstract Classification (Accuracy)
Keyword Classification (Accuracy)
Numeric Classification (Accuracy)
Ensemble (Accuracy)

Classification Parameters

	Numeric Classifier Parameters	
Cluster	n_estimators	max_features
4	10	2
15	10	6
17	8	2
20	12	2
22	6	4

Classification Parameters

	Ensemble Classifier Parameters	
Cluster	n_estimators	max_features
4	1	1
15	4	1
17	1	1
20	1	1
22	1	1

Feature Importance

Introduction
Preprocessing
Visualization
Classification
Results
Conclusion

	Numeric Feature Weight					
Cluster	Title	Abstract	Keyword	Year	Author	Page
	(length)	(length)	(length)		Count	Count
4	0.089	0.117	0.143	0.240	0.188	0.223
15	0.160	0.157	0.121	0.321	0.065	0.175
17	0.136	0.253	0.174	0.183	0.089	0.165
20	0.087	0.214	0.121	0.242	0.109	0.228
22	0.140	0.038	0.066	0.556	0.155	0.045

Feature Importance

	Ensemble Feature Weight			
Cluster	Numeric Classifier	Title Classifier	Abstract Classifier	Keyword Classifier
4	1	0	0	0
15	0.420	0.271	0.059	0.250
17	0.917	0	0.083	0
20	1	0	0	0
22	1	0	0	0

Classification Accuracy

	Accuracy (%)				
Cluster	Numeric	Title	Abstract	Keywords	Ensemble
4	60.4	45.8	56.3	58.3	60.4
15	63.0	64.2	65.4	55.6	61.7
17	64.2	59.1	59.6	62.5	68.5
20	64.3	42.9	35.7	38.1	64.3
22	64.3	0.5	0.5	0.5	64.3

– Overall Accuracy (All Clusters): 57.7%

– Overall Accuracy (4,15,17,20,22): 66.5%

Conclusion

Conclusions

Introduction

Preprocessing

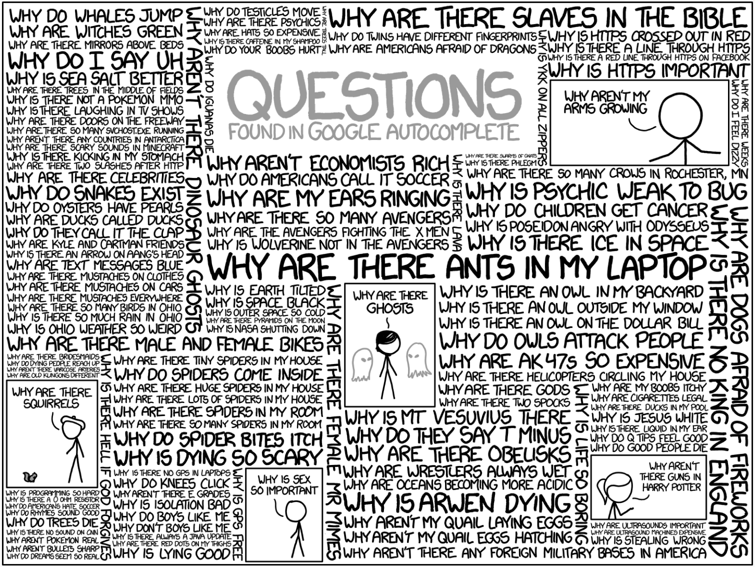
Visualization

Classification

Results

Conclusion

- Cleaning data is important
- Some clusters are better than others
- **66.5% prediction accuracy in optimal clusters**
- Future work: targeted dataset
- Controlling for year


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE TREES TALL
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

https://xkcd.com/1256