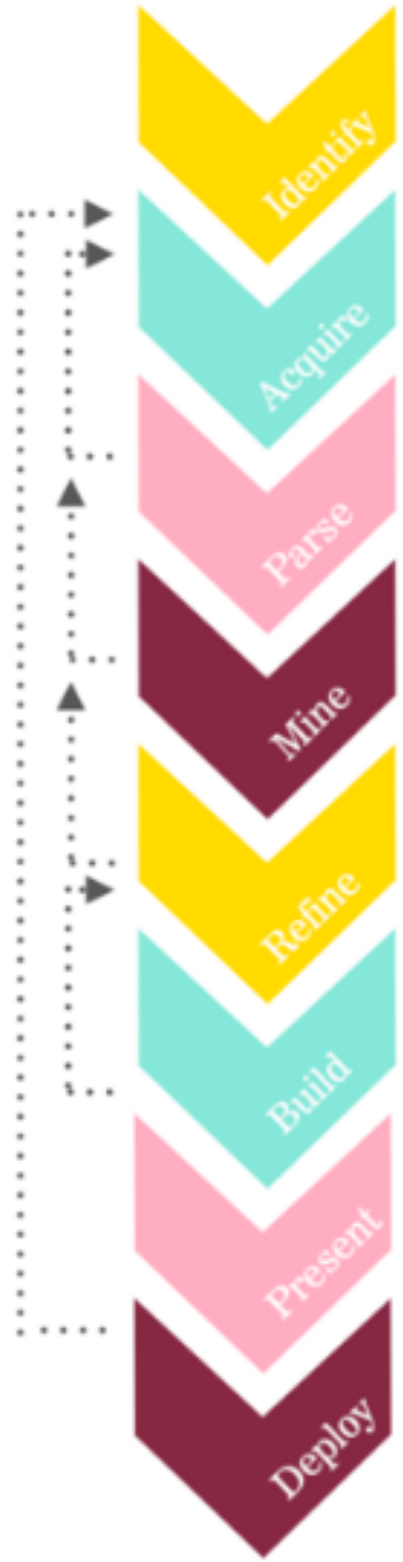
Ian Perry

Capstone Project

Project

DATA SCIENCE WORKFLOW



IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT RESULTS

- ☐ Summarize with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up issues for future analysis

DEPLOY AND VALIDATE

- ☐ Write unit tests and documentation
- ☐ Deploy stable production-ready code
- ☐ Retrain and validate models over time

DataSet / Cleaning

Rank	Title	Category	User	Views	Upload	#	+/-	URL	Duration	Descript...	Keyword	Keyword	Keyword
2	Speed of	Science &	Science...	345732	2016-09-11	413	1160/108	http://...	44:28	For more	yes	no	0.00%
3	The	Education	Real	22592	2016-11-15	171	213/11	http://...	50:09	"You feel	yes	no	0.00%
4	[Blow	Science &	Discovery	550	2016-11-17	0	53/2	http://...	0s	LIKE -------	yes	yes	8.50%
5	National	Science &	Star Staff	14277	2016-11-14	10	56/14	http://...	44:03	The	yes	no	0.00%
6	Anonym...	Nonprofits	Anonym...	734223	2016-07-24	3073	10538/298	http://...	58:00	Anonym...	yes	yes	2.01%
7	Living	Education	Real	122780	2016-11-11	483	829/64	http://...	52:02	An	yes	yes	0.32%

Successes

- Three laws
- Fascinating Topic
- A small difference
- NLTK

Challenges

- The API
- Data Capture
- Duplication
- Skewness in Dislikes

For example


This video is no longer available due to a copyright claim
Sorry about that.

Top Docs

Top Likes

- 3 - Beyonce
- 2 - Cyber Bully
- 1 - Suicide Forests in Japan.
  (Viewer Discretion)

Top Dislikes

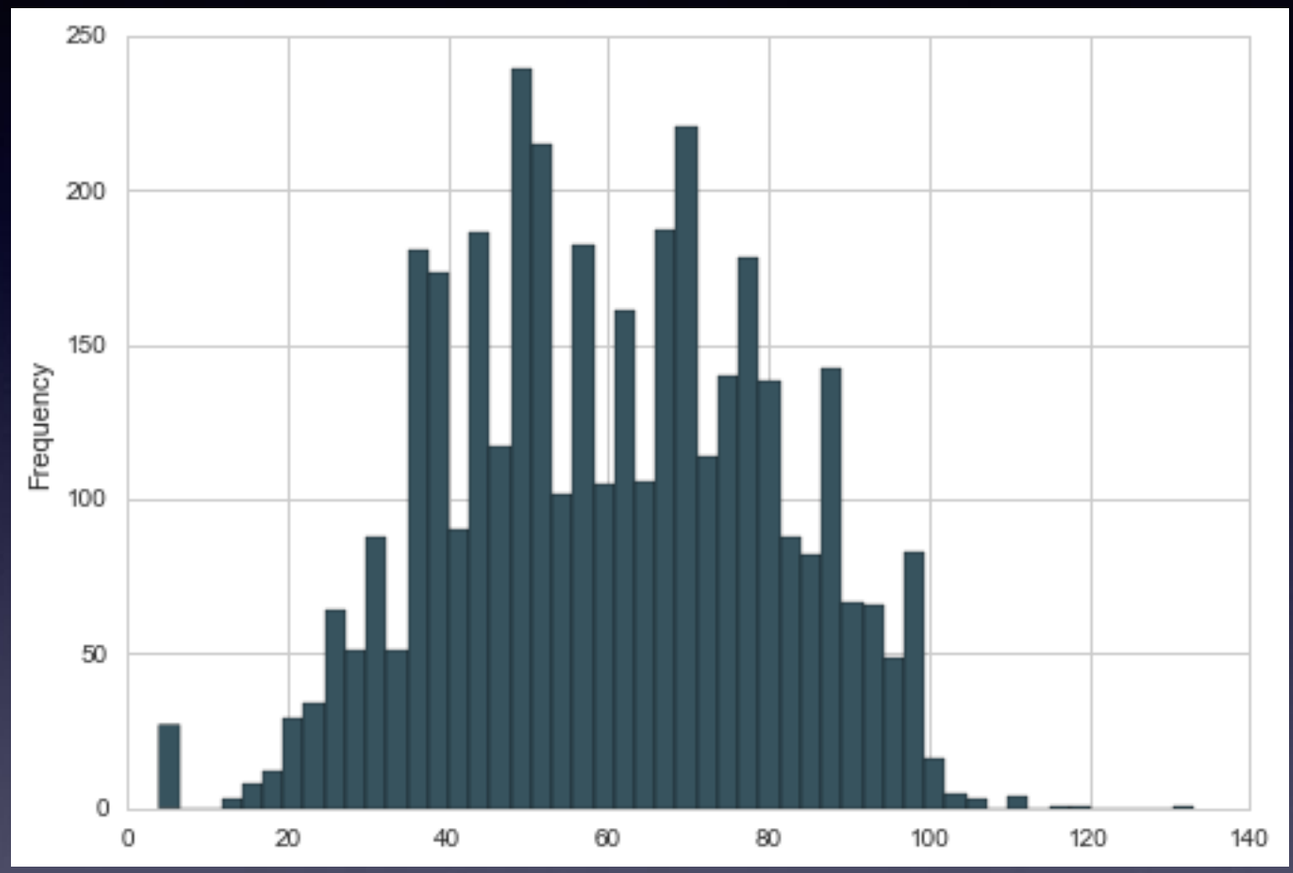
- 3 - Brexit
- 2 - Living as a baby
- 1 - Cult Leader Thinks He's Jesus

Top Cats

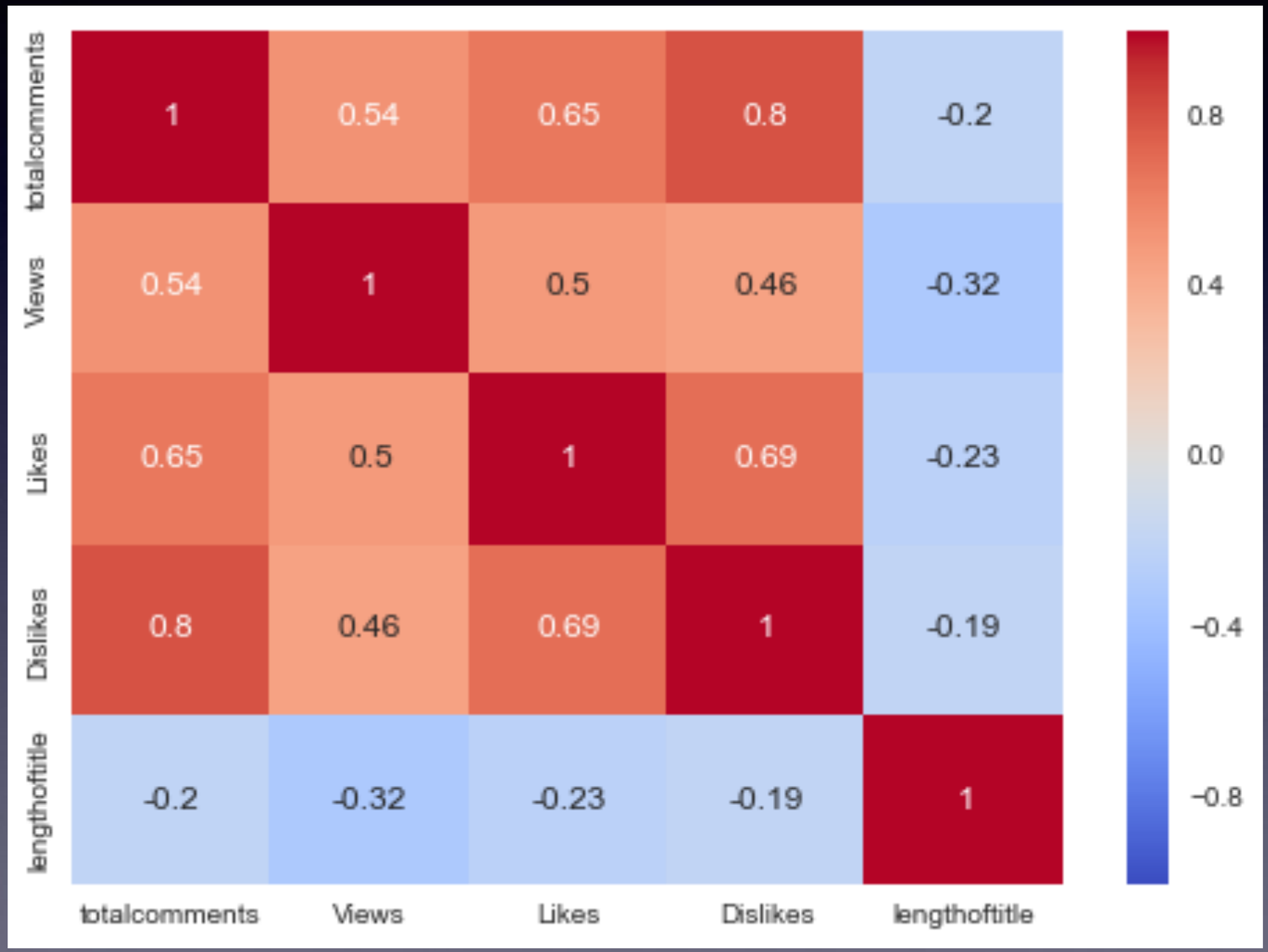
Top Categories

- 3 - Education
- 2 - Film and animation
- 1 - People and Blogs

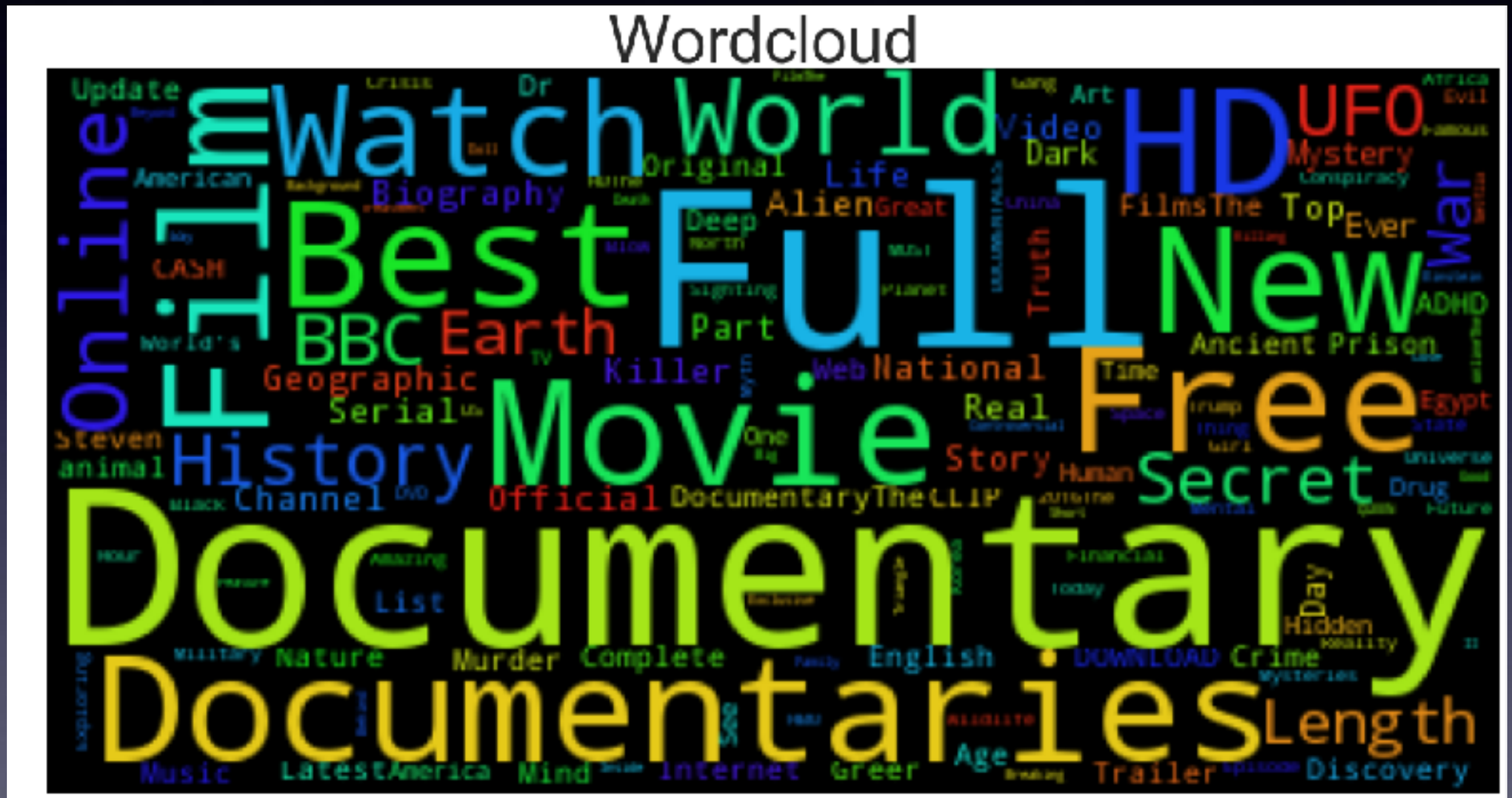
Title Length



Length of Title



Wordcloud


Wordcloud
Update
Dr
Art
Video
HD
UFO
Mystery
Conspiracy
Dark
Original
Life
American
Biography
Alien
Great
Deep
FilmsThe
Top
Ever
War
Best
Full
New
Film
CASH
Truth
ADHD
BBC
Earth
Part
Ancient
Prison
Geographic
Killer
Web
National
Time
Free
Movie
Real
Egypt
Serial
Steven
History
One
Story
Human
Secret
Drug
animal
Channel
Official
DocumentaryTheCLIP
Documentary
List
Financial
Day
Hidden
Nature
Murder
Complete
English
DOWNLOAD
Crime
Documentaries
Length
Music
LatestAmerica
Mind
Internet
Greer
Age
Trailer
Discovery

Modeling

	Bayes	LR	TFIDF
Accruacy	0.79	0.82	0.81
Precision	0.80	0.81	0.83
Recall	0.80	0.81	0.83
F1	0.80	0.81	0.83
Support	1143	1143	1258

Repeat

Questions