


# Click Bait or Not?




# Remember those TF-IDF Scores!


We can use them for more stuff! Like sorting stuff and learning patterns to sort clickbait from real news articles!



### 17 Facts You Won't Believe Are True


Hope you're sitting down for this one.


 **Jessica Misener**  
posted on Oct 24, 2014



### 42 Wrestlers You Won't Believe Actually Existed


#NeverForget

 **Nick Wray**  
posted on Nov 07, 2014



### Why BuzzFeed Doesn't Do Clickbait

You won't believe this one weird trick.

 **Ben Smith**  
posted on Nov 06, 2014



# Remember those TF-IDF Scores!

## 1. Find meaningful words or “features”.

We filled one document with clickbait headlines, and another with news headlines.

Then we used variants of **TF-IDF** to find the most meaningful words for each document.

things,	0.00166862287567
truth,	0.00114717822702
movies,	0.000938600367565
hollywood,	0.000834311437835
video,	0.000834311437835
movie,	0.000730022508106
everyone,	0.000677878043241
didn't,	0.000625733578377

Some get a much  
high score in  
**Clickbait**  
E.g. “Believe”

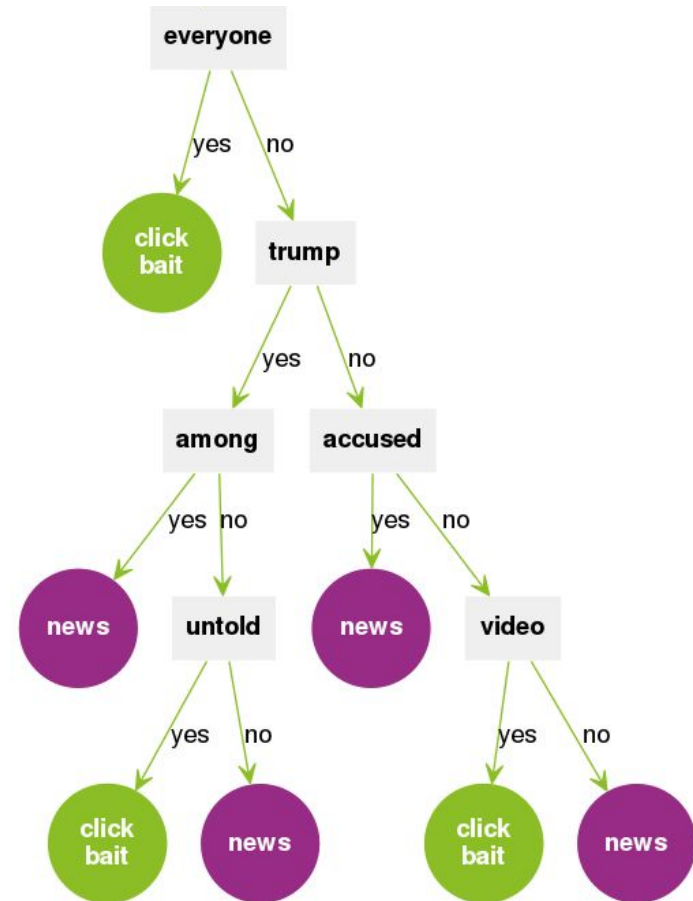
Some get a much  
high score in  
**news articles**  
E.g. “Education”



# Click Bait or Not?

We use those numbers to make a few different decision trees!

They sort clickbait and news article!



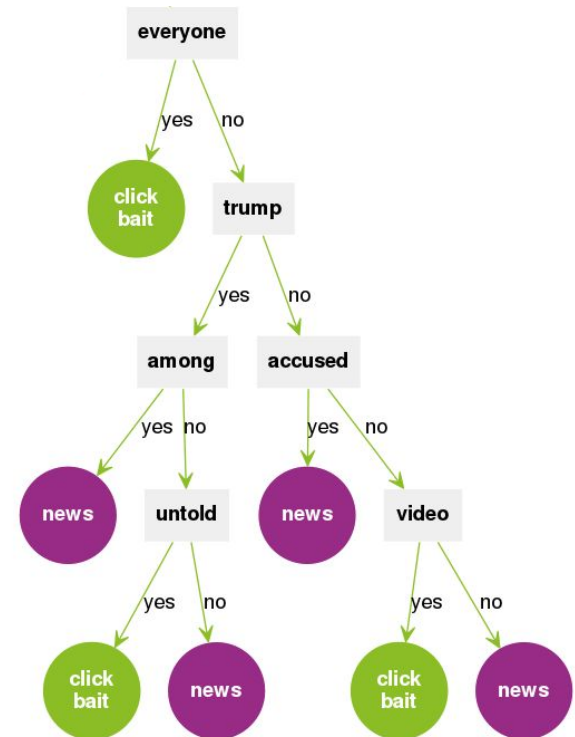
# But which decision tree is the best??

We're going to find out!

1. Get a story headline!
2. Run it through the decision tree based on the words in the headline
3. See if it gets it correct!
4. Repeat for a bunch of headlines of your choice, to see how it does on average!

Keep count in a table like this!

		Headline is actually...	
		Clickbait	News
Decision tree thinks headline is...	Clickbait	(True positives)	(False negatives)
	News	(False positives)	(True negatives)



# Find the best one!

		Headline is actually...	
		Clickbait	News
Decision tree thinks headline is...	Clickbait	(True positives)	(False negatives)
	News	(False positives)	(True negatives)

**We want the one with the best accuracy!**

$$\frac{\text{true positives} + \text{true negatives}}{\text{total number of headlines tested}} = \text{accuracy}$$

**When you have found the most accurate one for your set of headlines, put a sticker on it!!**

