Word Information

Word Information

Words carry information. Exactly how much information you can get from a word depends on 2 things.

1. How **often** the word appears.

1. How **common** the word is in general.

Word Information

This song contains the word **'the'** which appears **2** times.

'the'

1. Appears often
1. Not common

Not informative

Word Information

This song contains the word **‘the’** which appears **20** times.

‘the’

1. Appears often
1. Not common

Not informative

Word Information

This song contains the word **'love'** which appears **1** time.

'love'

1. Appears often
1. Not common

Not informative

Word Information

This song contains the word **'love'** which appears **22** times.

'love'

1. Appears often
1. Not common

Very informative

Word Information

This song contains the word **'love'** which appears **22** times.

'love'

1. Appears often
1. Not common

Very informative

Tells us this song is probably a love song.

How do we score this?

Words carry information. Exactly how much information you can get from a word depends on 2 things.

1. How **often** the word appears.

1. How **common** the word is in general.

How do we score this?

Words carry information. Exactly how much information you can get from a word depends on 2 things.

1. How **often** the word appears.

1. How **common** the word is in general.

How do we score this?

$$\text{Term frequency (TF)} = \frac{\text{Number of times a word is in a song}}{\text{Total number of words in a song}}$$

What is the TF score for the word 'birthday' in the Happy Birthday song?

*Happy **birthday** to you*
*Happy **birthday** to you*
*Happy **birthday** to Alex*
*Happy **birthday** to you*

How do we score this?

$$\text{Term frequency (TF)} = \frac{\text{Number of times a word is in a song}}{\text{Total number of words in a song}} = \frac{4}{16} = 0.25$$

What is the TF score for the word 'birthday' in the Happy Birthday song?

*Happy **birthday** to you*
*Happy **birthday** to you*
*Happy **birthday** to Alex*
*Happy **birthday** to you*

How do we score this?

Words carry information. Exactly how much information you can get from a word depends on 2 things.

1. How **often** the word appears.

1. How **common** the word is in general.

How do we score this?

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}}$$

How do we score this?

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}}$$

What is the IDF score for the word 'little' for the following playlist of songs?

- ***Mary had a little lamb***
- *Happy birthday*
- ***Twinkle twinkle little star***
- *The Australian anthem*

How do we score this?

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}} = \frac{4}{2} = 2$$

What is the IDF score for the word 'little' for the following playlist of songs?

- ***Mary had a little lamb***
- *Happy birthday*
- ***Twinkle twinkle little star***
- *The Australian anthem*

How do we score this?

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}}$$

What is the IDF score for the word 'birthday' for the following playlist of songs?

- *Mary had a little lamb*
- ***Happy birthday***
- *Twinkle twinkle little star*
- *The Australian anthem*

How do we score this?

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}} = \frac{4}{1} = 4$$

What is the IDF score for the word 'birthday' for the following playlist of songs?

- *Mary had a little lamb*
- ***Happy birthday***
- *Twinkle twinkle little star*
- *The Australian anthem*

How do we score this?

Gets bigger when the word is less common!

$$\text{Inverse document frequency (IDF)} = \frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}} = \frac{4}{1} = 4$$

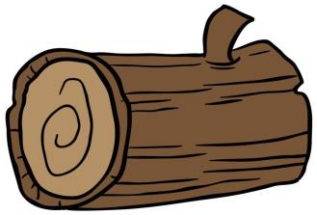
What is the IDF score for the word 'birthday' for the following playlist of songs?

- *Mary had a little lamb*
- ***Happy birthday***
- *Twinkle twinkle little star*
- *The Australian anthem*

Adding a log

$$\text{Inverse document frequency (IDF)} = \log\left(\frac{\text{Total number of songs}}{\text{Number of different songs the word appears in}}\right)$$

To stop our IDF score getting to large we add a log!



A little on log!

Putting it all together

We represent the usefulness of a word with a TF IDF score!

TF IDF = Term frequency **TF** x Inverse document frequency **IDF**

Appears often

Not common

Activity 1: 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!

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
```

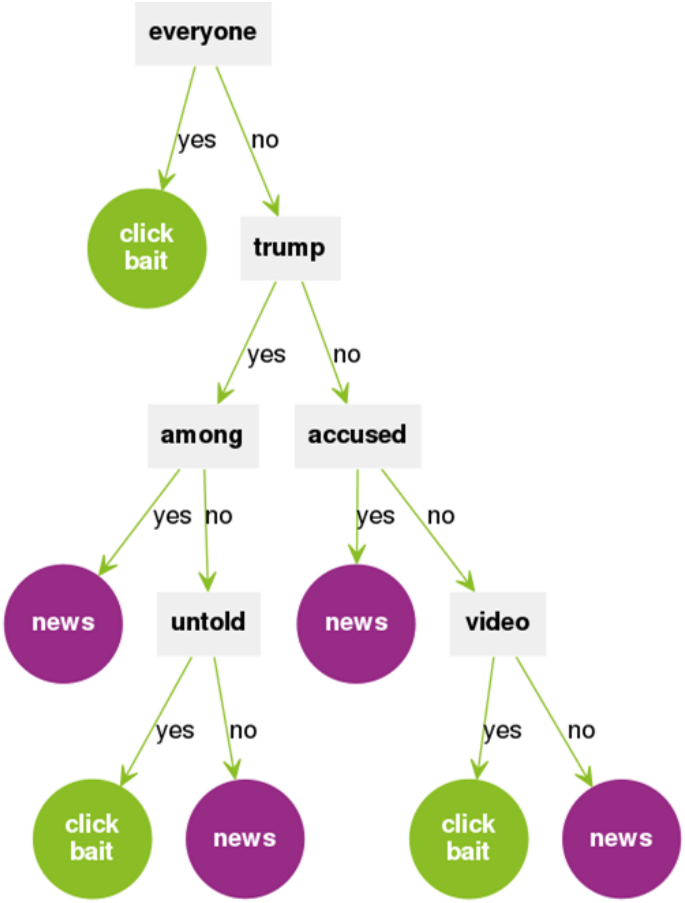
Some get a much high score in Clickbait E.g. "Believe"

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!



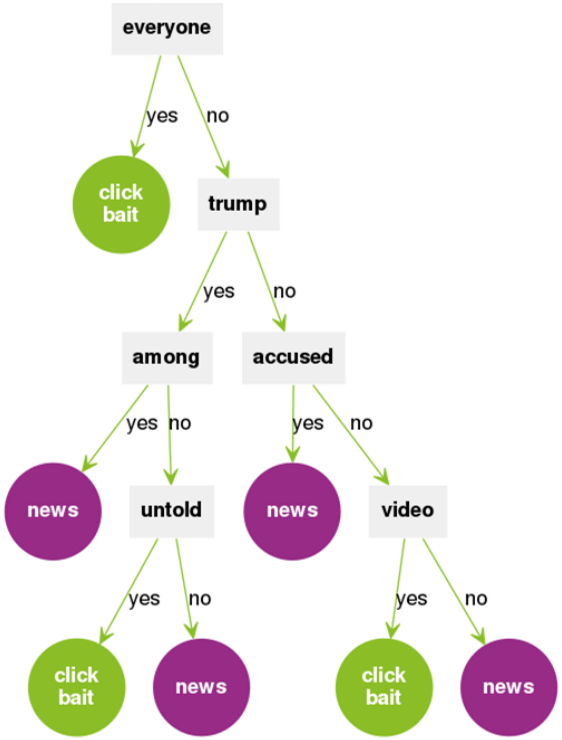
But which decision tree is the best??

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!

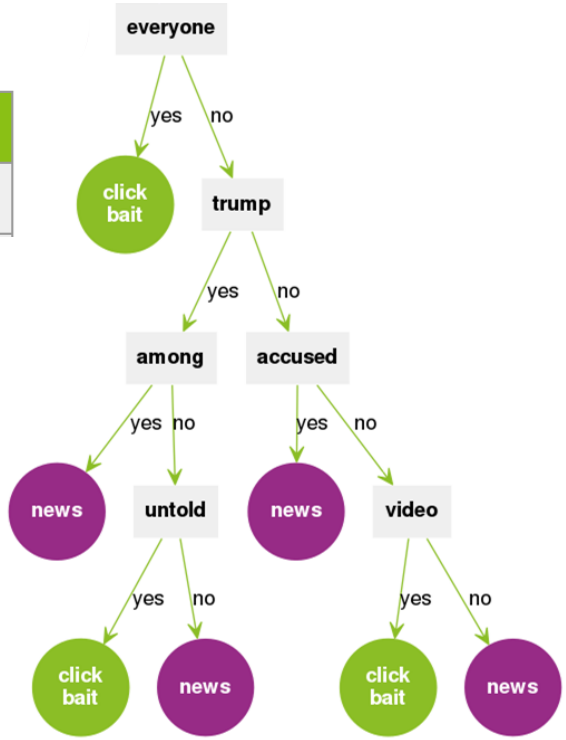
		Headline is actually...	
		Clickbait	**News**
Decision tree thinks headline is...	**Clickbait**	(True positives)	(False negatives)
	News	(False positives)	(True negatives)



Let’s try this!

Clickbait	Tree 1	Tree 2	Tree 3
The 100 Worst Endings Ever in Video Games			

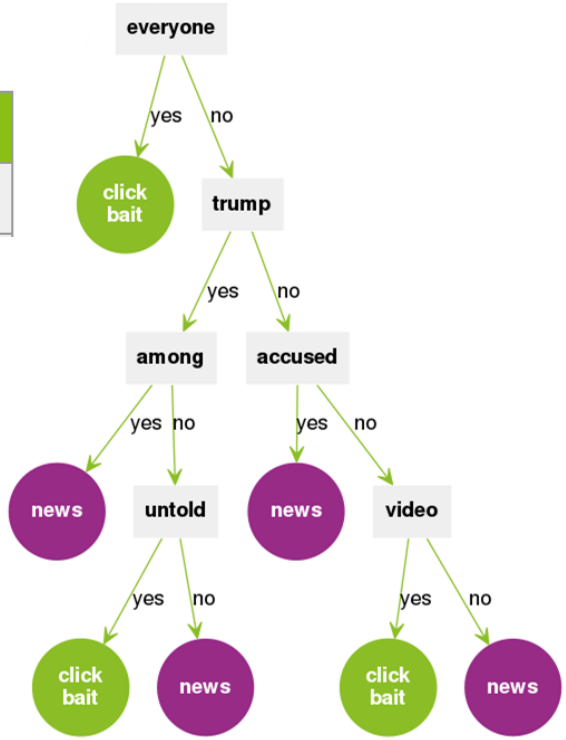
		Headline is actually...	
		Clickbait	News
Decision tree thinks headline is...	Clickbait	(True positives)	(False negatives)
	News	(False positives)	(True negatives)



Let’s try this!

Clickbait	Tree 1	Tree 2	Tree 3
The 100 Worst Endings Ever in Video Games	✔		

		Headline is actually...	
		Clickbait	News
Decision tree thinks headline is...	Clickbait	1	0
	News	0	0



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!

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

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

Activity 2: Similar Song Suggester

Find the best matching song for a phrase

The activity:

- You'll be given a phrase:
- You will need to calculate the TF IDFs for each word in a phrase when compared to a particular song.
- You can find all the data you need on the walls to calculate the TF IDFs!

(*We've done the logs for you!*)

Let’s try this!

Say you believe in love

Let’s try this!

Say you believe in love

Step 1) Find the IDF for each word in the phrase

Word	SAY	YOU	BELIEVE	IN	LOVE
IDF					

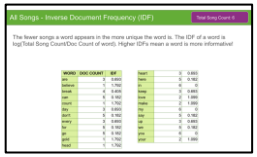
Let’s try this!

Say you believe in love

1

Step 1) Find the IDF for each word in the phrase

Word	SAY	YOU	BELIEVE	IN	LOVE
IDF	0.182	0	1.792	0	1.099


All Songs - Inverse Document Frequency (IDF)

Total Song Count: 5

The fewer songs a word appears in the more unique the word is. The IDF of a word is log(Total Song Count/Doc Count of word). Higher IDFs mean a word is more informative!

WORD	DOC COUNT	IDF			
are	3	0.693	heart	2	0.693
believe	1	1.792	hello	5	0.182
break	4	0.405	in	0	0
count	4	0.405	keep	3	0.693
count	1	1.792	leave	2	0.693
fly	3	0.693	letter	2	1.099
don't	5	0.182	my	0	0
every	3	0.693	see	5	0.182
for	5	0.182	up	5	0.405
go	5	0.182	we	5	0.182
gotcha	1	1.792	you	5	0
hand	1	1.792	good	2	1.099

Let’s try this!

Say you believe in love

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Song 1: Flowers

Total word count: _____

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$	IDF (from step 1)	TF-IDF (TF x IDF)
SAY				
YOU				
BELIEVE				
IN				
LOVE				
			Total TF-IDF	

Let’s try this!

Say you believe in love

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Song 1: Flowers

Total word count: 332

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$	IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3			
YOU	13			
BELIEVE	0			
IN	3			
LOVE	22			
			Total TF-IDF	

2

Let's try this!

Say you believe in love

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Song 1: Flowers

Total word count: 332

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$	IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3	0.009		
YOU	13	0.039		
BELIEVE	0	0		
IN	3	0.009		
LOVE	22	0.066		
			Total TF-IDF	

Let’s try this!

Say you believe in love

1

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Word	SAY	YOU	BELIEVE	IN	LOVE
IDF	0.182	0	1.792	0	1.099

Song 1: Flowers

Total word count: 332

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$	IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3	0.009	0.182	
YOU	13	0.039	0	
BELIEVE	0	0	1.792	
IN	3	0.009	0	
LOVE	22	0.066	1.099	
			Total TF-IDF	

Let's try this!

Say you believe in love

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Song 1: Flowers **Total word count: 332**

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$		IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3	0.009		0.182	0.02
YOU	13	0.039		0	0
BELIEVE	0	0	X	1.792	0
IN	3	0.009		0	0
LOVE	22	0.066		1.099	0.073
				Total TF-IDF	

Let's try this!

Say you believe in love

Step 2) Calculate the TF-IDF for each of the songs

Find the word count and total word count for each song on the posters!

Song 1: Flowers

Total word count: 332

	Word Count	TF $\frac{\text{Word Count}}{\text{Total word count for song}}$	IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3	0.009	0.182	0.02
YOU	13	0.039	0	0
BELIEVE	0	0	1.792	0
IN	3	0.009	0	0
LOVE	22	0.066	1.099	0.073
			Total TF-IDF	0.075

Add these!

Let’s try this!

Say you believe in love

Step 3) Calculate the total TF-IDF for each song using the tables below! The highest score is the best match!

Total TF-IDF’S

	Total TF-IDF
Dance the night	
Flowers	0.075
Let it go	
September	
Shake it off	
We don’t talk about Bruno	

Song 1: Flowers — **Total word count: 332**

	Word Count	TF (Word Count / Total word count for song)	IDF (from step 1)	TF-IDF (TF x IDF)
SAY	3	0.009	0.182	0.02
YOU	13	0.039	0	0
BELIEVE	0	0	1.792	0
IN	3	0.009	0	0
LOVE	22	0.066	1.099	0.073
			Total TF-IDF	0.075

Let’s try this!

Say you believe in love

Step 3) Calculate the total TF-IDF for each song using the tables below! The highest score is the best match!

Total TF-IDF’S

	Total TF-IDF
Dance the night	
Flowers	0.075
Let it go	
September	
Shake it off	
We don’t talk about Bruno	

Use these to work out the best song suggestion!

Best suggestion!

Largest TF-IDF Score
Best Song Title