There are various approaches for converting text into vector

- Label Encoding
- One Hot Encoding
- Bag Of Words
  Bag of n-grams
- TF-IDF
- Word Embeddings

Tesla

Elon Musk wants time to prepare for a trial over his contentious withdrawal from an agreement to buy Twitter for $44 billion, according to a filing in a Delaware chancery court by his attorneys on Friday.

Musk's team says the trial should wait until next year, after Twitter had requested expedited treatment and a hearing as early as this September.

Apple

Apple CEO Tim Cook says there's at least one question your iPhone can't answer: How do you build a life that provides both meaning and fulfillment?

Luckily, Cook himself has some advice on that front. At Gallaudet University's commencement ceremony in Washington D.C. last month,

Apple

Apple leaks have revealed that iPhone 14 Pro models will receive mul upgrades and higher pricing. But now a new report claims standard iP will also receive a price bump, despite being virtually unchanged from predecessors.

Speaking to The Sun, Dan Ives, head of popular analyst group Wedbus supply chain prices as the driving force behind the increase.

Tesla

The Tesla Model 3 Long Range topped the charts with a 74.2/100 "greenpower" rating. Model 3 LR has a battery size of 83.0 kWh and a charging speed of 18.0 miles/minute.

Meanwhile, the Lucid Air ranked second in the ratings with 70.5/100 greenpower. It's ir despite having the largest battery of 112 kWh among all the cars that made it to the list. than the Model 3 LR at 451 miles versus Tesla's 358 miles. However, the Air's charging s

Apple

As Apple seeks to bolster its subscription service to compete wi company is adding exclusive performances from artists, who w Apple's studios to cover classics and recreate their own hits.

On Friday, the company introduced Apple Music Sessions, feat from singers including Carrie Underwood and Tenille Townes.

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
article 1	[0	32	45	48	26	7	3	0	...]
article 2	[0	4	3	7	0	6	3	0	...]
article 3	[15	31	44	43	25	0	0	0	...]
article 4	[3	0	0	0	0	0	0	1	...]

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
article 1	[0	32	45	48	26	7	3	0	...]
article 2	[0	4	3	7	0	6	3	0	...]
article 3	[15	31	44	43	25	0	0	0	...]
article 4	[3	0	0	0	0	0	0	1	...]


musk that price market investor iphone itunes gigafactory ...
Apple
article 1
[0 32 45 48 26 7 3 0 ...]
Apple
article 2
[0 4 3 7 0 6 3 0 ...]
article 3
[15 31 44 43 25 0 0 0 ...]
article 4
[3 0 0 0 0 0 0 1 ...]


musk
that
price
market
investor
iphone
itunes
gigafactory
...
Apple
article 1
[0 32 45 48 26 7 3 0 ...]
Apple
article 2
[0 4 3 7 0 6 3 0 ...]
Tesla
article 3
[15 31 44 43 25 0 0 0 ...]
Tesla
article 4
[3 0 0 0 0 0 0 1 ...]

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	0	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]


musk that price market investor iphone itunes gigafactory ...
Apple
article 1
[0 32 45 48 26 7 3 0 ...]
Apple
article 2
[0 4 3 7 0 6 3 0 ...]
Tesla
article 3
[15 31 44 43 25 0 0 0 ...]
Tesla
article 4
[3 0 0 0 0 0 0 1 ...]


I'M NOT SAYING
IGNORE THEM COMPLETELY
BUT STILL IGNORE THEM SOMEHOW
imgflip.com

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	0	32	45	48	26	7	3	0	...
Apple article 2	0	4	3	7	8	6	3	0	...
Tesla article 3	15	31	44	43	25	0	0	0	...
Tesla article 4	3	0	0	0	0	0	0	1	...

that → 3

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

that → 3 gigafactory → 1 iphone → 2

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	0	32	45	48	26	7	3	0	...
Apple article 2	0	4	3	7	8	6	3	0	...
Tesla article 3	15	31	44	43	25	0	0	0	...
Tesla article 4	3	0	0	0	0	0	0	1	...

that → 3 gigafactory → 1 iphone → 2

$$Document\ Frequency\ (DF) = Number\ of\ times\ term\ \mathbf{t}\ is\ present\ in\ all\ docs$$

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

$$\left(\frac{\text{Total Documents}}{\text{Number of documents term } t \text{ is present in}}\right)$$

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

that → 4/3 -> 1.33

$$\left(\frac{\text{Total Documents}}{\text{Number of documents term } t \text{ is present in}}\right)$$

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

that → 4/3 -> 1.33 gigafactory → 4/1 → 4 iphone → 4/2 → 2

$$\left(\frac{\text{Total Documents}}{\text{Number of documents term } t \text{ is present in}}\right)$$

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	32	45	48	26	7	3	0	...]
Apple article 2	[0	4	3	7	8	6	3	0	...]
Tesla article 3	[15	31	44	43	25	0	0	0	...]
Tesla article 4	[3	0	0	0	0	0	0	1	...]

that → 4/3 -> 1.33 gigafactory → 4/1 → 4 iphone → 4/2 → 2

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

Apple

article 1

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
$Term\ count$	[0	32	45	48	26	7	3	0	...]
$IDF(t)$	[0	0.12	0.12	0.12	0.12	0.3	0.3	0	...]

that → 4/3 gigafactory → 4/1 iphone → 4/2

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

$$\left(\frac{Total\ Number\ of\ time\ term\ t\ is\ present\ in\ doc\ A}{Total\ number\ of\ tokens\ in\ doc\ A}\right)$$

- $$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

$$TF(t,d) = \left(\frac{Total\ Number\ of\ time\ term\ t\ is\ present\ in\ doc\ A}{Total\ number\ of\ tokens\ in\ doc\ A} \right)$$

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in} \right)$$

$$TF(t,d) = \left(\frac{Total\ Number\ of\ time\ term\ t\ is\ present\ in\ doc\ A}{Total\ number\ of\ tokens\ in\ doc\ A} \right)$$

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in} \right)$$

$$TF - IDF = TF(t,d) * IDF(t)$$

	musk	that	price	market	investor	iphone	itunes	gigafactory	...
Apple article 1	[0	0.05	0.01	0.05	0.05	0.9	0.8	0	...]
Apple article 2	[0	0.002	0.008	0.01	0.02	0.9	0.8	0	...]
Tesla article 3	[0.99	0.05	0.01	0.05	0.05	0	0	0	...]
Tesla article 4	[0.95	0	0	0	0	0	0	0.87	...]

TF – IDF Representation (or Vectorizer)

sklearn

If `smooth_idf=True` (the default), the constant "1" is added to the numerator and denominator of the idf as if an extra document was seen containing every term in the collection exactly once, which prevents zero divisions: idf(t) = log [(1 + n) / (1 + df(t))] + 1.

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

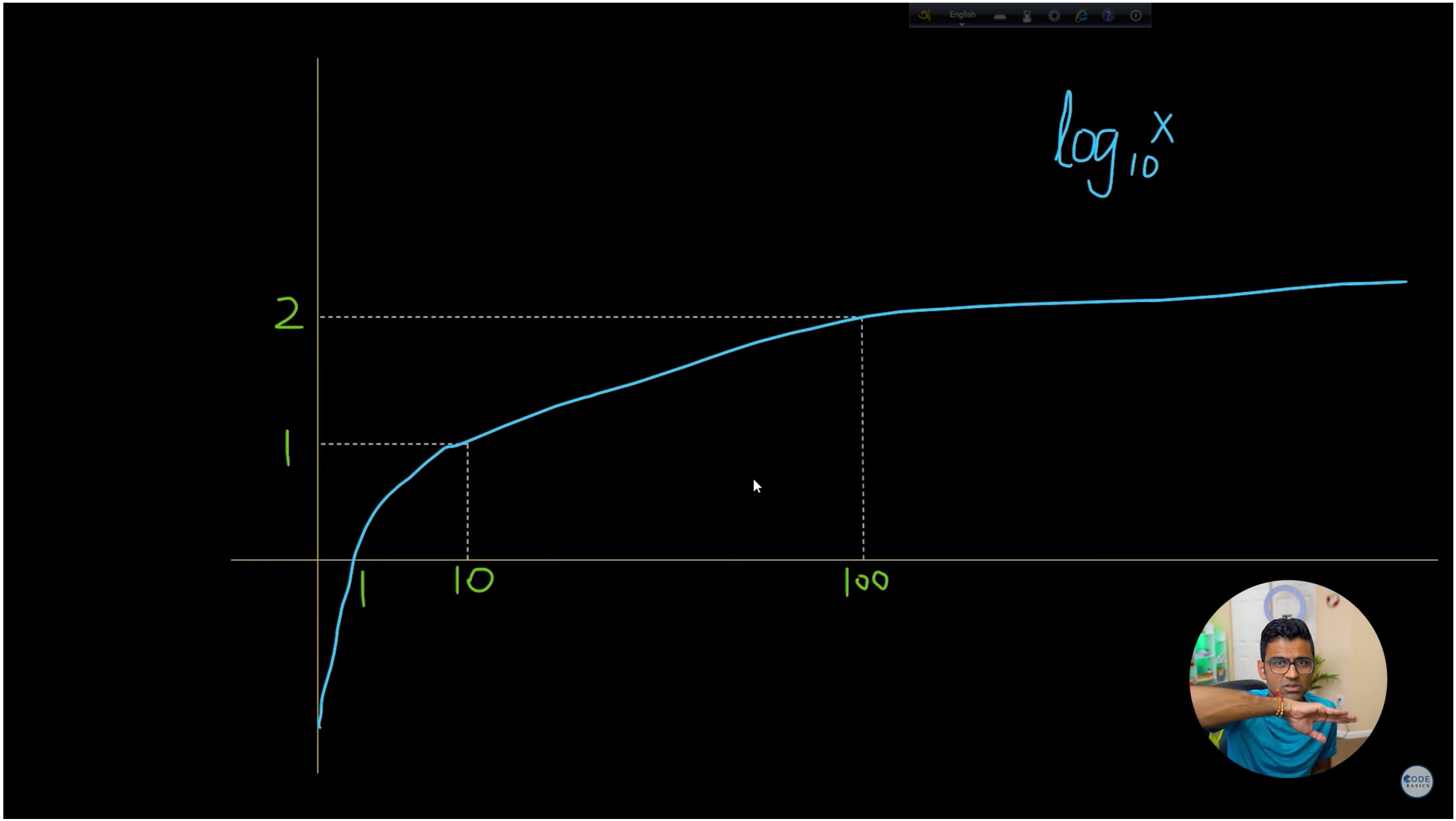
Why do we use log in IDF

$$IDF(t) = log\left(\frac{Total\ Documents}{Number\ of\ documents\ term\ t\ is\ present\ in}\right)$$

Here is the intuition: If term frequency for the word 'computer' in doc1 is 10 and in doc2 it's 20, we can say that doc2 is more relevant than doc1 for the word 'computer.

However, if the term frequency of the same word, 'computer', for doc1 is 1 million and doc2 is 2 millions, at this point, there is no much difference in terms of relevancy anymore because they both contain a very high count for term 'computer'.

Just like Debasis's answer, adding log is to dampen the importance of term that has a high frequency, e.g. Using log base 2, the count of 1 million will be reduced to 19.9!


English
$\log_{10} x$
2
1
1
10
100

Limitations of tf-idf model

Limitations of tf-idf model

As n increased, dimensionality, sparsity increases

Limitations of tf-idf model

As n increased, dimensionality, sparsity increases

Doesn't capture relationship between words

Limitations of tf-idf model

As n increased, dimensionality, sparsity increases

Doesn't capture relationship between words

Doesn't address out of vocabulary (OOV) problem