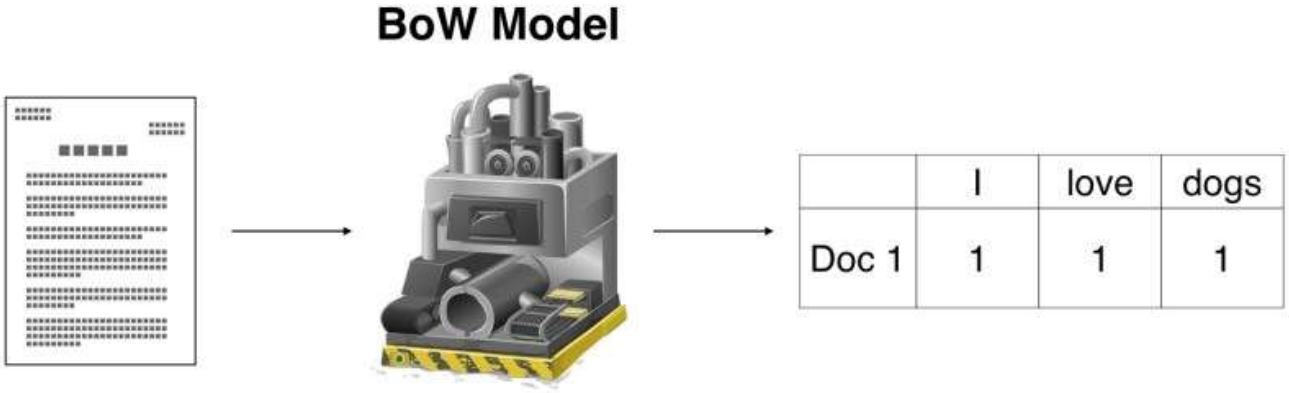
Bag-of-words Model

You can think of the bag-of-words (BoW) model as a machine which takes as input a set of documents and outputs a table containing the frequency counts of each word in each document.
In the simplest case, let's suppose that our set of documents consists of only one document – a sentence:

I love dogs.

Let's apply our BoW model to this document.



	I	love	dogs
Doc 1	1	1	1

The BoW outputs a table wherein each row corresponds to a document and each column represents a unique word. The entries are the counts of each word in each document.

We can't really derive any insight from the table, since we're only considering one document. Each entry in the table is 1, since the first document obviously contains all unique words in the first document.

The BoW model was created as a means to compare a set of documents. There really is no point if you only have a single document.

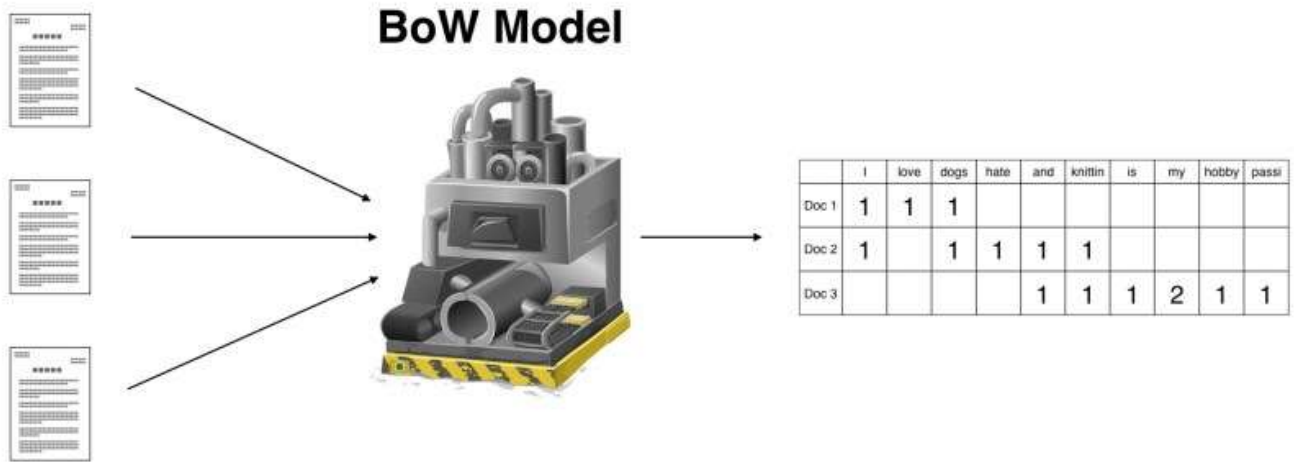
Suppose our set of documents consists of three sentences:

1. I love dogs.

2. I hate dogs and knitting.

3. Knitting is my hobby and my passion.

Applying the BoW to this set, we get a table with three rows.



Let's zoom in on the table.

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

OK. This is a lot more interesting.

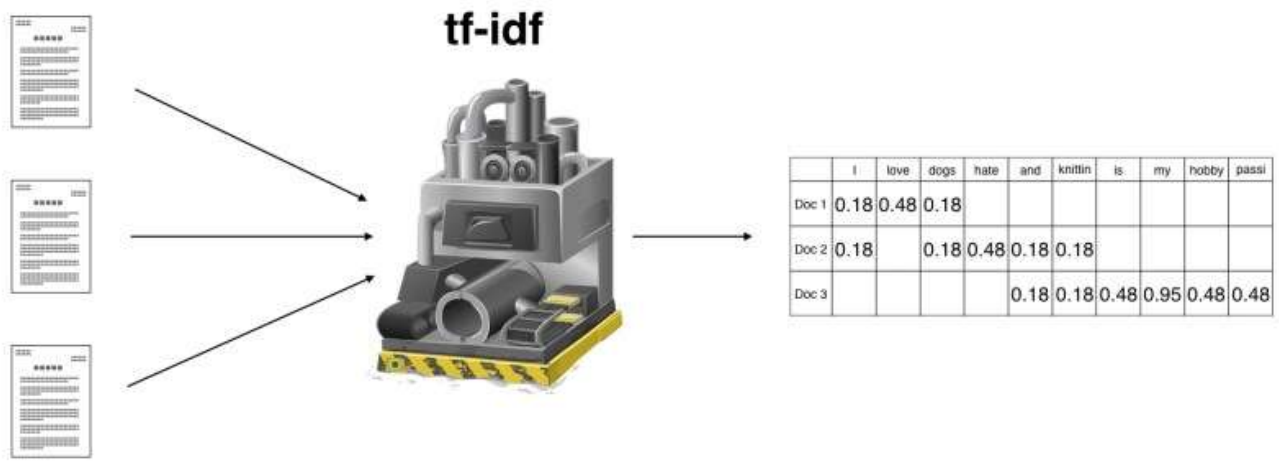
BoW only takes into consideration the frequency of words in a document. A document's important words, in the BoW-sense, are words that occur a lot of times in said document.

We can see in the table that the word **love** appears only in the first document, **hate** only appears in the second, and **hobby** and **passion** only

appear in the third. Using the BoW model, we can identify the important – that is, signature – words in the different documents by visual inspection. The question now is, can we actually **quantify** the signature words? Well, yes we can. By applying another transformation.

tf-idf

The BoW model is a perfectly acceptable model to convert raw text to numbers. However, if our purpose is to identify signature words in a document, there is a better transformation that we can apply.



	I	love	dogs	hate	and	knittin	is	my	hobby	passi
Doc 1	0.18	0.48	0.18							
Doc 2	0.18		0.18	0.48	0.18	0.18				
Doc 3					0.18	0.18	0.48	0.95	0.48	0.48

tf-idf is shorthand for term frequency – inverse document frequency. So, two things: **term frequency** and **inverse document frequency**.
Term frequency (tf) is basically the output of the BoW model. For a specific document, it determines how important a word is by looking at how frequently it appears in the document. Term frequency measures the local importance of the word. If a word appears a lot of times, then the word must be important. For example, if our document is "I am a cat lover. I have a cat named Steve. I feed a cat outside my room regularly," we see that the words with the highest frequency are **I**, **a**, and **cat**. This agrees with our intuition that high term frequency = higher importance since the document is all about my fascination with cats.
The second component of tf-idf is inverse document frequency (idf). For a word to be considered a signature word of a document, it shouldn't appear that often in the other documents. Thus, a signature word's document frequency must be low, meaning its inverse document frequency must be high. The idf is usually calculated as

$$\text{idf(W)} = \log \frac{\#(\text{documents})}{\#(\text{documents containing word W})}.$$

The tf-idf is the product of these two frequencies. For a word to have high tf-idf in a document, it must appear a lot of times in said document and must be absent in the other documents. It must be a signature word of the document.

Let's zoom in on the output of our three-sentence example.

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	**0.48**	0.18							
Doc 2	0.18		0.18	**0.48**	0.18	0.18				
Doc 3					0.18	0.18	**0.48**	**0.95**	**0.48**	**0.48**

We can see here that tf-idf highlights **love** as a signature word in the first document, **hate** in the second, and **is**, **my**, **hobby** and **passion** in the third.