Word Embeddings (TFIDF)

Tim Metzler
Department of Computer Science
HBRS / ST.A.

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Why?
Make embeddings that make it easy to distinguish between documents.
Used in information retrieval (find relevant documents to a search query)

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Two parts:
- tf → How often does a term appear in a document

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Two parts:
- tf → How often does a term appear in a document
- df → In how many documents does a term appear

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Two parts:
- tf → How often does a term appear in a document
- df → In how many documents does a term appear
- idf → Inverse of df. idf = N/df

Term Frequency – Inverse Document Frequency (TFIDF)

Idea:
Take a corpus (collection of documents), count occurences of terms within documents.
Normalize over documents.

Two parts:
- tf → How often does a term appear in a document
- df → In how many documents does a term appear
- idf → Inverse of df. idf = N/df

Intuition:
A term is important if it appears often in a document.
A term is important if it only appears in a few documents.

Term Frequency – Inverse Document Frequency (TFIDF)

Example corpus:

```
corpus = [
    "A dog is an animal. A dog is not a cat.",
    "A cat is an animal.",
    "My dog is playful.",
    "I like animals. Linux is not an animal.",
    "Cat is a linux command. Dog is not.",
    "Dog is not a linux command.",
    "My dog likes linux. my dog is playful"
]
```

Term Frequency – Inverse Document Frequency (TFIDF)

Count occurences of words (terms) in each document (term frequency tf)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1

Term Frequency – Inverse Document Frequency (TFIDF)

Count in how many documents a term appears (document frequency df)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Document Frequency	**4**	**3**	**3**	**1**	**3**	**2**	**5**	**1**	**6**	**1**	**1**	**4**	**2**	**4**	**2**

Term Frequency – Inverse Document Frequency (TFIDF)

Take inverse of document frequency by dividing number of documents by df (idf)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Inverse Document Frequency	**6/4**	**6/3**	**6/3**	**6/1**	**6/3**	**6/2**	**6/5**	**6/1**	**6/6**	**6/1**	**6/1**	**6/4**	**6/2**	**6/4**	**6/2**

Term Frequency – Inverse Document Frequency (TFIDF)

Take inverse of document frequency by dividing number of documents by df (idf)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Inverse Document Frequency	**1.5**	**2**	**2**	**6**	**2**	**3**	**1.2**	**6**	**1**	**6**	**6**	**1.5**	**3**	**1.5**	**3**

Term Frequency – Inverse Document Frequency (TFIDF)

Normalize idf with logarithm (log10)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Inverse Document Frequency (log)	**0.18**	**0.3**	**0.3**	**0.78**	**0.3**	**0.48**	**0.08**	**0.78**	**0**	**0.78**	**0.78**	**0.18**	**0.48**	**0.18**	**0.48**

Term Frequency – Inverse Document Frequency (TFIDF)

Normalize idf with logarithm (log10)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
I like animals. Linux is not an animal.	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0
Cat is a linux command. Dog is not.	1	0	0	0	1	1	1	0	2	0	0	1	0	1	0
Dog is not a linux command.	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Inverse Document Frequency (log)	**0.18**	**0.3**	**0.3**	**0.78**	**0.3**	**0.48**	**0.08**	**0.78**	**0**	**0.78**	**0.78**	**0.18**	**0.48**	**0.18**	**0.48**

Term Frequency – Inverse Document Frequency (TFIDF)

Normalize idf with logarithm (log10)

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	3	1	1	0	1	0	2	0	2	0	0	0	0	1	0
A cat is an animal.	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0
My dog is playful.	0	0	0											0	1
I like animals. Linux is not an animal.	0	1	1											1	0
Cat is a linux command. Dog is not.	1	0	0											1	0
Dog is not a linux command.	1	0	0	0	0	1	1		1	0	0	1	0	1	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	2	0	1	0	1	1	2	0	1
Inverse Document Frequency (log)	**0.18**	**0.3**	**0.3**	**0.78**	**0.3**	**0.48**	**0.08**	**0.78**	**0**	**0.78**	**0.78**	**0.18**	**0.48**	**0.18**	**0.48**

idf of 0 → can not be used to distinguish documents.
Appears in all of them!

Term Frequency – Inverse Document Frequency (TFIDF)

Multiply tf with idf

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	.54	.30	.30	0	.30	0	.16	0	0	0	0	0	0	.18	0
A cat is an animal.	.18	.30	.30	0	.30	0	0	0	0	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	.08	0	0	0	0	0	.48	0	.48
I like animals. Linux is not an animal.	0	.30	.30	.78	0	0	0	.78	0	.78	0	.18	0	.18	0
Cat is a linux command. Dog is not.	.18	0	0	0	.30	.48	.08	0	0	0	0	.18	0	.18	0
Dog is not a linux command.	.18	0	0	0	0	.48	.08	0	0	0	0	.18	0	.18	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	.16	0	0	0	.78	.18	.96	0	.48

Term Frequency – Inverse Document Frequency (TFIDF)

Document embeddings with words as dimensions

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	.54	.30	.30	0	.30	0	.16	0	0	0	0	0	0	.18	0
A cat is an animal.	.18	.30	.30	0	.30	0	0	0	0	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	.08	0	0	0	0	0	.48	0	.48
I like animals. Linux is not an animal.	0	.30	.30	.78	0	0	0	.78	0	.78	0	.18	0	.18	0
Cat is a linux command. Dog is not.	.18	0	0	0	.30	.48	.08	0	0	0	0	.18	0	.18	0
Dog is not a linux command.	.18	0	0	0	0	.48	.08	0	0	0	0	.18	0	.18	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	.16	0	0	0	.78	.18	.96	0	.48

Term Frequency – Inverse Document Frequency (TFIDF)

Word embeddings with documents as dimensions

	a	an	animal	animals	cat	command	dog	i	is	like	likes	linux	my	not	playful
A dog is an animal. A dog is not a cat.	.54	.30	.30	0	.30	0	.16	0	0	0	0	0	0	.18	0
A cat is an animal.	.18	.30	.30	0	.30	0	0	0	0	0	0	0	0	0	0
My dog is playful.	0	0	0	0	0	0	.08	0	0	0	0	0	.48	0	.48
I like animals. Linux is not an animal.	0	.30	.30	.78	0	0	0	.78	0	.78	0	.18	0	.18	0
Cat is a linux command. Dog is not.	.18	0	0	0	.30	.48	.08	0	0	0	0	.18	0	.18	0
Dog is not a linux command.	.18	0	0	0	0	.48	.08	0	0	0	0	.18	0	.18	0
My dog likes linux. My dog is playful.	0	0	0	0	0	0	.16	0	0	0	.78	.18	.96	0	.48