ecnológico
Monterrey

NLP 2

Inteligencia artificial avanzada para la ciencia de datos II Modulo 5 NLP 2

TF/IDF

Until this point

Analyzing the frequency of tokens in the **corpus.**

Help us to present descriptive statistics of the main words or phrases.

BoW constructs a vocabulary from this analysis and new inputs are **vectorized** based on the tokens frequency inside the input.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

What's the problem?

By intuition frequency is related to higher probability.

If a token appears many times in the corpus, new phrases in that context probably will include those tokens.

However, **BoW** is not normalized and rare words are most of the time **ignored** to vectorization (new phrases with rare tokens produces the same vectors).

Example:
A medical vocabulary from a corpus produces the following dimensions:

Token	Frequency
Infection	200
Blood	150
analysis	100
cardiopulmonary	5
resuscitations	2

Someone search:
“Cardiopulmonary infections”

What is the resulting vector?

Document relevance

Suppose we will process medical books in a corpus, so we might be able to compare vectors for retrieve (search) problem, there are two medical documents to process.

A medical research for **infections**:

A study for common **infections, cardiopulmonar infections,**general **infection** protocols and **infections** in surgery.

Pulmons and hearth illness

A complete **cardiopulmonar** guide **for infections**

First document (blue) is not so **relevant** for **“cardiopulmonar infections”**, second document is very relevant, but there is a problem...

Document vectors

A medical research for **infections**:

A study for common **infections, cardiopulmonar infections,**general **infection** protocols and **infections** in surgery.

Pulmons and hearth illness

A complete **cardiopulmonar** guide **for infections**

5	0	0	0	1

1	0	0	0	1

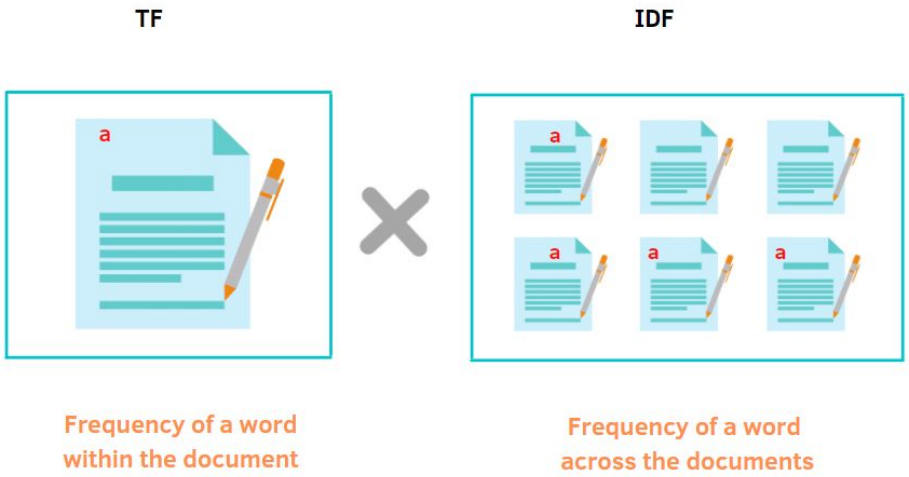
Following **Bow** vectors, first document actually is more relevant than second...

TF/IDF is the answer

TF/IDF statistical model, gives a statistical normalized encoding process, so **rare** and not so frequent **words** have more relevance/meaning than **general** words.

This is important for retrieving problems and also ML models.

Rare but important (differentiators) features gives information for classification problems.



Term Frequency (TF)

TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

$$TF = \frac{Number\ of\ times\ a\ word\ "X"\ appears\ in\ a\ Document}{Number\ of\ words\ present\ in\ a\ Document}$$

For example

A medical research for **infections**:

A study for common **infections, cardiopulmonar infections,**general **infection** protocols and **infections** in surgery.

Infections = 5/13
Cardiopulmonar = 1/13

pulmons and hearth illness

A complete **cardiopulmonar** guide **for infections**

Infections = 1/7
Cardiopulmonar = 1/7

Inverse Dense Frequency(IDF)

IDF represents how **"special"** (not so common) a word is, it divides the number of documents and the number of documents that includes that word. Finally a normalization function is apply.

$$IDF = log\left(\frac{Number\ of\ Documents\ present\ in\ a\ Corpus}{Number\ of\ Documents\ where\ word\ "X"\ has\ appeared}\right)$$

Example:

Documents in the Corpus= 3500

Token	Frequency	Documents with this word	IDF
Infection	200	120	log(3500/120)
Blood	150	80	log(3500/80)
analysis	100	95	log(3500/95)
cardiopulmonary	5	1	log(3500/1)
resuscitations	2	1	log(3500/1)

TF/IDF

Finally by multiplying both metrics we are able to produce a more significant vectors.

$$TF\ IDF = TF * IDF$$

Infections = (5/13)*log(3500/120)

Cardiopulmonar = (1/13)*log(3500/1)

A medical research for **infections**:

A study for common **infections, cardiopulmonar infections,**general **infection** protocols and **infections** in surgery.

0.56	0	0	0	0.272

Infections = (1/7)*log(3500/120)

Cardiopulmonar = (1/7)*log(3500/1)

Pulmons and hearth illness

A complete **cardiopulmonar** guide **for infections**

0.209	0	0	0	0.50

Thanks

Do you have any questions?

emmanuel.paez@tec.mx
Slack #module-5-nlp-1