Text Analytics

24 April 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today(in this session)

- Sentiment Analysis
- Converting text to numbers
- DTM and TF-IDF matrix
- Some more text cleaning exercises and examples
- Cosine Similarity

TF-IDF Motivation

Consider the following 3 sentences -

- S1 = Text Analytics is boring boring boring
- S2 = Analytics is interesting
- S3 = We want interesting sports analytics

TF-IDF Motivation

Consider the following 3 sentences -

- S1 = Text Analytics is boring boring boring
- S2 = Analytics is interesting
- S3 = We want interesting sports analytics

We can choose to remove the stopwords, convert everything to lowercase and construct the following matrix. We call this DTM or Document Term Matrix.

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

TF-IDF Motivation

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

TF-IDF Motivation

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1

We see that analytics and sports is getting the same weightage in S-3, whereas "sports" is exclusive to S-3, "analytics" can be found in all sentences.

In TF-IDF Matrix, we increase the weightage of the words that are exclusive to a document/sentence and decrease the weightage of the words that are common to many sentences.

TF-IDF Motivation

DF = Document Frequency(computed for each term),
IDF = Inverse Document Frequency(computed for each term),
TF = Term Frequency (essentially the DTM matrix),
n = number of documents

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1

TF-IDF Motivation

DF = Document Frequency(computed for each term),
IDF = Inverse Document Frequency(computed for each term),
TF = Term Frequency (essentially the DTM matrix),
n = number of documents

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1

As the name suggests, we would multiply the elements in TF with the corresponding IDF.
Several methods have been proposed in literature for the formula of IDF, one of the common ones is -

$$\text{IDF} = 1 + \ln\left(\frac{1+n}{1+DF}\right)$$

TF-IDF

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1
IDF	$1+\ln(1)$	$1+\ln(2)$	$1+\ln(4/3)$	$1+\ln(2)$	$1+\ln(2)$	$1+\ln(2)$

TF-IDF

	analytics	boring	interesting	sports	text	want
S-1	1	3	0	0	1	0
S-2	1	0	1	0	0	0
S-3	1	0	1	1	0	1
DF	3	1	2	1	1	1
IDF	1+ln(1)	1+ln(2)	1+ln(4/3)	1+ln(2)	1+ln(2)	1+ln(2)

We then multiply the TFs with the corresponding IDFs to get-

	analytics	boring	interesting	sports	text	want
S-1	1*1	3*1.693	0*1.287	0*1.693	1*1.693	0*1.693
S-2	1*1	0*1.693	1*1.287	0*1.693	0*1.693	0*1.693
S-3	1*1	0*1.693	1*1.287	1*1.693	0*1.693	1*1.693

Finally, we convert every row vector to a unit vector.

TF-IDF Matrix finally

From the previous slide,

	analytics	boring	interesting	sports	text	want
S-1	1*1	3*1.693	0*1.287	0*1.693	1*1.693	0*1.693
S-2	1*1	0*1.693	1*1.287	0*1.693	0*1.693	0*1.693
S-3	1*1	0*1.693	1*1.287	1*1.693	0*1.693	1*1.693

After normalization of each row,

TF-IDF Matrix						
	analytics	boring	interesting	sports	text	want
S-1	0.1836	0.9326	0	0	0.3109	0
S-2	0.6134	0	0.7898	0	0	0
S-3	0.3452	0	0.4445	0.5845	0	0.5845

Sentiment Analysis - VADER

https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399

Cosine Similarity

$\vec{a}, \vec{b}$

$$\boxed{\cos\theta = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\cdot||\vec{b}||}}$$

How do we measure similarity of two documents?

S1 = "Phone is good, phone is good"
S2 = "Phone is not good"
S3 = "It is a good phone"

Ph	good	not
2	2	0
1	1	1
1	1	0

(Removing the stopwords except "not")
Euclidean distance from DTM would suggest that distance of S3 and S1 is $\sqrt{2}$, and distance of S3 and S~~1~~2 is 1.

Cosine Similarity

Thank you for your attention