AIOutsider

POWERED BY

Section 3

Text Representation

Want more? Head over to www.AIOutsider.com

Section Overview

- Text Mining & NLP
- Sentiment Analysis
- Google Colab
- Dataset Overview
- Descriptive Statistics

- Text Normalization
- Features Cleaning
- Tokenization
- Stemming
- Lemmatization

- Text Representation
- Negative/Positive
- Bag-of-Words
- TF-IDF

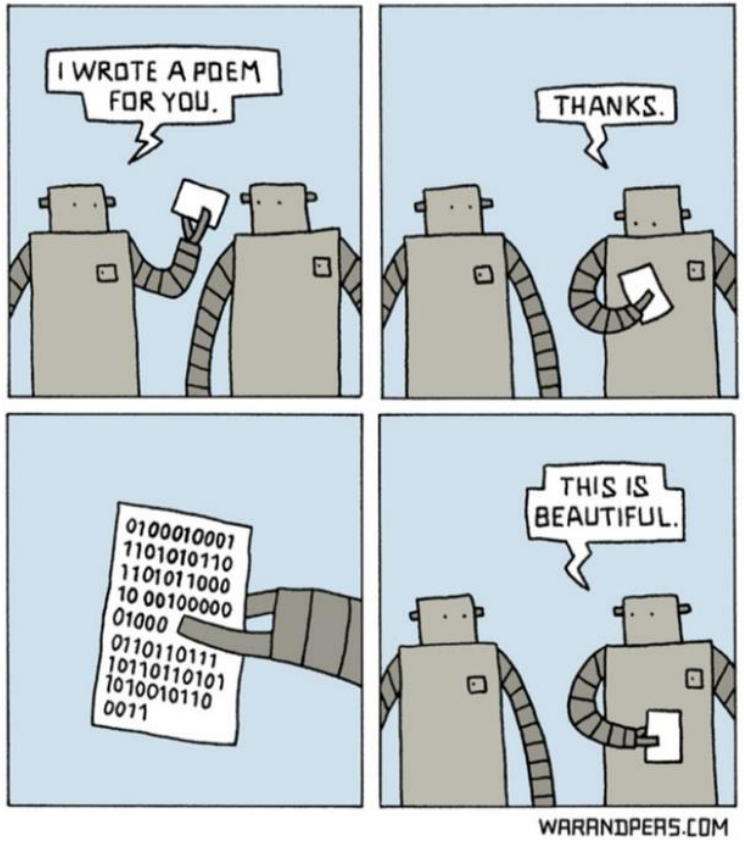
AIOutsider.com

WHY REPRESENTING?

WHY REPRESENTING TEXT?

Definition

Representing text numerically allows it to be understood by ML models



Aa

- Positive/Negative Frequencies
- Bag of Words
- TF-IDF

10101
01011
10101

AIOutsider.com

POSITIVE/NEGATIVE

POSITIVE AND NEGATIVE FREQUENCIES

TWEET + « I am glad I got hired »

TWEET + « This is great »

TWEET - « This is bad »

TWEET - « I am sad I got fired »

	Freq(w, 1)	Freq(w, 0)
I	2 ←	2 ←
am	1	1
glad	1	0
got	1 ←	1 ←
hired	1	0
fired	0 ←	1 ←
this	1 ←	1 ←
is	1 ←	1 ←
great	1	0
bad	0 ←	1 ←

NEW TWEET « I got fired, this is bad »

$$X = \left[\text{Pos Freq} \ , \ \text{Neg Freq} \right]$$

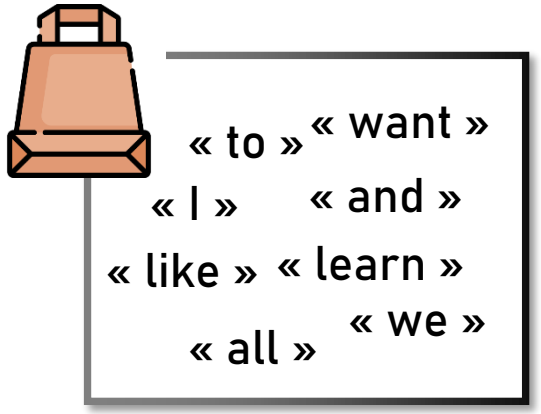
$$X = \left[\Sigma \text{Freq}(w, 1) \ , \ \Sigma \text{Freq}(w, 0) \right]$$

$$X = \left[5 \ , \ 7 \right]$$

AIOutsider.com

BAG OF WORDS

BAG OF WORDS REPRESENTATION



TWEET 1 « I like to learn »

TWEET 2 « We all like and all want to learn »

words	I	all	we	like	and	want	learn	to
features	1	2	3	4	5	6	7	8
VECTOR 1	1	0	0	1	0	0	1	1
VECTOR 2	0	2	1	1	1	1	1	1

MATRIX DIMENSION = (# of tweets, # unique words in corpus)

AIOutsider.com

TF-IDF

TF-IDF

TF « TERM FREQUENCY » — **IDF** « INVERSE DOCUMENT FREQUENCY »

$$tf_{w,d} = \frac{n_{w,d}}{\sum_k n_{w,d}}$$

$$idf_w = \log\left(\frac{N}{df_w}\right)$$

TWEET 1
« I like my cat »

TWEET 2
« I love my dog »

features	TF d_1	TF d_2	IDF	TF-IDF d_1	TF-IDF d_2
I	w_1 1/4	1/4	w_1 Log(2/2) = 0	0	0
like	w_2 1/4	0	w_2 Log(2/1) = 0.3	0.075	0
love	w_3 0	1/4	w_3 Log(2/1) = 0.3	0	0.075
my	w_4 1/4	1/4	w_4 Log(2/2) = 0	0	0
cat	w_5 1/4	0	w_5 Log(2/1) = 0.3	0.075	0
dog	w_6 0	1/4	w_6 Log(2/1) = 0.3	0	0.075

FOLLOW US!

AIOutsider.com

twitter.com/AIOutsider_

Disclaimer