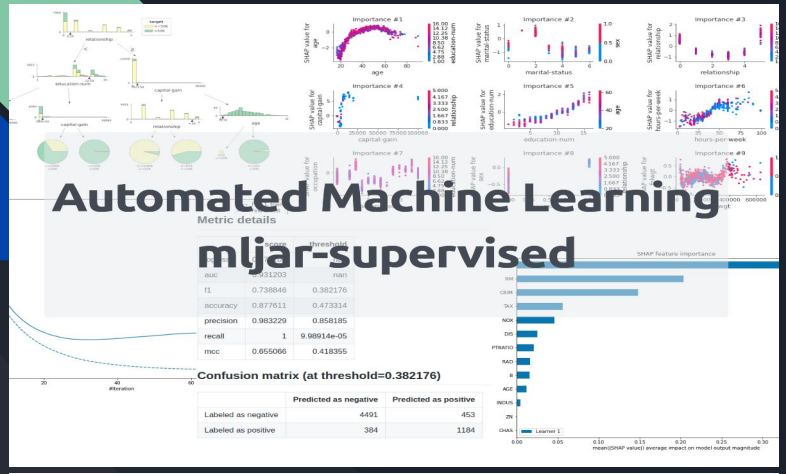
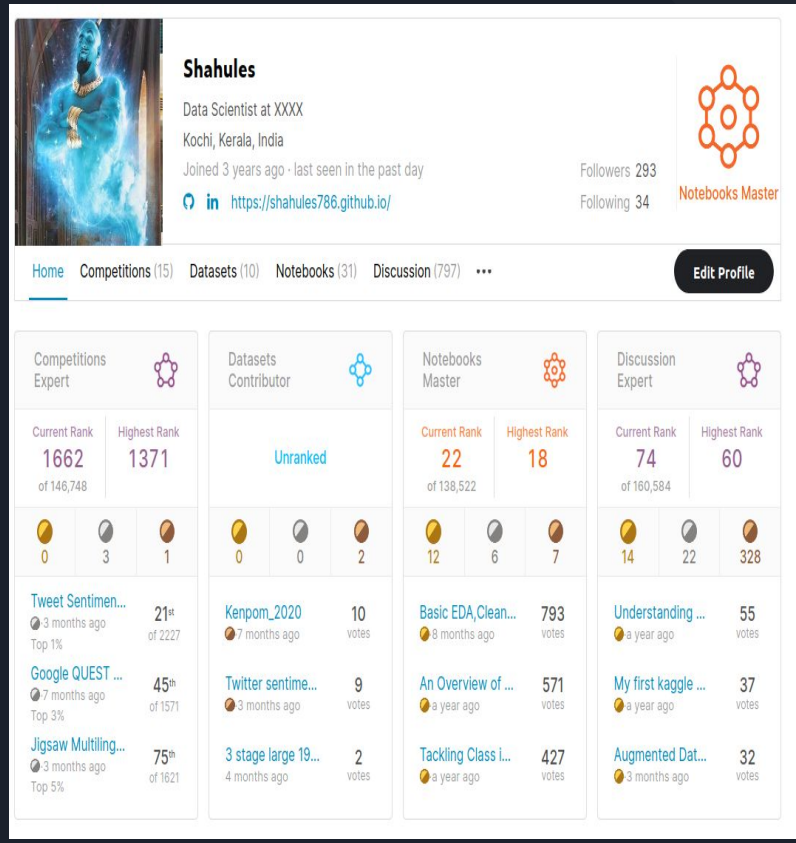
Text Representation in NLP

Sep 13 2020

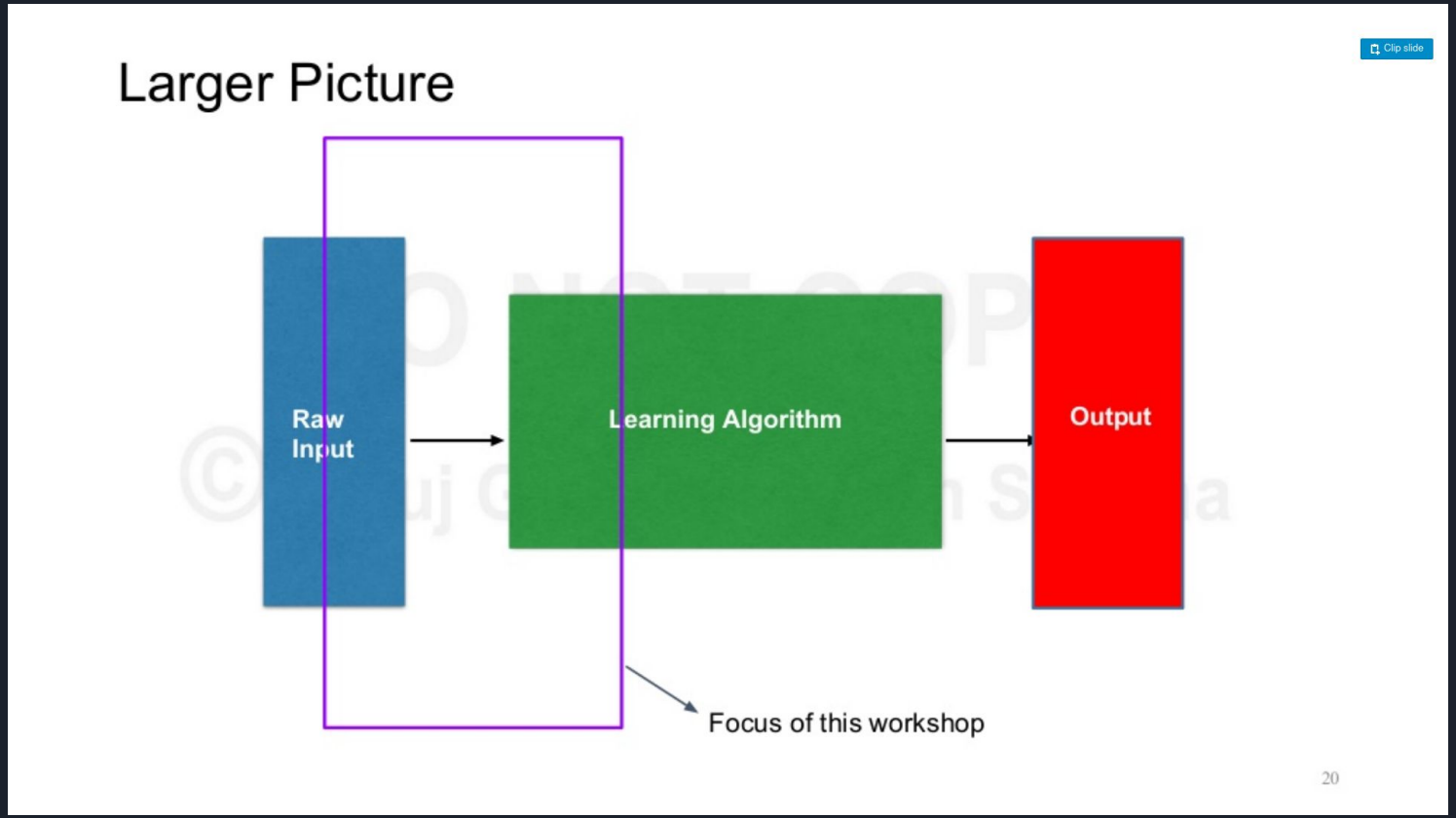
Who am I ?


Automated Machine Learning
mljar-supervised
Metric details
Confusion matrix (at threshold=0.382176)
SHAP feature importance

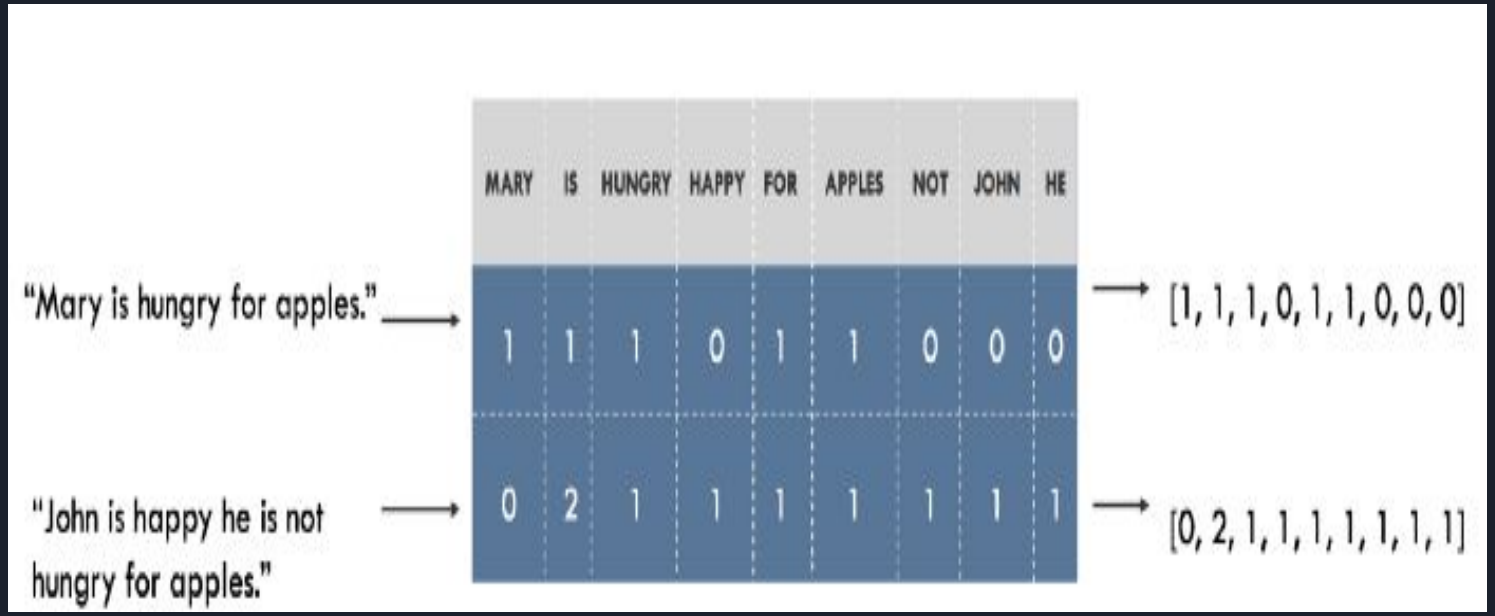

Author: Shahul ES
In-Depth Articles for Machine Learning Practitioners
Hyperparameter Tuning in Python: a Complete Guide 2020
Data Augmentation in NLP: Best Practices From a Kaggle Master
Tabular Data Binary Classification: All Tips and Tricks from 5 Kaggle Competitions
Text Classification: All Tips and Tricks from 5 Kaggle Competitions
Document Classification: 7 Pragmatic Approaches for Small Datasets
Exploratory Data Analysis for Natural Language Processing: A Complete
Topics
Recent posts


Shahules
Data Scientist at XXXX
Kochi, Kerala, India
Joined 3 years ago · last seen in the past day
https://shahules786.github.io/
Followers 293
Following 34
Notebooks Master
Home
Competitions (15)
Datasets (10)
Notebooks (31)
Discussion (797)
Edit Profile
Competitions Expert
Current Rank 1662 of 146,748
Highest Rank 1371
0 3 1
Tweet Sentimen... 3 months ago Top 1% 21st of 2227
Google QUEST ... 7 months ago Top 3% 45th of 1571
Jigsaw Multiling... 3 months ago Top 5% 75th of 1621
Datasets Contributor
Unranked
0 0 2
Kenpom_2020 7 months ago 10 votes
Twitter sentime... 3 months ago 9 votes
3 stage large 19... 4 months ago 2 votes
Notebooks Master
Current Rank 22 of 138,522
Highest Rank 18
12 6 7
Basic EDA,Clean... 8 months ago 793 votes
An Overview of ... a year ago 571 votes
Tackling Class i... a year ago 427 votes
Discussion Expert
Current Rank 74 of 160,584
Highest Rank 60
14 22 328
Understanding ... a year ago 55 votes
My first kaggle ... a year ago 37 votes
Augmented Dat... 3 months ago 32 votes

NLP Workflow



Bag of words


MARY IS HUNGRY HAPPY FOR APPLES NOT JOHN HE
"Mary is hungry for apples."
1 1 1 0 1 1 0 0 0
[1, 1, 1, 0, 1, 1, 0, 0, 0]
"John is happy he is not hungry for apples."
0 2 1 1 1 1 1 1 1
[0, 2, 1, 1, 1, 1, 1, 1, 1]

What’s the main issue?

- The words such as ‘is’, ‘the’, etc are called stopwords in Natural language processing.
- These words get more importance in bag of words model .

So, what next?

Inverse Document Frequency (IDF)

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

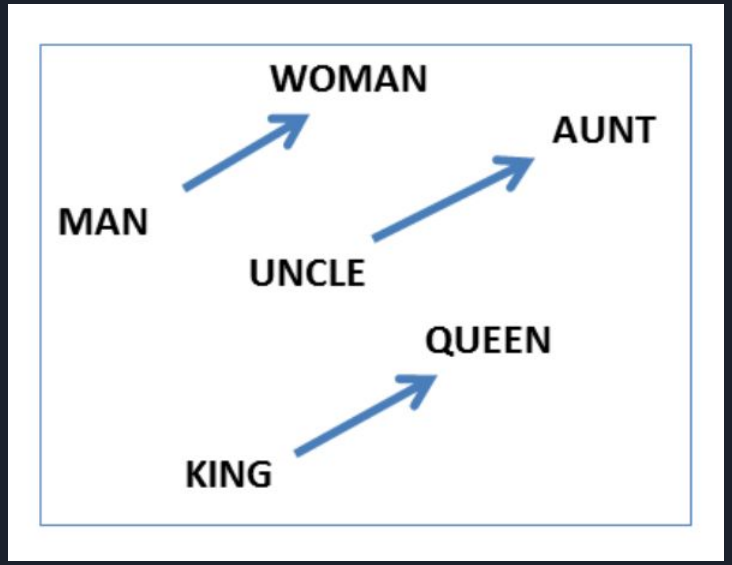
Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

- IDF('movie',) = log(3/3) = 0
- IDF('is') = log(3/3) = 0
- IDF('not') = log(3/1) = log(3) = 0.48
- IDF('scary') = log(3/2) = 0.18
- IDF('and') = log(3/3) = 0
- IDF('slow') = log(3/1) = 0.48

So, What’s the issue?

- The size of matrix increases as the size of vocabulary increases.
- It does not capture position in text,semantics,co-occurrence in sentences.
- Out of vocabulary words, what happens when a new word occur during test?

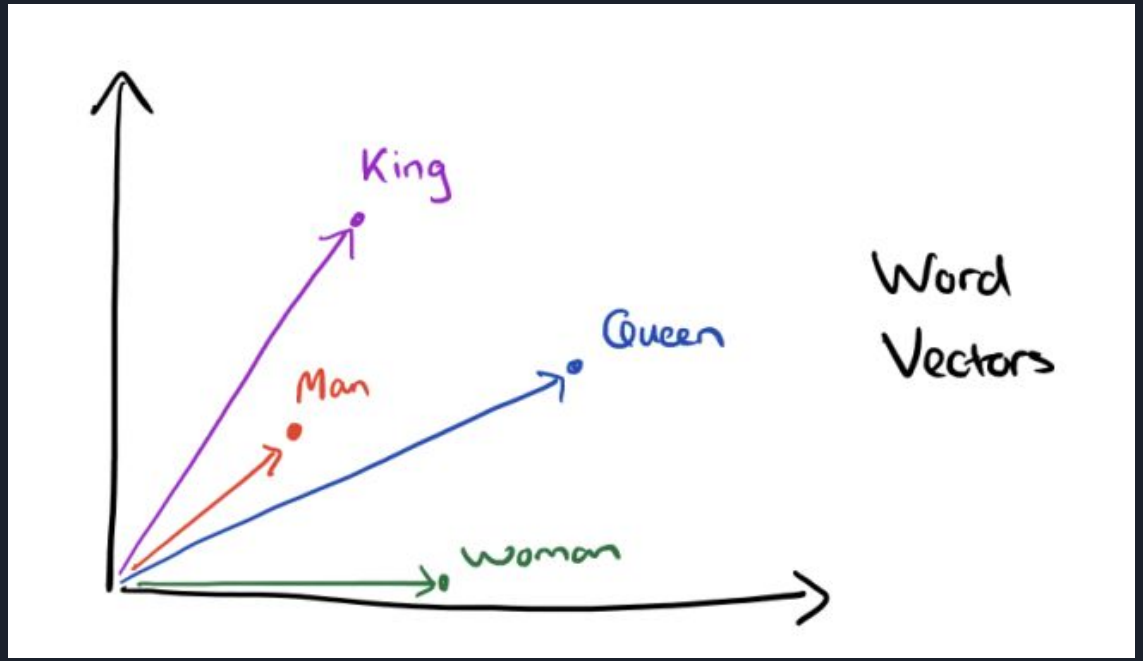
Word embeddings



- A word embedding is a learned representation for text where words that have the same meaning have a similar representation.
- It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems.
- Here each word is represented using a n dimensional vector.

Pre trained word embeddings

- Word2vec
- Glove
- Fastext



So,What's the issue again?

- Word embeddings are non contextual embeddings, meaning it does not take context in which word is used to create the representation.
- Out of vocabulary embeddings.

For example,

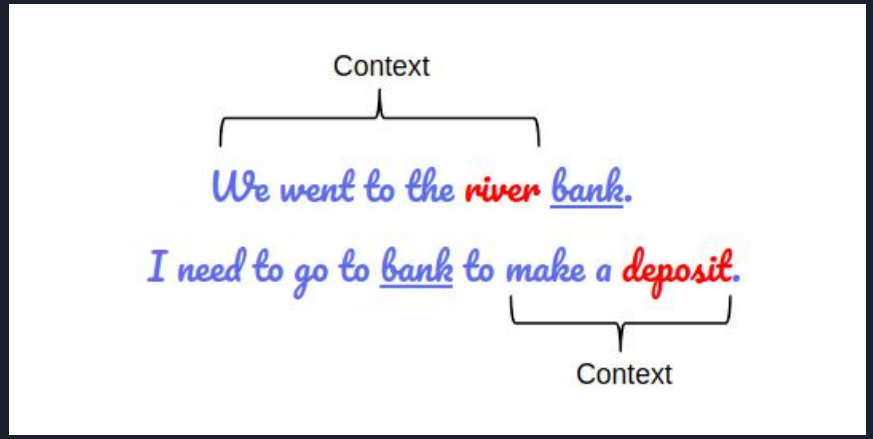
- I went to the bank for an enquiry.
- I was in the river bank.

Here the word "bank" is used in different context, but will have same representation using word embeddings.

BERT

- BERT stands for Bidirectional Encoder Representations from Transformers.
- It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context.
- As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks.

So,what's the final result?



We have very different representation for these two sentences which is finally what we wanted to have.