Breakaway Session 4

Africa Data Hub

Contents

Africa Data Hub

01 - Overview

02 - Machine Learning Summary

03 - Machines + Text

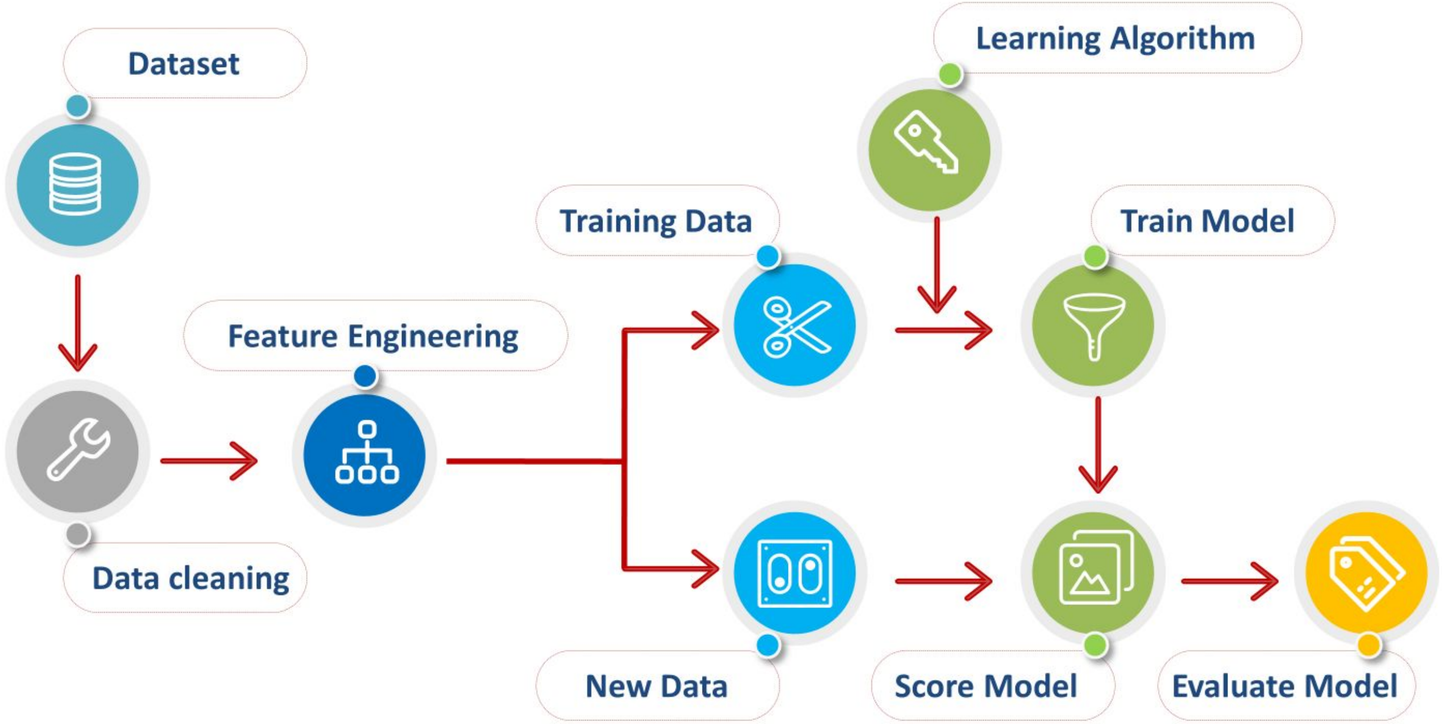
04 - Word Embeddings

05 - Applications

06 - Questions + Practical

Machine Learning Summary

Africa Data Hub



Natural Language Processing

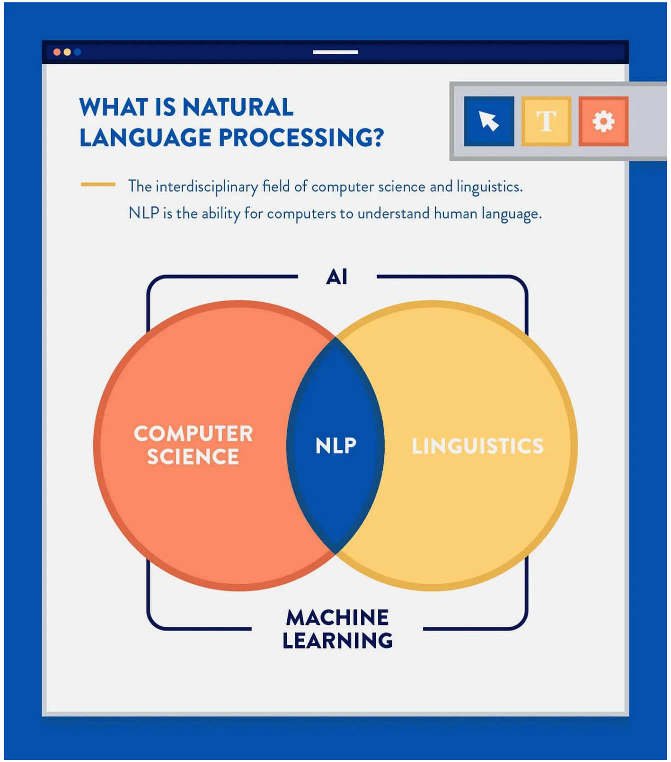
NLP is an interdisciplinary field concerned with the interactions between computers and natural human languages, i.e. speech or text.

It sits at the intersection between linguistics and computer science

Machines do not comprehend language like we do

Algorithmic engagement with language provides exciting opportunities in modern life

There are also important risks and biases that we as practitioners should be aware of



WHAT IS NATURAL LANGUAGE PROCESSING?

The interdisciplinary field of computer science and linguistics.
NLP is the ability for computers to understand human language.

Machines + Text – Building Blocks

Africa Data Hub

First, we need to supply machines with the building blocks of language

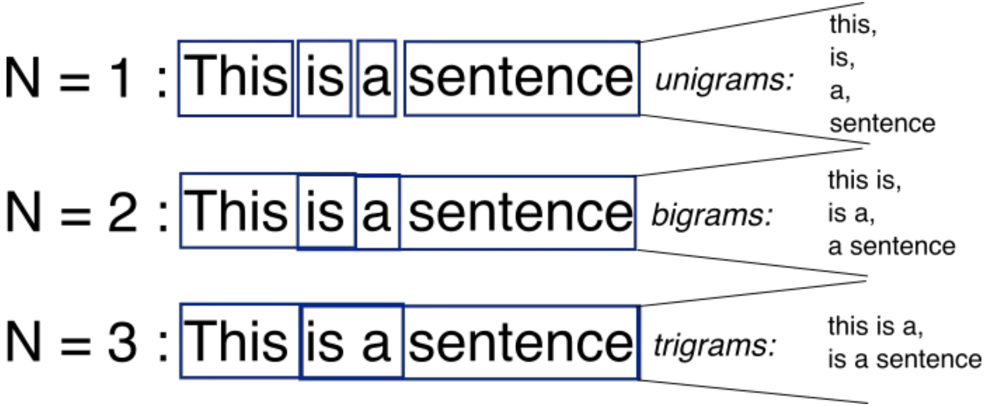
n-grams are contiguous sequences of n items from a given sample of text or speech

This approach can be used to describe many things other than words

n-grams typically are collected from a text or speech corpus

The construction of the corpus will have a powerful influence on things built on it

Longer n-grams capture additional information the machine can utilize in approximating language



Machines + Text - BoW

The bag of words representation is a simple method to represent words, or n-grams

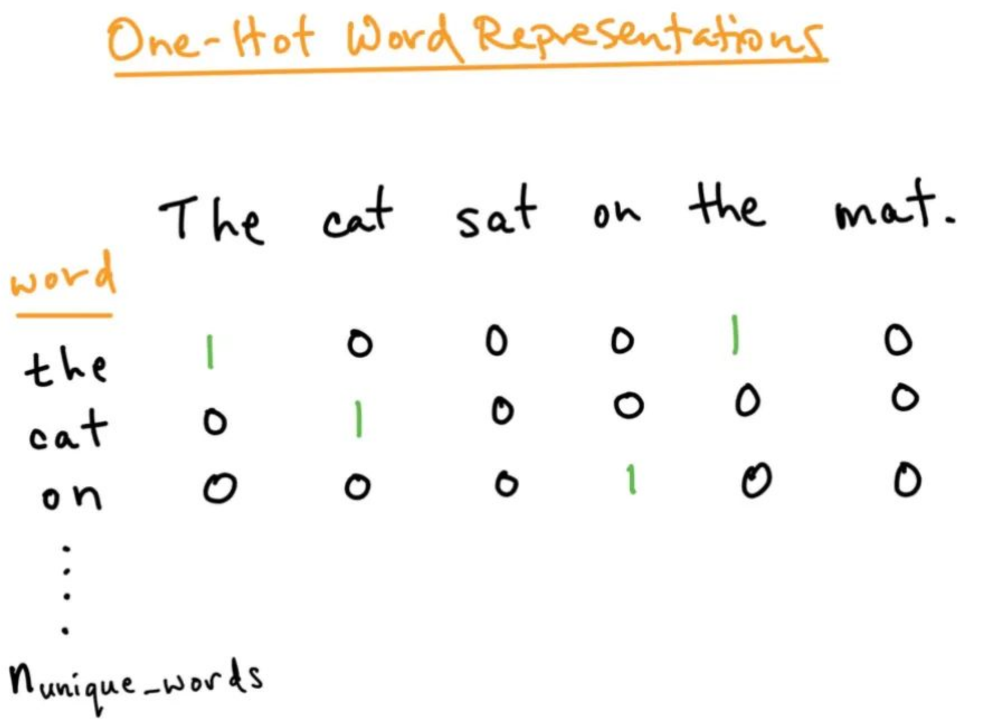
The approach assumes the following:

- Grammar is disregarded
- Word order is disregarded
- Multiplicity is maintained

This method is commonly used for feature generation

It allows for metric derivations and is often used in document classification

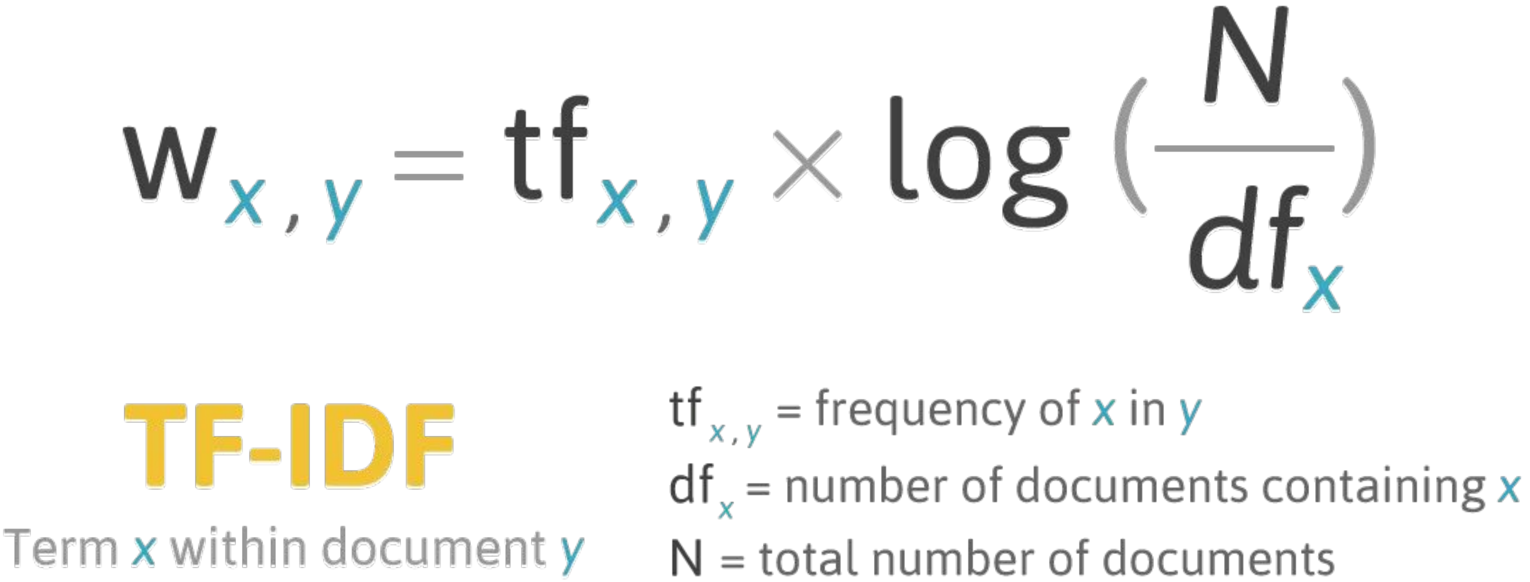
It does not capture nuanced interactions between words



One-Hot Word Representations

word	The	cat	sat	on	the	mat.
the	1	0	0	0	1	0
cat	0	1	0	0	0	0
on	0	0	0	1	0	0
⋮						
n_{unique_words}						

Machines + Text - TF-IDF



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Word Embeddings - Static

Word embeddings seek to represent the meaning of words

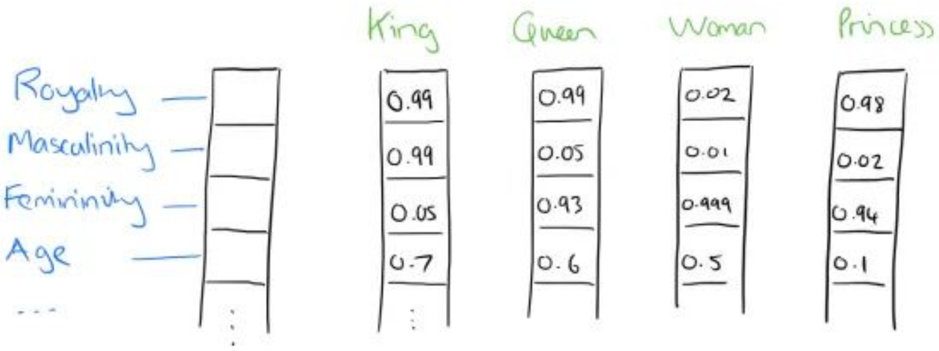
There are two major classes:

- Static embeddings represent a word with a set of words that appear nearby, e.g. **word2vec**

- Contextualized embeddings consider contextual information, e.g. **BERT**

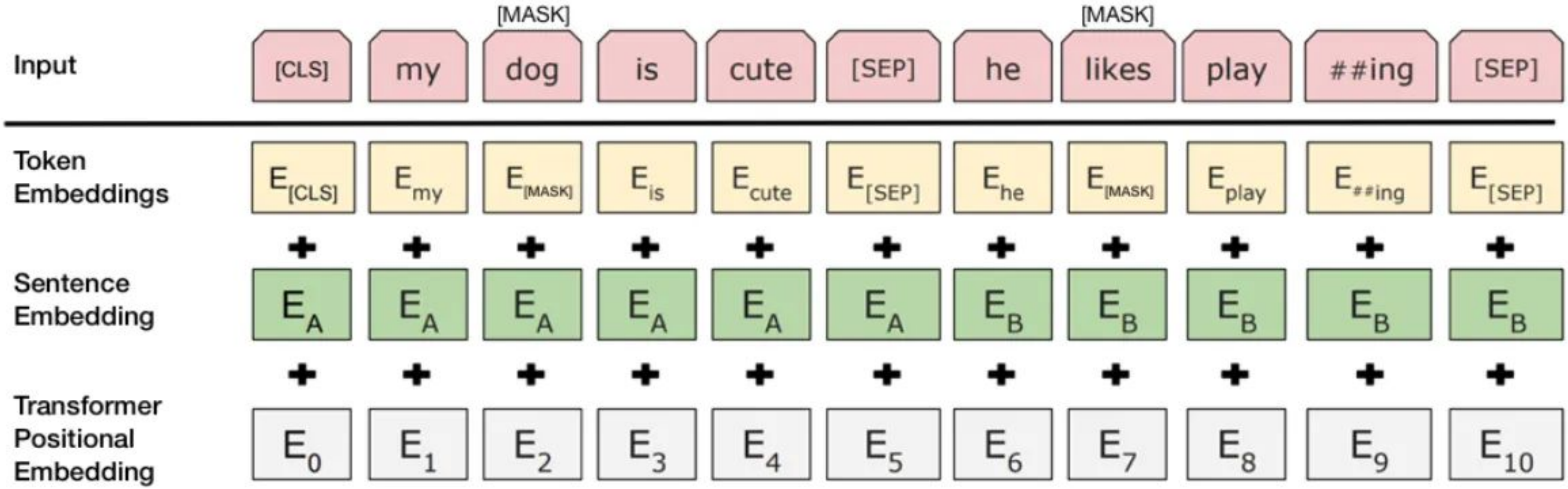
A drawback of word2vec is in dealing with polysemic words, e.g. “**bank**” in the sentence “Tom left the bank and played on the bank of river”

Bias needs to be monitored in associations within the vectors



	King	Queen	Woman	Princess
Royalty	0.99	0.99	0.02	0.98
Masculinity	0.99	0.05	0.01	0.02
Femininity	0.05	0.93	0.999	0.94
Age	0.7	0.6	0.5	0.1
...	⋮	⋮		

Word Embeddings – Contextual



			[MASK]					[MASK]			
Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	$E_{[MASK]}$	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	$E_{[MASK]}$	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Sentence Embedding	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Transformer Positional Embedding	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Applications + Opportunities

Text Classification (e.g. spam detection in Gmail).

Part of Speech (POS) tagging

Named Entity Recognition (NER)

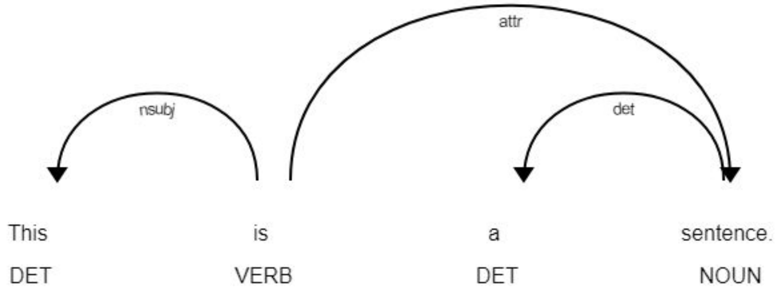
Sentiment Analysis

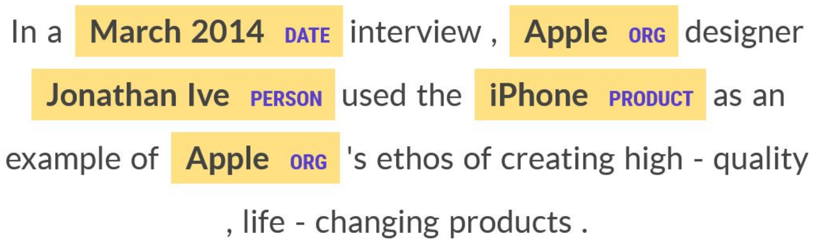
Coreference Resolution

Machine Translation

Question Answering

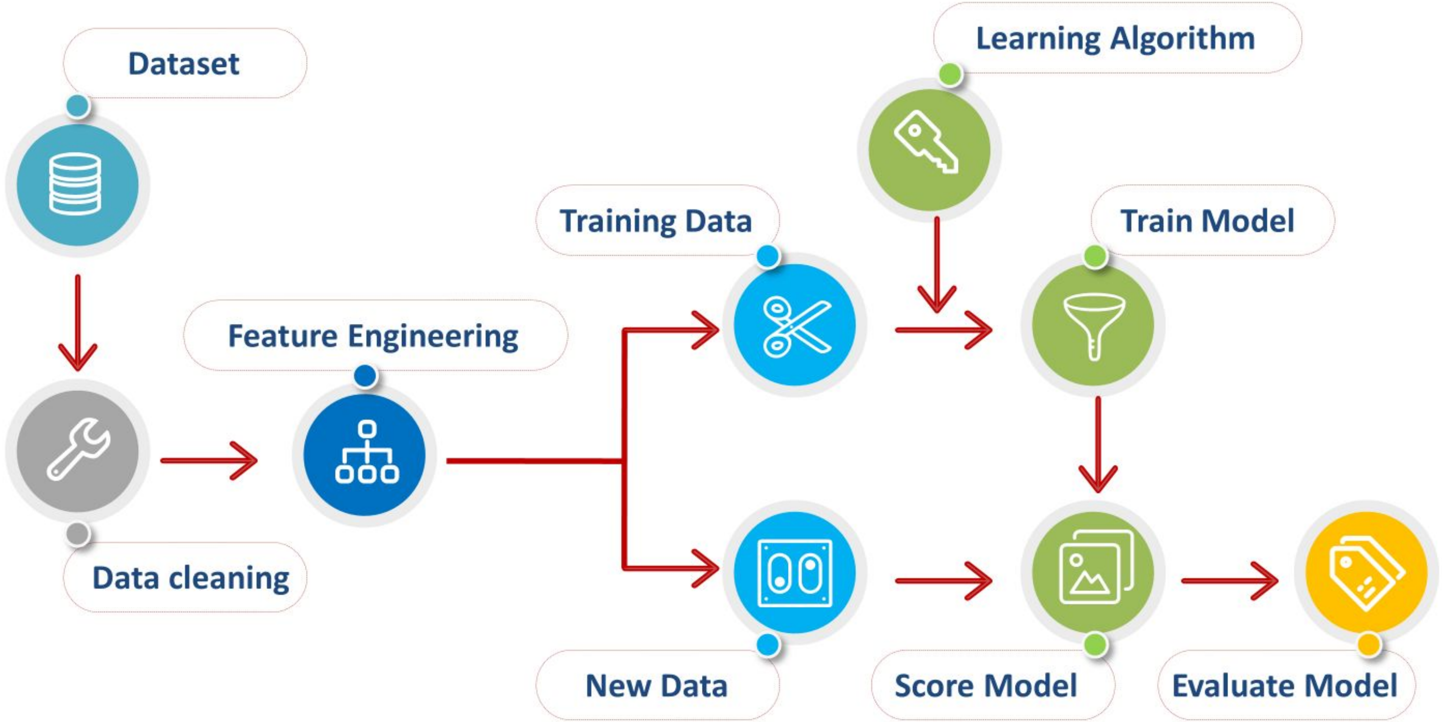
Low-resource language focus




In a **March 2014** DATE interview , **Apple** ORG designer **Jonathan Ive** PERSON used the **iPhone** PRODUCT as an example of **Apple** ORG 's ethos of creating high - quality , life - changing products .

Practical Application

Africa Data Hub



Dataset

Data cleaning

Feature Engineering

Training Data

New Data

Learning Algorithm

Train Model

Score Model

Evaluate Model

THANK YOU!

info@africadatahub.org

Visit https://www.africadatahub.org/

@Africa_DataHub

Africa Data Hub