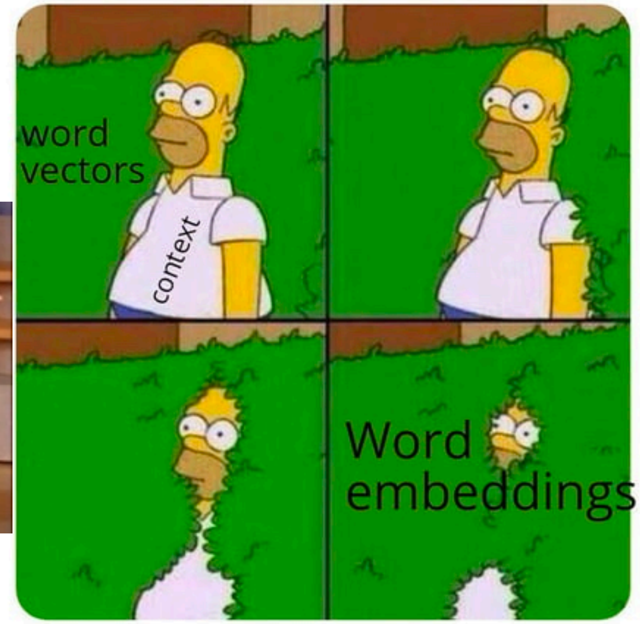
April 17, 2023.

DSML : NLP module.

Introduction to NLP

Word embeddings in a nutshell



Class starts @ 9:05

When you penalize your Natural Language Generation model for large sentence lengths

Computer Vision / Deep Learning : Main ideas :

1] Low-level computational blocks : Conv, Pooling, Activation.

2] Transfer learning.

3] Base task : Image classification

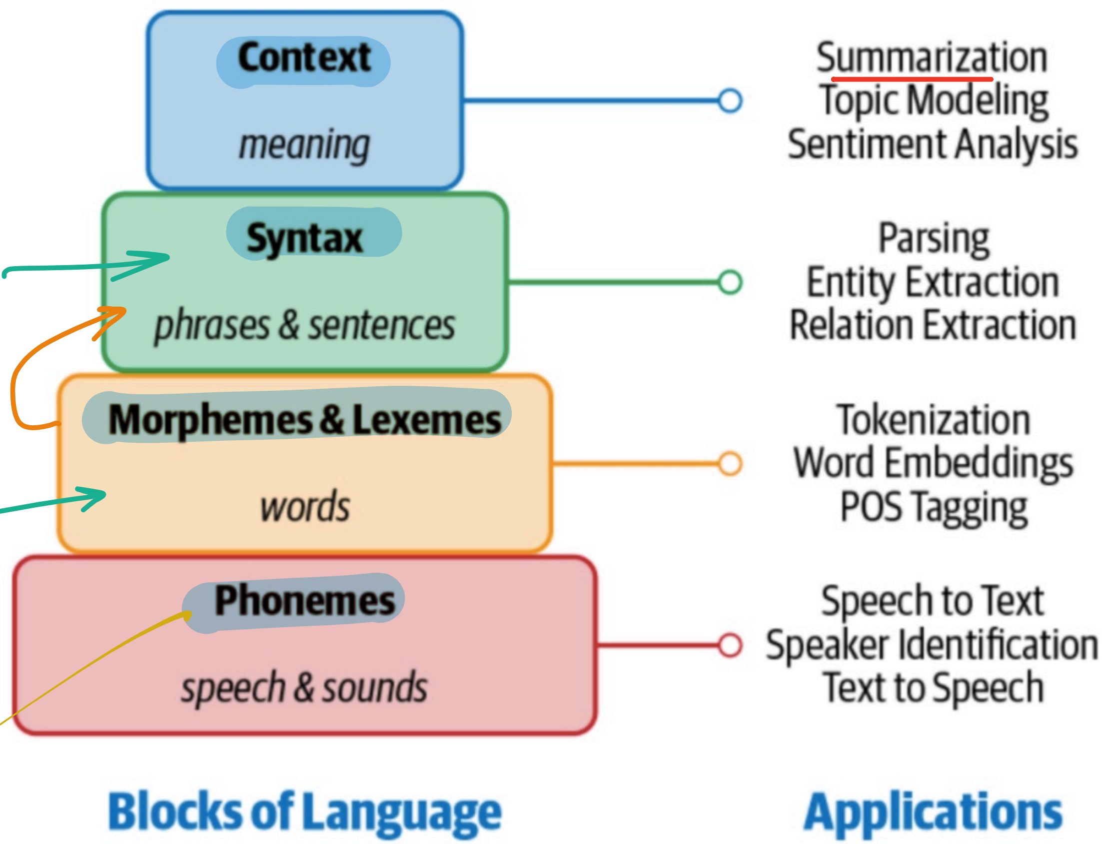
4] Architectures : skip connection, Inception modul, 1x1 convs. ROI, RPN etc.

* BERT, word2vec, RNN, LSTM, Transformer.

Natural language processing

→ Words, Sounds.

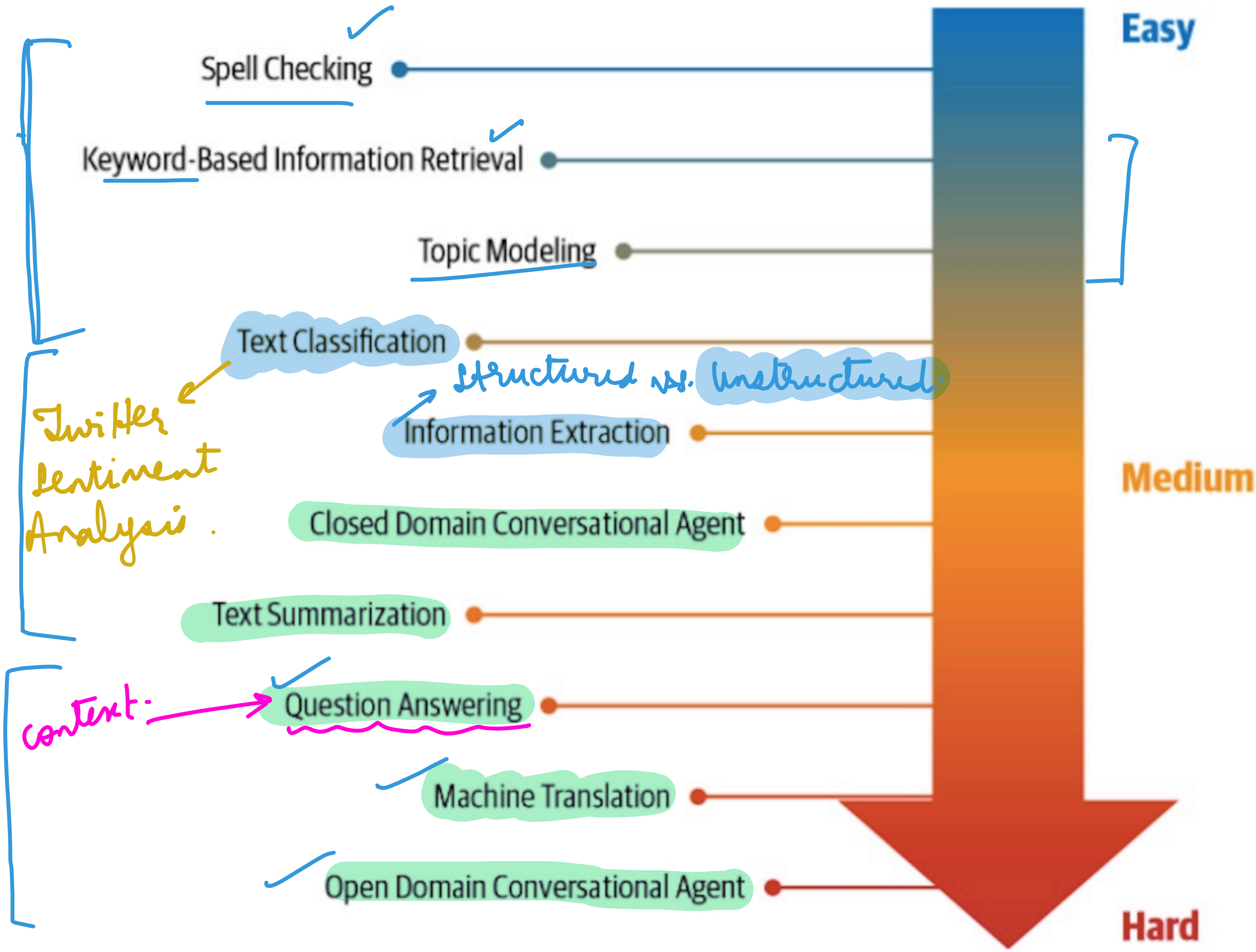
He usually lost to make her feel better. That day, he won.



} Speech signal processing.

smallest unit which represents sound.

NLP: Hierarchy of difficulty.



* pre - suffix meaning before.

x, y, z → suffixes meaning the same thing.

morphemes.

un, im, dis, mis.

Broad Approaches to NLP Problems:

1] Heuristics based Approach.

2] Rule-based approaches.

3] Optimization / learning.

→ Develop a Heuristics based approach to solve Twitter sentiment analysis.

Our first NLP problem:

Twitter dataset - Covid-19 tweets.

Objective: Sentiment analysis on Covid-19 tweets.

positive

negative.

Regular expressions.

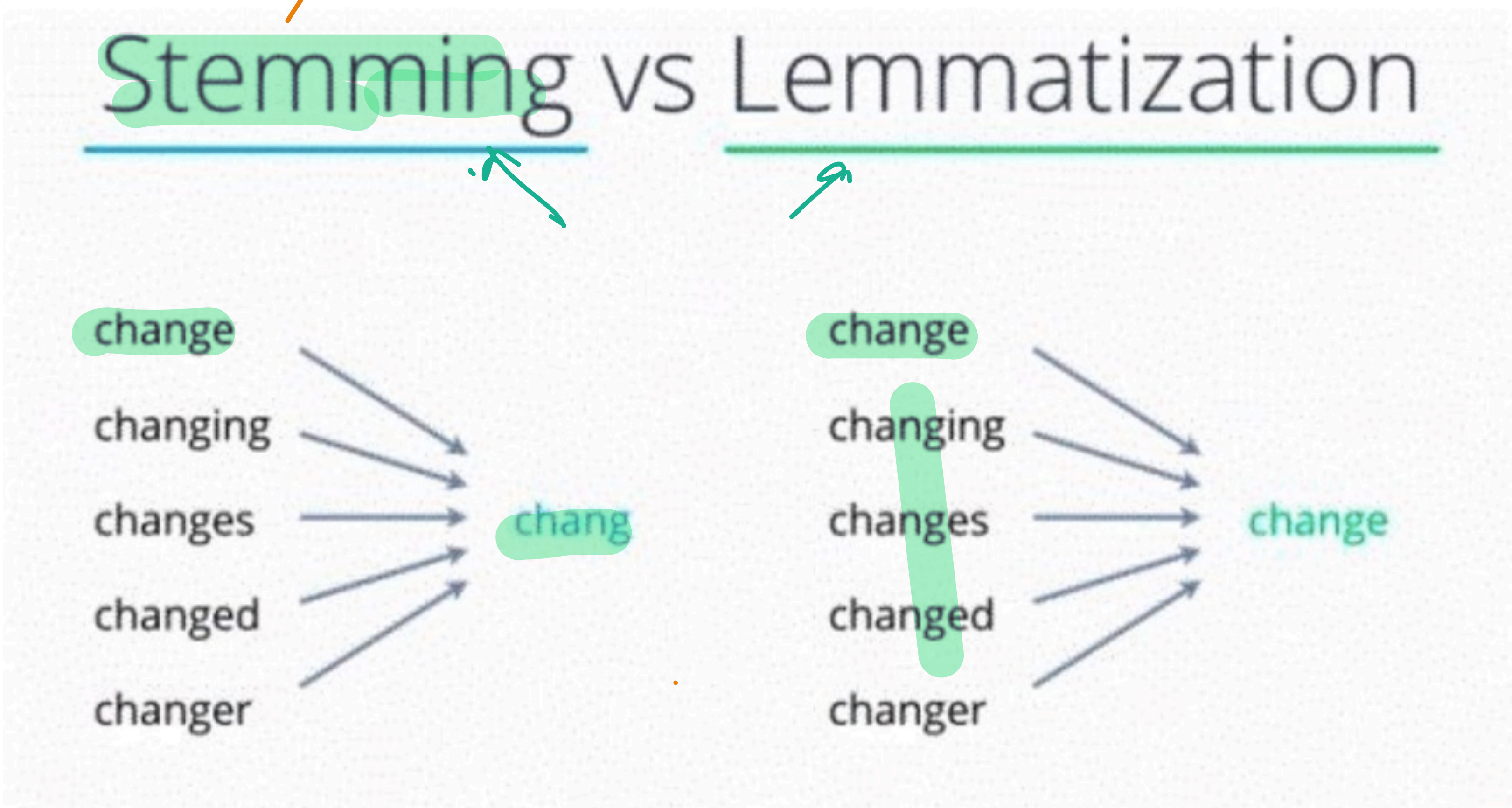
↳ Computer Science : Theory of Computation / compiler theory.

* What is a regular expression?
  → It is a set of pre-defined rules about characters or text, which help us "cluster" them into unnamed clusters

* Application : Tokenization

The process of extracting individual words from a tweet.

These methods help us further reduce words to its root form.



Converting tweets to feature vectors.

Feature vector: All tweets have variable lengths.

$$[\,1,\ \underline{\text{count\_positive}},\ \underline{\text{count\_negative}}\,] \in \mathbb{R}^3.$$

Every tweet → vectorization.

Eg: {

("eat", 0) : 67, ← counts.

("eat", 1) : 63,

("hell, 0) : 117,

⋮

-ve

+ve

Feature extraction

freqs: dictionary mapping from (word, class) to frequency

$$X_m = [1, \sum_w freqs(w,1), \sum_w freqs(w,0)]$$

Features of tweet m

Bias

Sum Pos. Frequencies

Sum Neg. Frequencies

$$D = \{ (\bar{x}_i, \bar{y}_i) \}_{i=1}^{n}$$

$\bar{x}_i \in \mathbb{R}^d$ $\quad$ $d = 3$

$\bar{y}_i \in \mathbb{R}$.

ANN, logistic / Linear Regression.

DT → Decision trees.

NB → Naive Bayes.

What could we have done better?

* We could have gone for sparse (BoW) representation.
* Use alternative representation (vectorizers, embeddings etc.)
* Normalizing our vector representation.
* Counts of synonym words. (No semantics captured in the representation).
* We have not used contextual information in any way.

Problem : Driver Assistance.

Driver drowsiness.

Road actions monitoring.

Trip monitoring.

object detection.

Pedestrians,
Emergency vehicles.
Free space