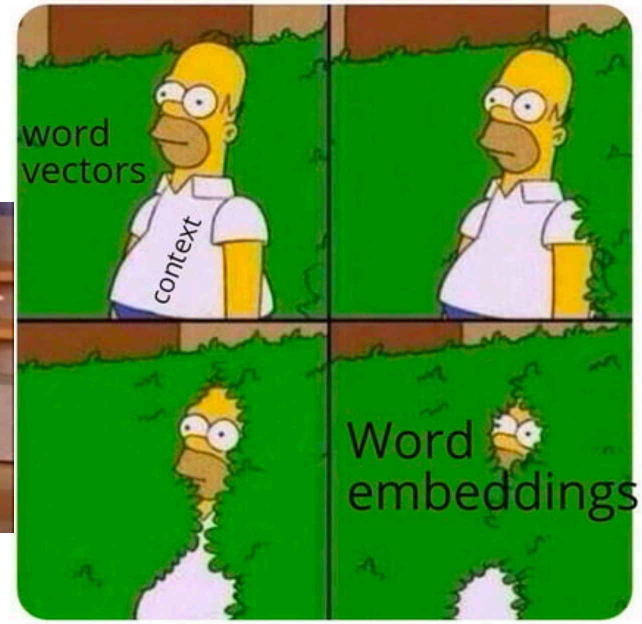
April 28, 2023.

DSML : NLP module.

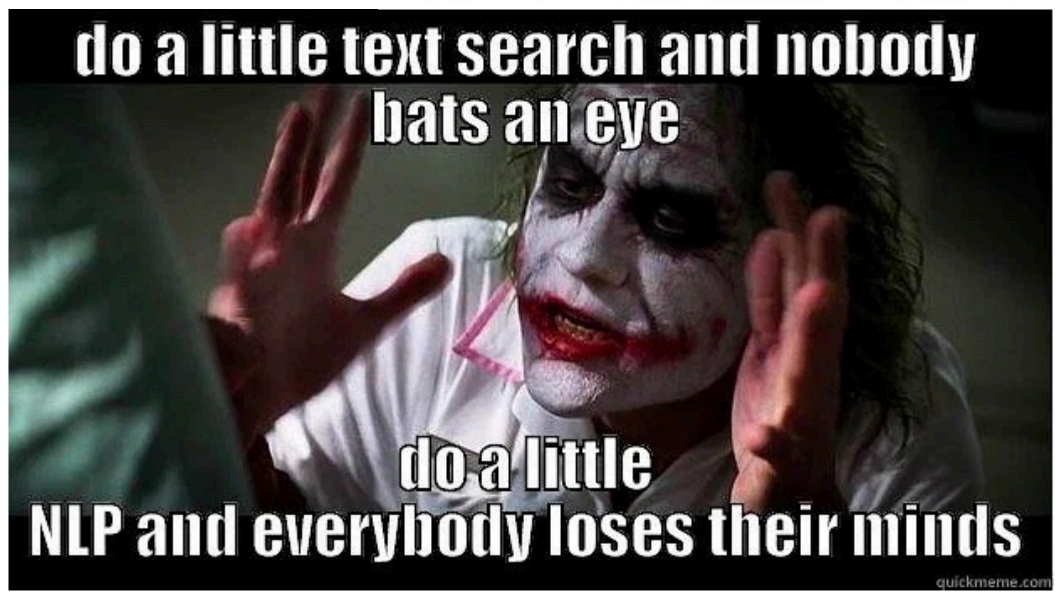
Language Modeling.

Class starts @ 9:05

Word embeddings in a nutshell



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



Recap:

* Corpus → Collection of documents.
* Documents → Collection of one or more sentences.
* Sentences → Collection of words.

* Converting text datasets to vectors:
  (i) Document → Vector 1, 2.
  (ii) Word → vector. 3, 4.

* Methods for vectorization:
  (i) Bag-of-words
  (ii) TF-IDF
  (iii) Continuous Bag-of-words
  (iv) Skip-gram.

Agenda:

* Language modeling - teaching a computer how to form sentences.

* Approaches: What could be done here?

* Techniques: Unigram, Bigram, Tri-gram, n-gram.

* Core concept: Conditional Probability.

* But first... business case!!

Naïve approaches - How to solve the problem?

* Conditional prob. → Naïve Bayes.

* Word 2 Vec → Cosine similarity.
  ↳ problem: similar words, maybe not so useful.

* Take CBOW → modify to produce the last word, we could have a NN. solution.

Conditional Prob.

A and B are independent if $P(A \mid B) = P(A)$

A, B

$P(A)$, $P(B)$.

$P(A \cap B)$

✓ $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}.$$

How to apply conditional probability to predict the next word?

Vocabulary

$$\left\{ \begin{array}{l} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n. \end{array} \right.$$

* The cat is ____ .

w_1 w_2 w_3 w_4.

w_4 is nothing but the word which maximizes the following probability

$$w_4 = \underset{i}{\arg\max}\; P(w_i \mid w_1 = \text{The}, w_2 = \text{cat}, w_3 = \text{is}).$$

Problems with probabilities:

1] $P(w_k \mid w_{k-1}\, w_{k-2}\, w_{k-3} \cdots w_1)$

$$= \frac{P(w_k \cap w_{k-1} \cap w_{k-2} \cap \cdots)}{P(w_{k-1} \cap w_{k-2} \cap \cdots)}$$

low values. → log probabilities.

2] $P(w) = 0$ → happens when we have not seen w before.

↪ Laplace smoothing.

Joint probability

$$P(w_1 \cap w_2 \cap w_3 \cap \ldots \cap w_n)$$ 0 ?

↳ "What is the probability of seeing → $w_1, w_2, w_3, \ldots, w_n$ in exactly this order?"

My Pet.

My pets name is ___

My pet is a ___

My pet's color is ___

① $P(\underbrace{w_1, w_2,} w_3, \ldots w_n)$

$$= \underline{P(w_1)} \cdot \underline{P(w_2)} \cdot \underline{P(w_3)} \cdot \underline{P(w_4)} \ldots P(w_n).$$

Naive Bayes approach: All word occurrences are independent of the others.

the.

corpus [cats

My dog the the the . . . .

① $P(\underbrace{w_1, w_2}_{A}, \underbrace{w_3}_{B})$

$\xrightarrow{(a)} P(w_1) \cdot P(w_2) \cdot P(w_3)$

$\rightarrow \underbrace{P(w_3 \mid \cancel{w_1}, w_2)} \cdot P(w_1, w_2).$

$P(w_3 \mid w_2) \cdot P(w_2 \mid w_1) \cdot P(w_1).$

"Bigram"

w_1 w_2

The man saw a girl looking through his telescope.

w_1 w_2 w_1 w_2

N-Gram → A window size of N is selected to calculate the probability.

The man ___

Bigram :

I have three books with me.

<start> → new token / new word we are using to mark the start of the sentence.

$P(w \mid \langle start \rangle)$

Trigram :

$P(w \mid \langle start \rangle, \langle start \rangle)$

<a href = "http:\\ . . . >