MAKING SENSE OF
UNSTRUCTURED TEXT

Python & spaCy

PRESENTED BY CHASE GRECO

CHASE GRECO

DATA SCIENTIST
COSTAR GROUP

- LinkedIn: /chasegreco
- Github: /zerthick

COVERED TODAY

MAIN POINTS OF DISCUSSION

What is CoStar?
Case Study: Apartments.com
Traditional Approaches to NLP
Modern(ish) Approaches to NLP
Intro to spaCy
Demo
Questions

Commercial Real Estate Analytics

CoStar exists so that commercial real estate professionals have a clearinghouse of commercial real estate and multifamily information that help them make better, faster decisions, connect with the right people, and get more deals done.

- 200,000 Professionals use CoStar for Information
- C.R.E. Profressions spend 40% of their time collecting and managing data - Ernst & Young
- 3600 Employees
- 80 Offices in 7 Countries
- Worth $17 Billion

CASE STUDY APARTMENTS.COM

4 Star

This is a great first apartment. The amenities of the complex are awesome. Great gym and pool. The appliances in the apartment need to be updated and the floors are loud, but the size is perfect for 2 people.

1 Star

When I moved into this apartment things were clean. When I moved out, I cleaned everything very thoroughly, however, they found one spec of dirt inside a cabinet and they charged me $35, claiming that I didn't clean it.

CAN WE INFER THE POSTIVITY OR NEGATIVITY OF A REVIEW?

Hint: The answer is yes

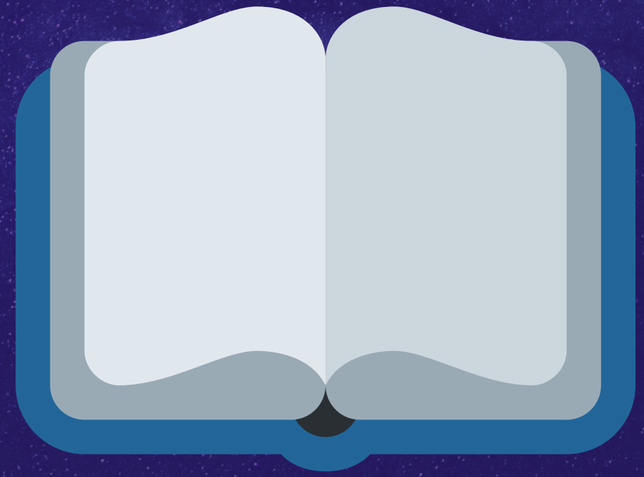
GETTING COMPUTERS TO UNDERSTAND TEXT

Traditional Approach - One-Hot Encoded Representation

Corpus (News)

Enummerate Words

Dictionary



EXAMPLE

Sentence

The quick brown fox jumps over the lazy dog.

Tokens

brown, dog, fox, jumps, lazy, over, quick, the

Vector

ant	...	apple	...	brown	...	fox	...	giraffe	...	lazy	...	zebra
0	...	0	...	1	...	1	...	0	...	1	...	0

CHALLENGES

THINGS TO KEEP IN MIND

Long

The vectors created by this approach are as long as the total number of words in your dictionary!

Sparse

The majority of the features in your vector for any given sentence will be zero!

Noisy

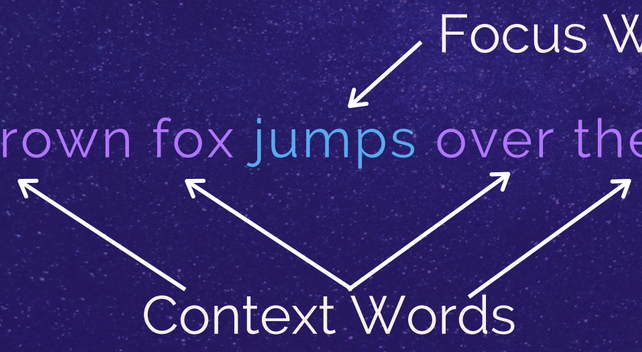
Many of the tokens do not provide any meaningful information!

A (MORE) MODERN APPROACH

Enter word2vec

- Vectors have a fixed length and are **dense**
- Based on the concept of understanding the **context** of a word

Example



Can we predict the presence of a context word given a focus word?

A (MORE) MODERN APPROACH

Example

The quick brown fox jumps over the lazy dog.

(the, quick)
(the, brown)

A (MORE) MODERN APPROACH

Example

The quick brown fox jumps over the lazy dog.

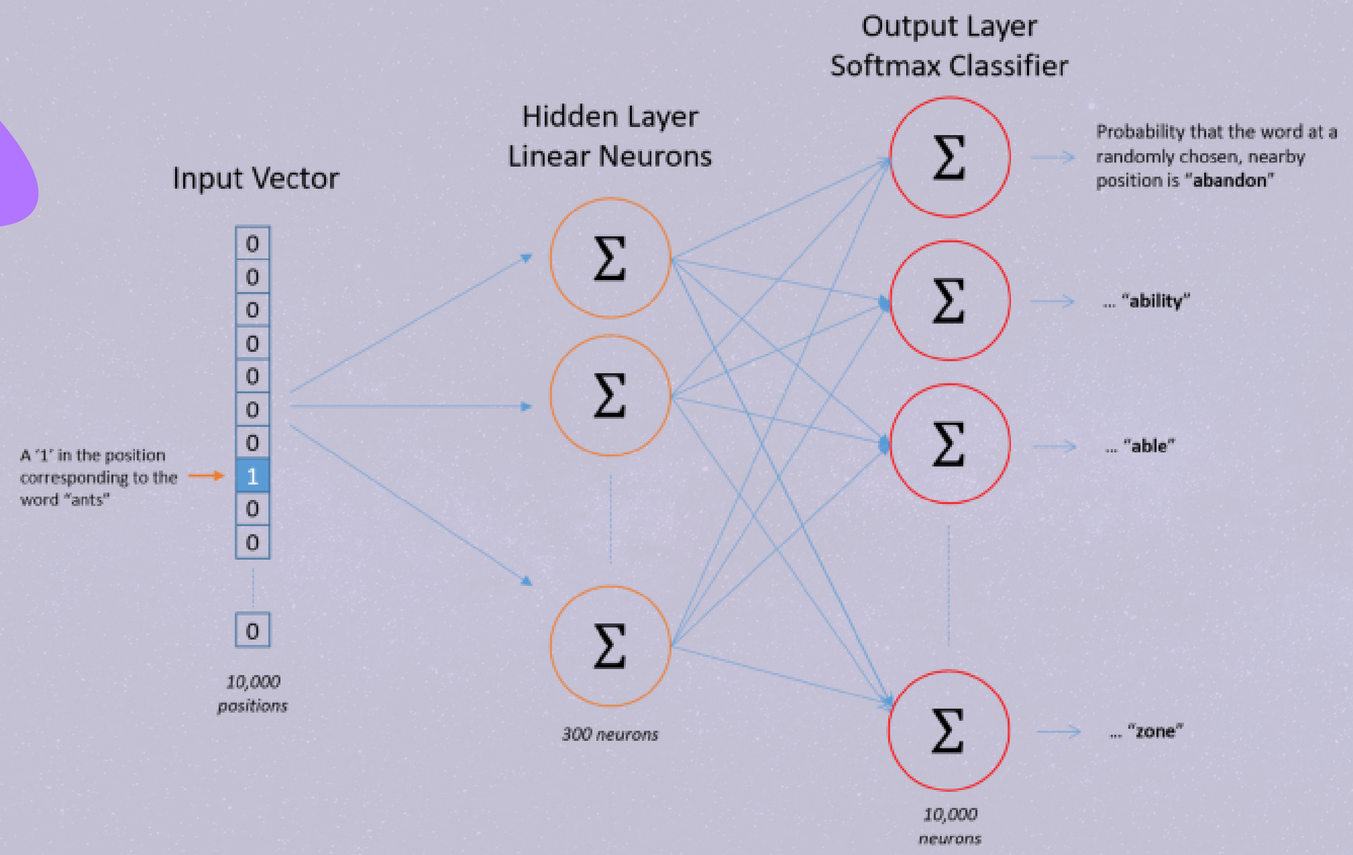
(the, quick)
(the, brown)
(quick, the)
(quick, brown)
(quick, fox)

A (MORE) MODERN APPROACH

Example

The quick brown fox jumps over the lazy dog.

(the, quick)
(the, brown)
(quick, the)
(quick, brown)
(quick, fox)
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)



https://becominghuman.ai/how-does-word2vecs-skip-gram-work-f92e0525def4

DENSE WORD EMBEDDINGS!

Examples

- Fox : [-0.34868 , -0.07772 , 0.17775 , -0.094953, ...]
- Dog : [-4.0176e-01, 3.7057e-01, 2.1281e-02, -3.4125e-01, ...]
- Quick : [-4.4563e-01, 1.9151e-01, -2.4921e-01, 4.6590e-01, ...]

DEEP LEARNING WITH TEXT

A FOUR STEP PROCESS

EXPLOSION.AI/BLOG/DEEP-LEARNING-FORMULA-NLP

Embed

Map long, sparse vectors into dense continuous vectors (word2vec)

Encode

Encode a word vector sequence into a sentence representation

Attend

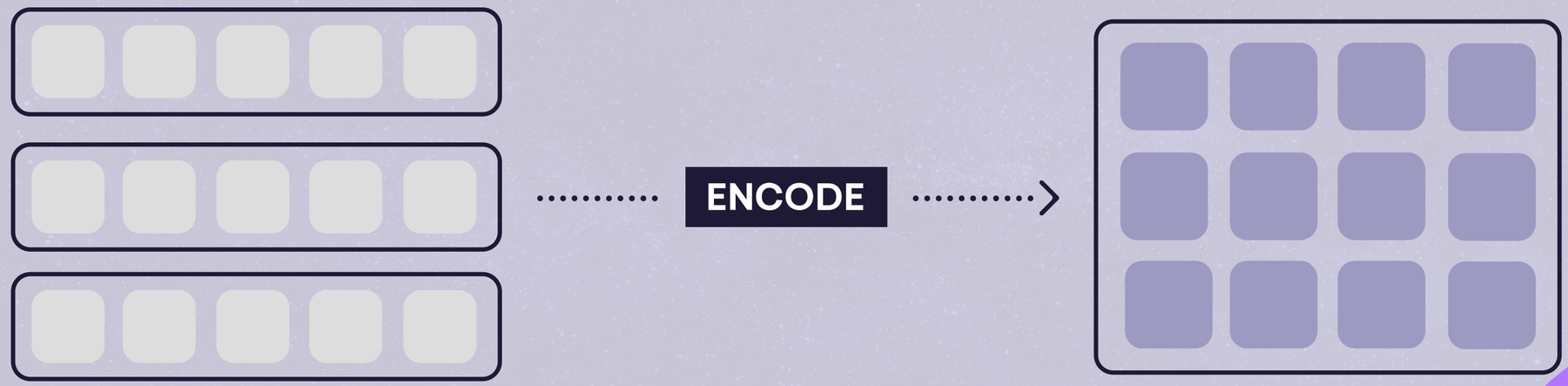
Reduce sentence representation into a single vector

Predict

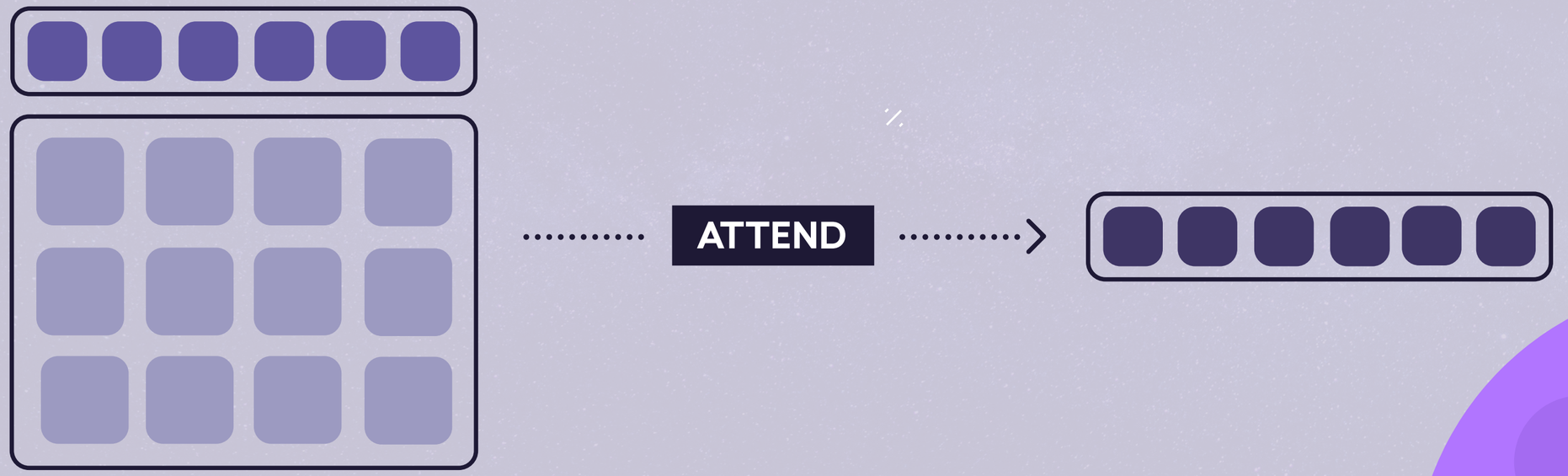
Presentations are communication tools that can be used as lectures.

EMBED

ENCODE



ATTEND



PREDICT

SPACY

INDUSTRIAL-STRENGTH NATURAL LANGUAGE PROCESSING

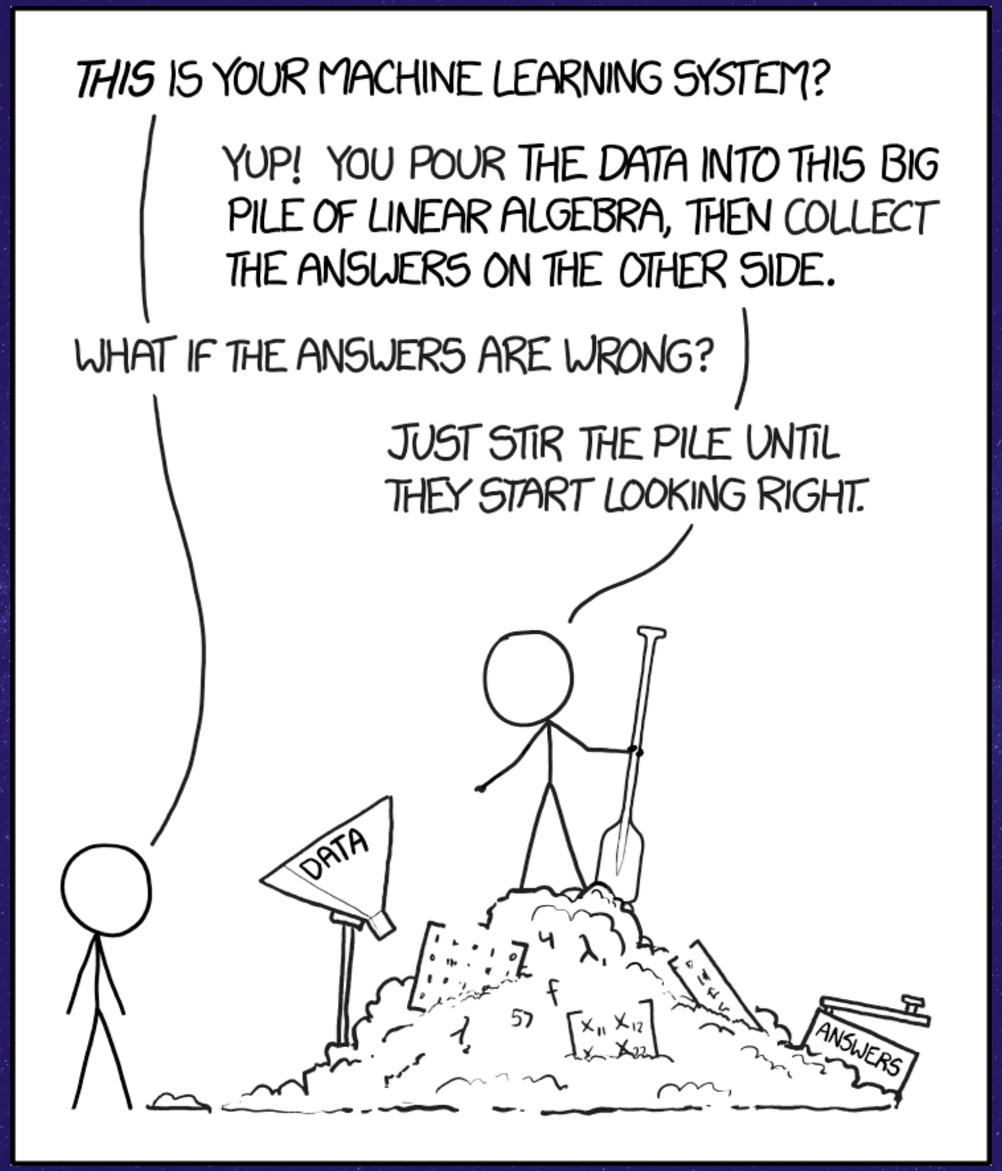
Python framework for performing a variety of NLP tasks, such as text classification, named entity recognition (NER), and part of speech tagging. Comes with **pre-trained** models to get started quickly.

Making Sense of Unstructured Text

DEMO

PYTHON WITH SPACY

IMDB Review Classification



QUESTIONS?