DigitalOcean

Demystifying Natural Language Processing

Building ML Models and How to Leverage Them By Example

David vonThenen

@davidvonthenen

SCaLE

DigitalOcean

David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy...
- There is storage for that!

@davidvonthenen

Agenda

- "Hello World": Question vs Sentence
- Named Entity Recognition (NER)
  - Obtaining/Finding the Data
  - Grooming and Formatting the Data
  - Processing Data and Building the Model
- Demo: Multiple NLP Models
- Resources For You!!
- Q&A

SCALE

DigitalOcean

Our First NLP Model

Machine Learning Terms, Basics, Etc

Level Set with ML Models

- Data(set)
  - Domain of Problem, "Examples"
  - Search/Pattern Amongst
- Tokenzier
  - BERT uncased, DeBERTa, etc
- ML Framework
  - PyTorch, Tensorflow, fastai

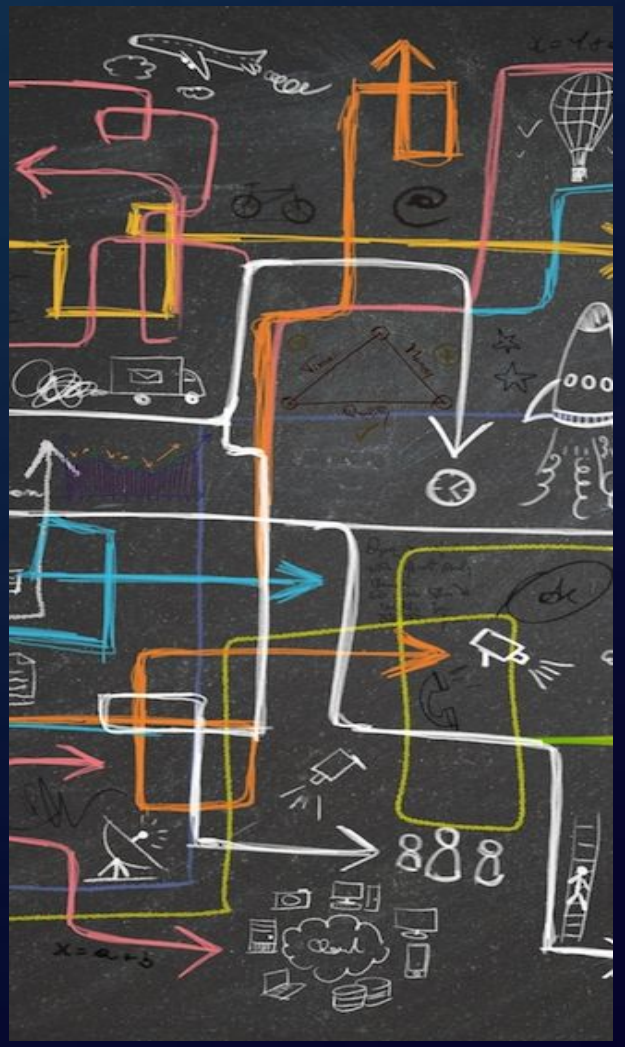
- Tensor – A Measurement (Multi-Dimensional Matrix of Measured Data)
- Some Popular Supporting Libraries:
  - pandas, NumPy, etc

Building Your First NLP Model

- Classification Models Easier to Understand
- Starter Model: Sentence or Question?
- Off-the-Shelf/Curated Datasets
  - Data... Lots of Data
  - Stanford Question Answering Dataset (SQuAD)
- Classify the Data:
  - Yes or No, 1 or 0

More Complex That You Think...

- While This Seems Straightforward
  - Couldn't You Just Look For "?"
- Consider These Examples:
  - Is this an example sentence?
  - My name is John Doe.
  - How are you doing my friend
  - Tell me about the history of the United States.



DigitalOcean

Demo: Question Classifier

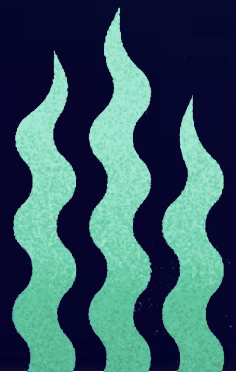
Qs vs non-Qs Recap

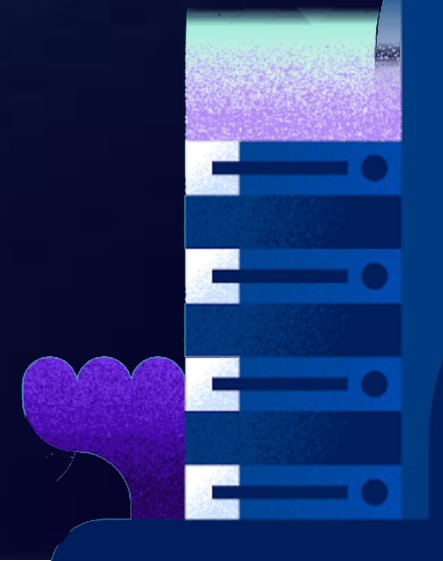
- People Don't Conform to Language Rules
- Things to Consider. Not All...
  - Questions End With a Question Mark
  - Sentences End With a Period
- More Complex Than We Think
  - Not All Question Start With:
    - Who, What, When, Where, Why, How
  - Some "Questions" End With a Period

SCALE

DigitalOcean

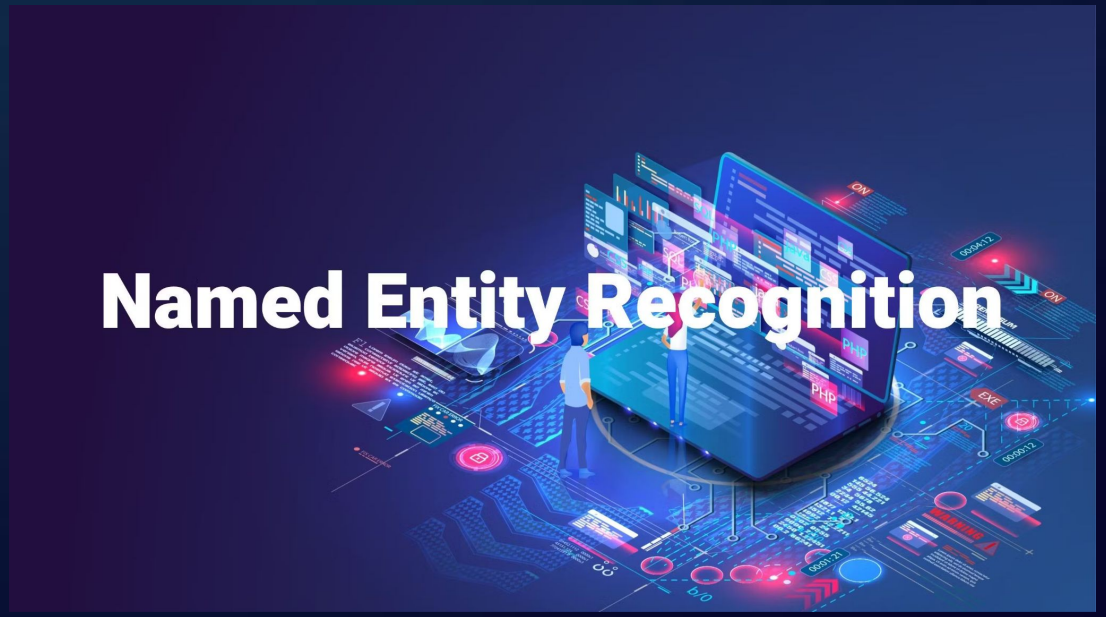
Building NLP Models Named Entity Recognition And Redaction





What Are Named Entities?

- Extracting and Classifying "Things" Mentioned in Unstructured Text into Predefined Categories
- Typically Means:
  - Personally Identifiable Info
    - Name, Age, SSN, IP Address
  - Protected Health Info
    - Blood Type, Drug, Injury
  - Payment Card Industry
    - Credit Card #, CVV
- More Basic, It's Just a Label



Obtaining/Finding the Data

- Most Difficult Part is Getting the Data
- Look Everywhere...
  - GitHub – Entity Recognition Repo[1]
  - Huggingface
  - Kaggle – Projects w/ Datasets
  - Academic Torrents
- and Get Creative...
  - Any CoNLL[2] Formatted Dataset
  - Ask Researchers! Some Will Share!
  - Synthetic Data – Becare With This!

Grooming and Formatting

- CoNLL Format Desirable Due to Availability
  - "Standard" Widely Available Format
- The Simplistic View...
  - Capture Words in Sentences
  - Each Word is Labelled
  - Labels Apply to Multiple Words
    - United States of America
- Label = Classification!
  - PII, PHI, PCI SSC

CoNLL Format – Good

Word	Part of Speech	Syntactic Chunk	Entity Tag
United	NNP	I-NP	**B-ORG**
Nations	NNP	I-NP	**I-ORG**
official	NN	I-NP	**O**
Ekeus	NNP	I-NP	**B-PER**
heads	VBZ	I-VP	**O**
for	IN	I-PP	**O**
Baghdad	NNP	I-NP	**B-LOC**
.	.	O	**O**

CoNLL Format – Bad

Word	Part of Speech	Syntactic Chunk	Entity Tag
United	NNP	I-NP	O
Nations	NNP	I-NP	O
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-LOC
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-PER
.	.	O	O

Processing and Building

- After Data Is Formatted, We Need Structure!
- Word, "Tag Map" Or...

Word	O (No Entity)	B-ORG	I-ORG	B-TIME	...
United	0	1	0	0	...
Nations	0	0	1	0	...
is	1	0	0	0	...

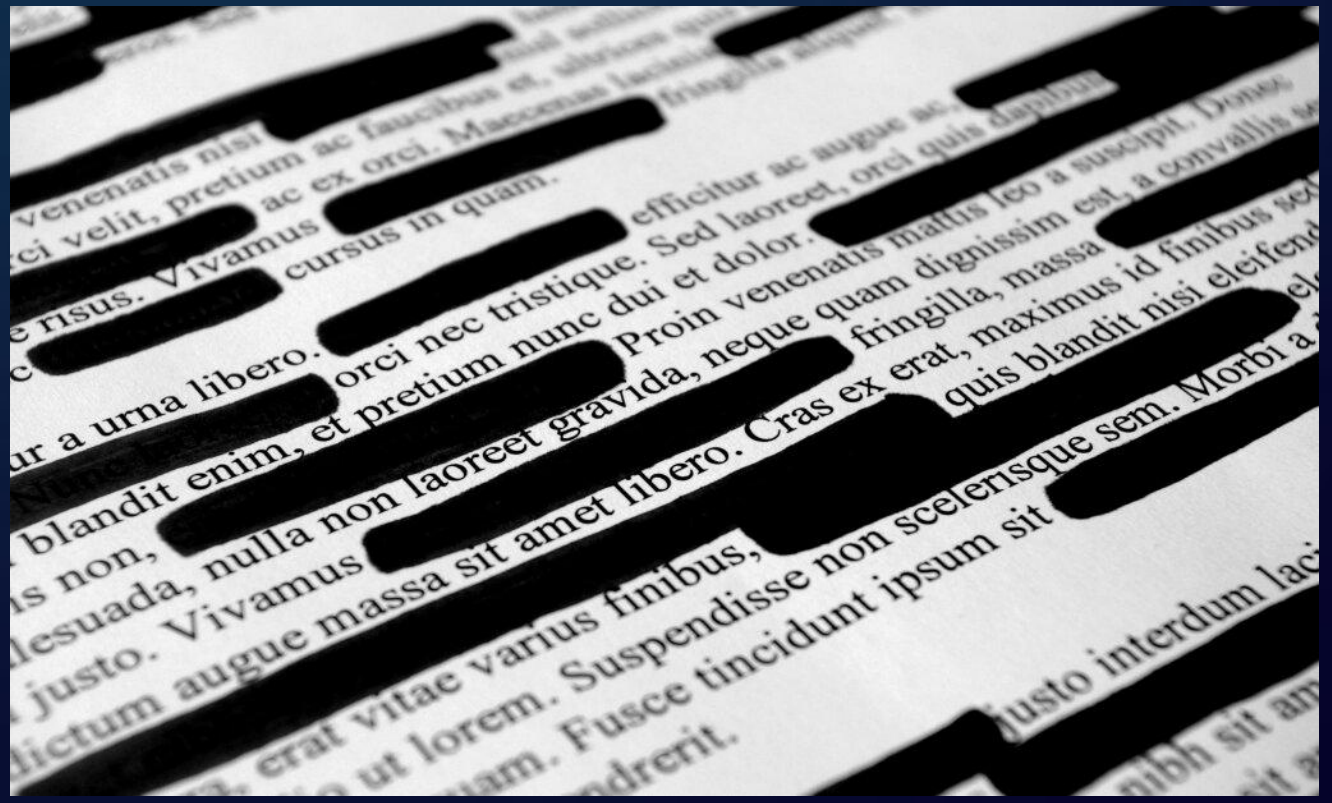
- Tokenizer = bert-base-uncased
- Each Sentence Composed of a Tensors for:
  - Tokens, Entity Labels, Attention Mask (Padding)

DigitalOcean

Demo: NER Classification

What About Redaction?

- Censoring/Obscuring Text for Legal/Security Purposes
- Popular Examples:
  - X-Files
  - Who Killed JFK?
- Remember These?
  - PII, PHI, PCI SSC
- How to Redact?
  - Identify Based By Label
  - Replace: Word -> *****



NER and Redaction Recap

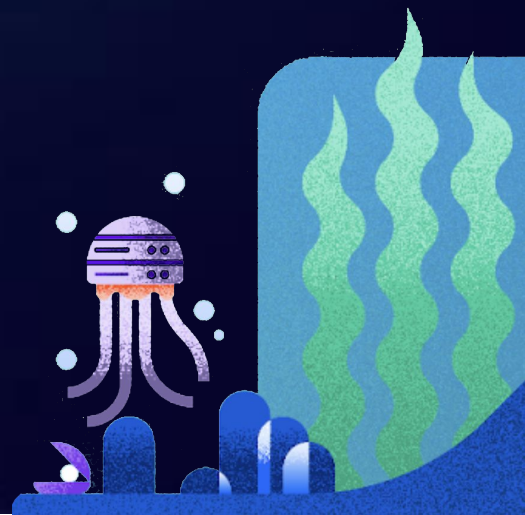
- Find Datasets and Start With Low-Hanging Fruit
  - Custom Data(sets)?
- Most Difficult: Grooming the Data
  - Does Data Accurately Reflect the Problem
  - Fix the Data! Correct the Errors
- Structure the Data for ML Training
- Generate the Model, Does It Work?
- Rinse and Repeat, Always Outliers
- Iterative Improvements, Refinement

SCALE

DigitalOcean

Real World Application

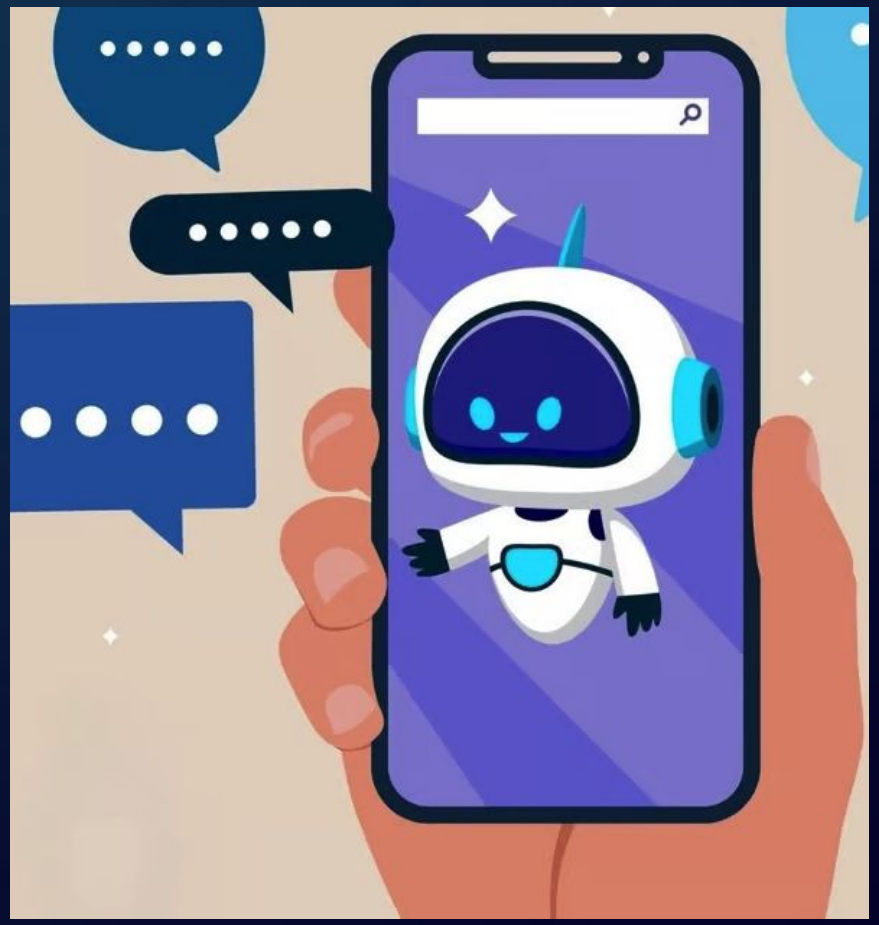
Mining Data From a Virtual Meeting



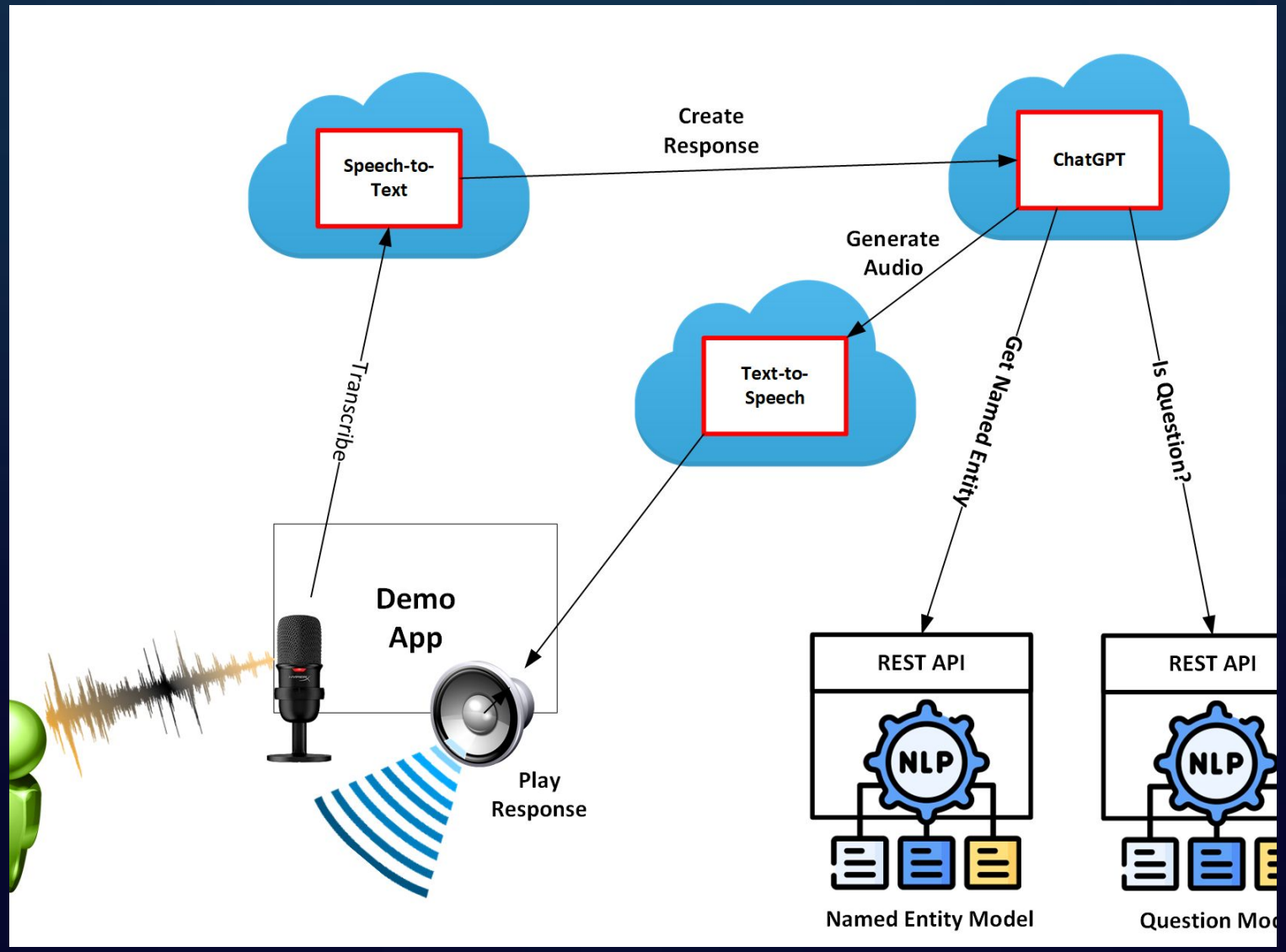
AI Assistant in a Virtual Meeting

Scenario: Voice Assistant to Help in a Meeting

- Simple Data Mining
- Can Answer Questions
- Details On Discussed Topics
- Topic Steering, Leading Convo
- Help Assist Attendees With:
  - Call Center Assist
  - Sales Assist
  - And More!



Voice AI Assistant Architecture



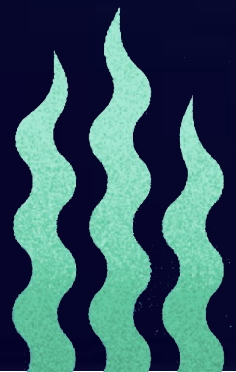
DigitalOcean

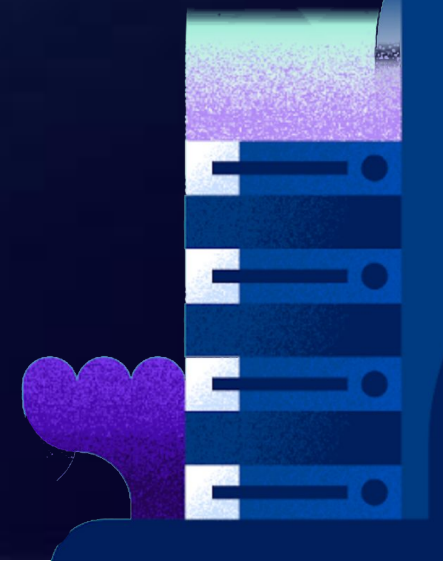
Demo: Mining Meeting Data

SCALE

DigitalOcean

Resources





Resources

[CLICK HERE] for All Material Contained in this Session [CLICK HERE]

Building These Models on Your Own

- Google Colab Notebook Instructions
- Kaggle Notebook Instructions
- Full Laptop Deploy Instructions

Other Resources:

- OpenAI API Playgroud
- Deepgram Speech-to-Text: API and Docs
- Deepgram Text-to-Speech: API and Docs
- Juan Diego Rodriguez – Named Entity Repo

DigitalOcean

Thank You!

David vonThenen

@davidvonthenen

https://linktr.ee/davidvonthenen