20.04.22

OPENCAMPUS.sh

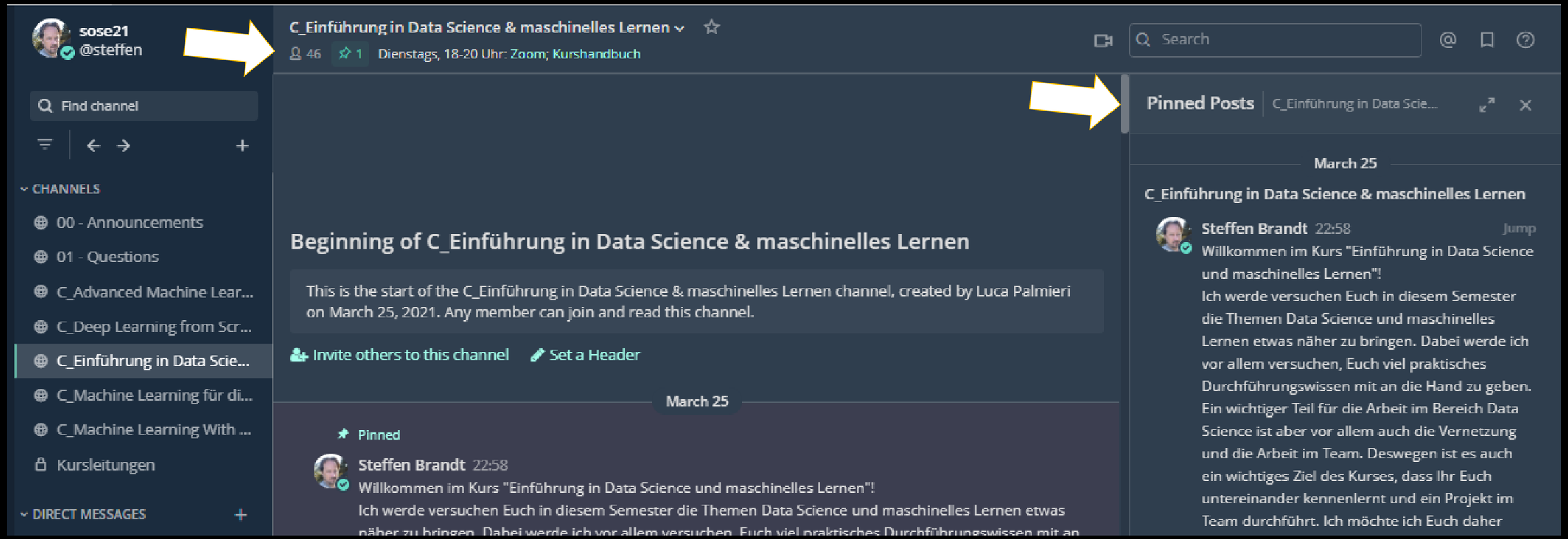
Transformers for Natural Language Processing and Beyond

GENERAL INTRODUCTION

- Organizational Matters
- Structure of the Course
- The Role of NLP in AI
- Basic Introduction to Transformers

INTRODUCTION ROUND

CHAT


sose21
@steffen
Find channel
CHANNELS
00 - Announcements
01 - Questions
C_Advanced Machine Lear...
C_Deep Learning from Scr...
C_Einführung in Data Scie...
C_Machine Learning für di...
C_Machine Learning With ...
Kursleitungen
DIRECT MESSAGES
C_Einführung in Data Science & maschinelles Lernen
46
1
Dienstags, 18-20 Uhr: Zoom; Kurshandbuch
Search
Beginning of C_Einführung in Data Science & maschinelles Lernen
This is the start of the C_Einführung in Data Science & maschinelles Lernen channel, created by Luca Palmieri on March 25, 2021. Any member can join and read this channel.
Invite others to this channel
Set a Header
March 25
Pinned
Steffen Brandt 22:58
Willkommen im Kurs "Einführung in Data Science und maschinelles Lernen"!
Ich werde versuchen Euch in diesem Semester die Themen Data Science und maschinelles Lernen etwas
Pinned Posts
C_Einführung in Data Scie...
March 25
C_Einführung in Data Science & maschinelles Lernen
Steffen Brandt 22:58
Jump
Willkommen im Kurs "Einführung in Data Science und maschinelles Lernen"!
Ich werde versuchen Euch in diesem Semester die Themen Data Science und maschinelles Lernen etwas näher zu bringen. Dabei werde ich vor allem versuchen, Euch viel praktisches Durchführungswissen mit an die Hand zu geben. Ein wichtiger Teil für die Arbeit im Bereich Data Science ist aber vor allem auch die Vernetzung und die Arbeit im Team. Deswegen ist es auch ein wichtiges Ziel des Kurses, dass Ihr Euch untereinander kennenlernt und ein Projekt im Team durchführt. Ich möchte ich Euch daher

COURSE HANDBOOK

opencampus.sh Machine Learning Program

EDU-Platform Chat

Search...

opencampus.sh Machine Learning Program

Course Kick-Off

How do I choose a course?

FAQ

COURSES

Einführung in Data Science und maschinelles Lernen

Bedingungen für ein Leistungszertifikat oder ECTS

Vorbereitung

Woche 1 - Einführung in Data Science

Woche 2 - Grafische Darstellung von Daten

Woche 1 - Einführung in Data Science

CONTENTS

Diese Woche werdet Ihr...

eine Einführung zu den folgenden Themen bekommen:

- Was ist Data Science?
- R vs. Python vs. SPSS vs. ...
- Wozu RStudio?
- Datenstrukturen in R

Lernressourcen

Präsentation zu Woche 1 - Einführung 201114_Einführung.pdf - 5MB

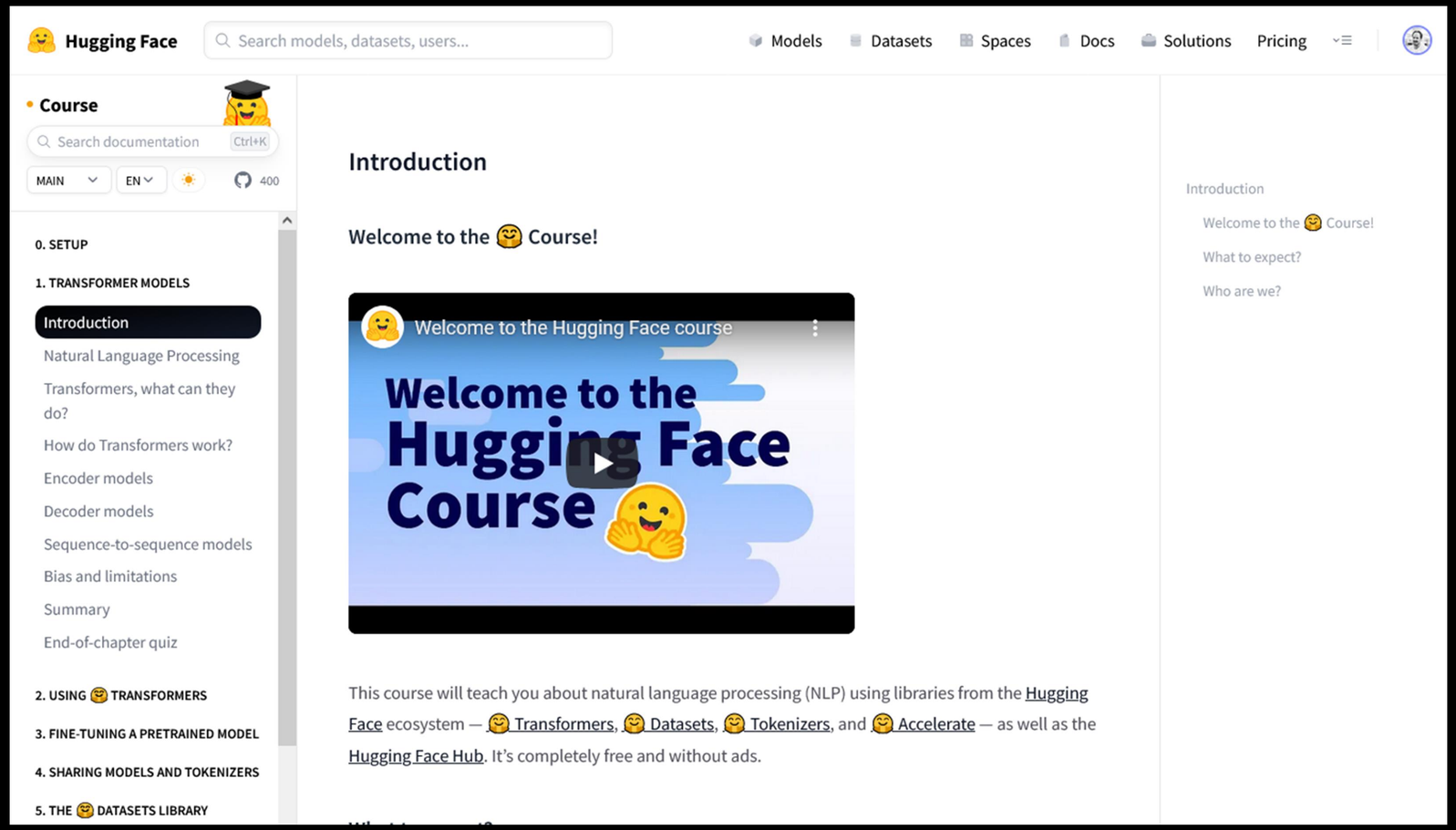
ORGANIZATIONAL MATTERS

- Use your full names in the zoom meetings!
- Scan the QR-Code if you participate in presence.
- Complete your profile in the mattermost chat with your full name and a photo.
- Please write me if you will not go on with the course!

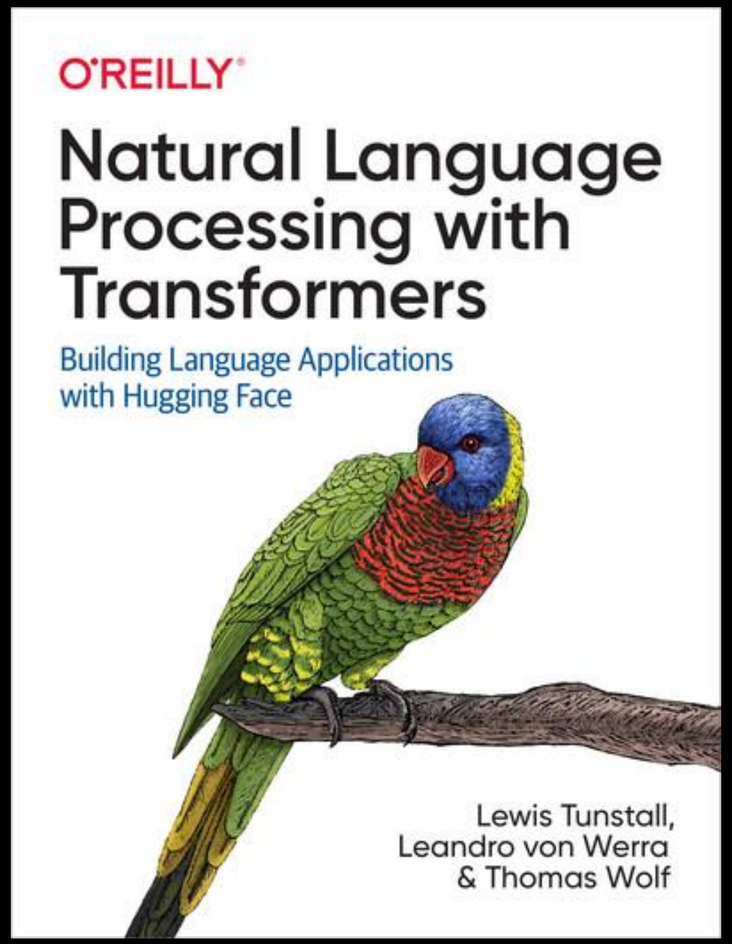
Coding.Waterkant

20. 04. **10: 00 - 12: 00**	GENERAL INTRODUCTION Remote via Zoom
27. 04. **10: 00 - 12: 00**	THE SELF-ATTENTION MECHANISM Digital - via Zoom Meeting
04. 05. **10: 00 - 12: 00**	PROMPT DESIGN Zoom + Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
11. 05. **10: 00 - 12: 00**	INTRODUCTION TO THE HUGGING FACE API Remote via Zoom
18. 05. **10: 00 - 12: 00**	FINE-TUNING PRETRAINED MODELS Remote via Zoom
25. 05. **10: 00 - 12: 00**	THE HUGGING FACE DATASETS LIBRARY Remote via Zoom
01. 06. **10: 00 - 12: 00**	APPLYING TRANSFORMERS FOR DIFFERENT TASKS Remote via Zoom
08. 06. **10: 00 - 12: 00**	JOINT CODING In presence in Kiel or remote via Zoom
15. 06. **10: 00 - 12: 00**	PRESENTATION OF THE FINAL PROJECTS Remote via Zoom

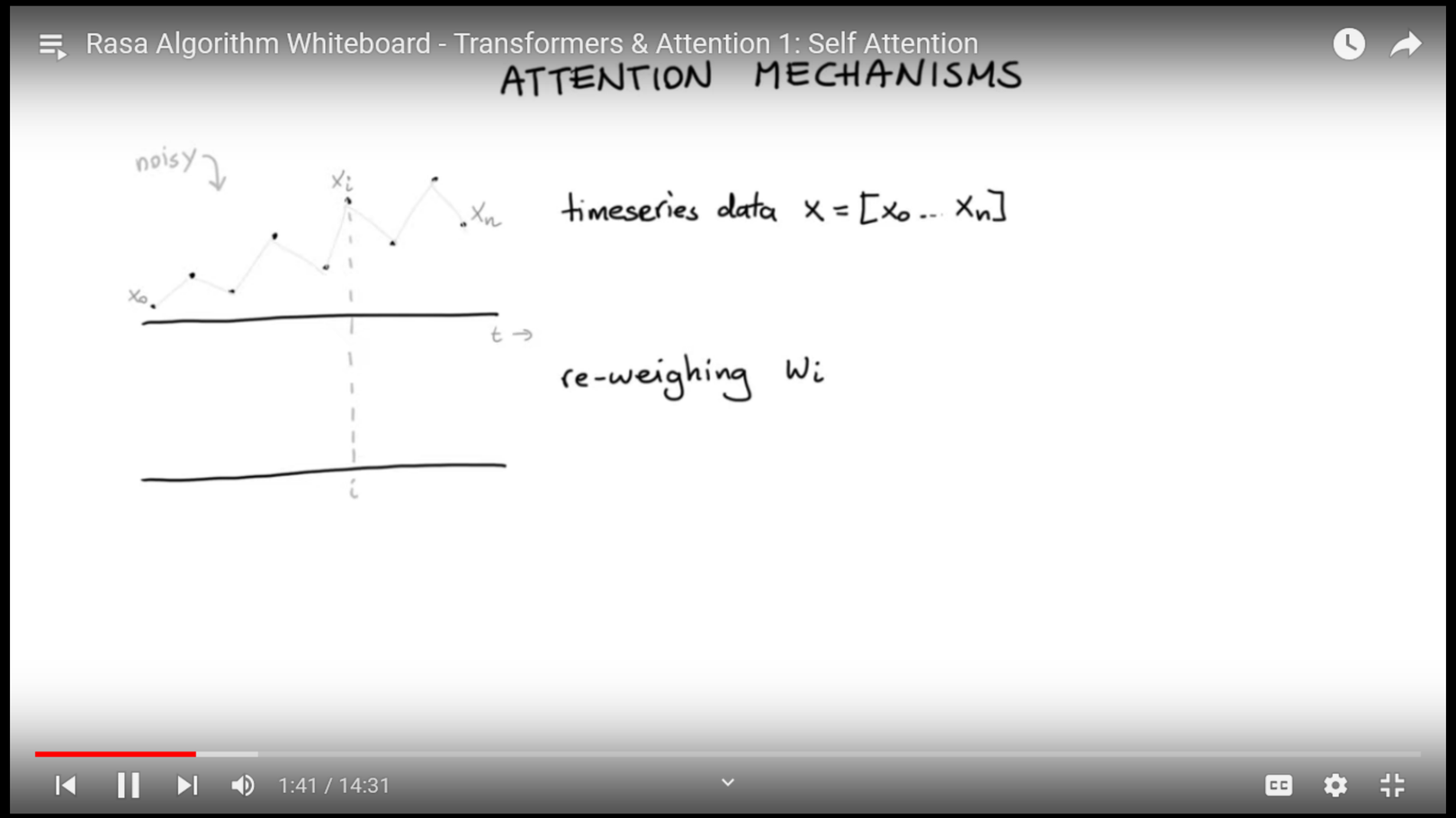
HUGGING FACE TRANSFORMERS COURSE



NLP WITH TRANSFORMERS



RASA ALGORITHM WHITEBOARD



BREAKOUT

- Which Ai impressed you the most in the last few month?
- Do you know transformer models – if so, which?

APPLICATION EXAMPLES

- Code Generation: Co-Pilot in VS-Code from GitHub and OpenAI (GPT model)
- Search: Google (Bert model)
- Prediction of Protein folding: Alpha Fold from Deepmind
- Image Generation: DALL-E 2 by OpenAI

THE TRADITIONAL ROLE OF NLP IN AI

- **A subfield of linguistics, computer science, and artificial intelligence**
- **Dealing with the interaction between computers and human language**
- **The goal is a computer capable of "understanding" text contents, including contextual nuances**
- **The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.**

Natural language processing. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=1048621730

THE ROLE OF NLP FOR ARTIFICIAL GENERAL INTELLIGENCE (AGI)

Touring Test:

- **Test for intelligence in a computer, requiring that a human being should be unable to distinguish the machine from another human being by using the replies to questions put to both.**

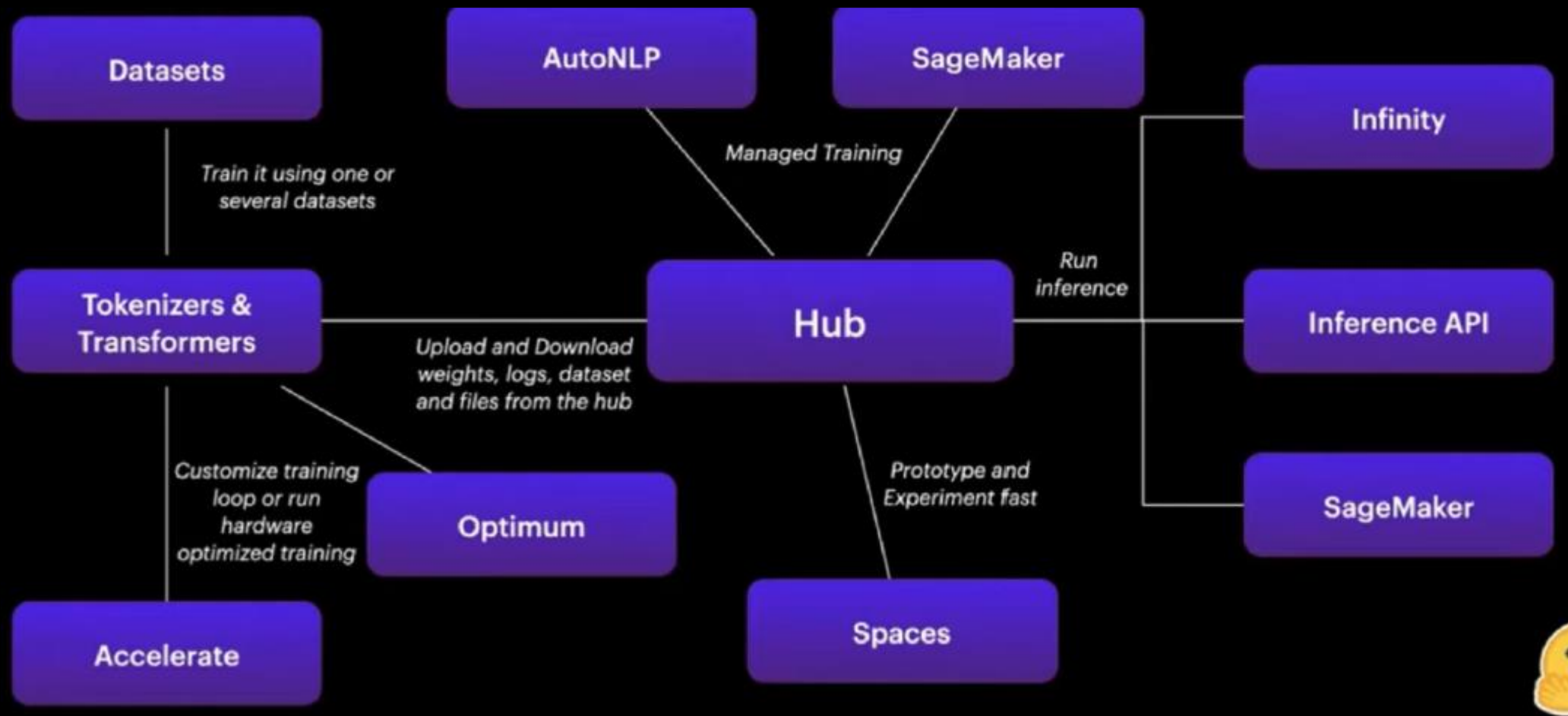
AGI RESEARCH

DeepMind

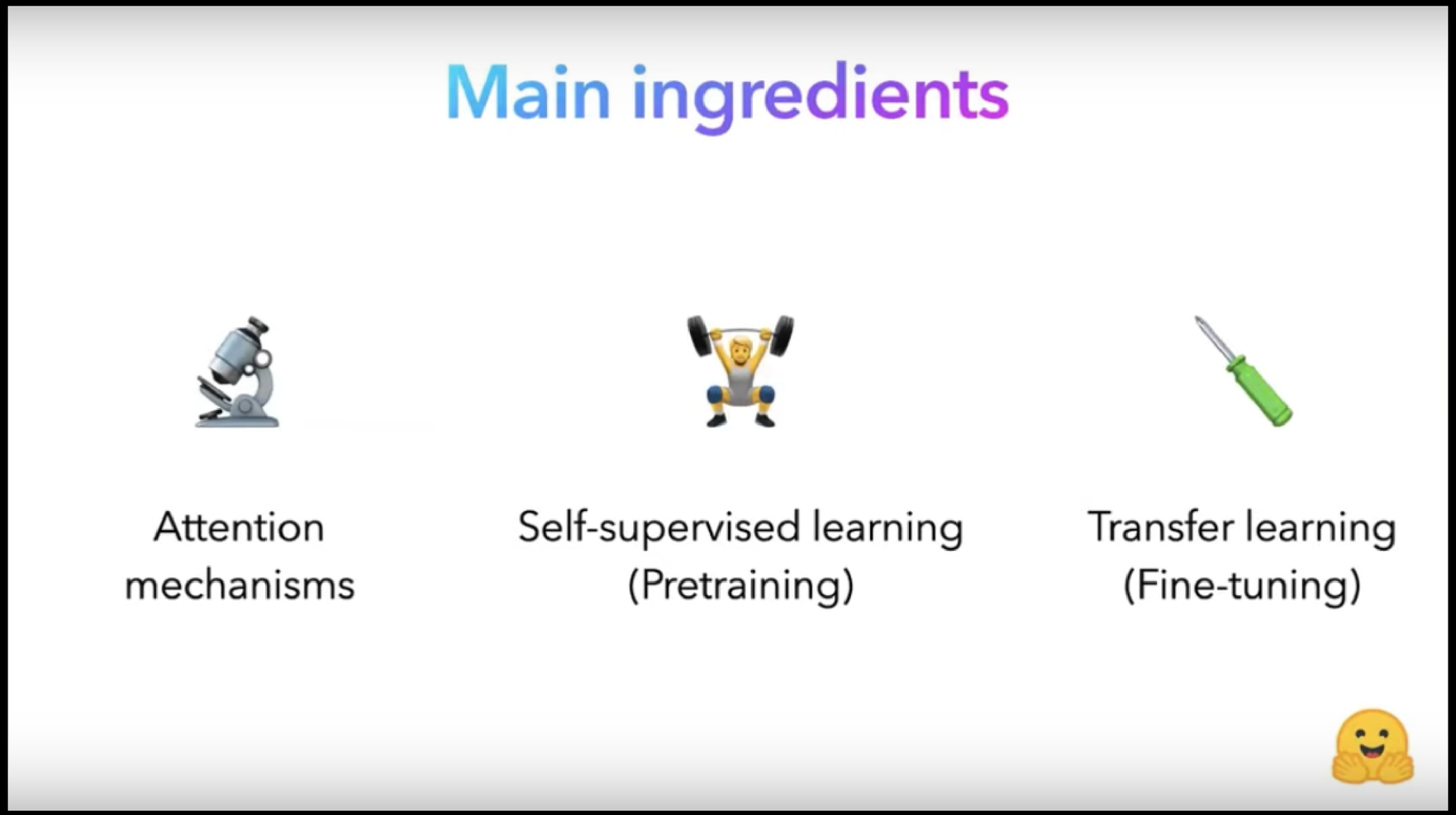
OpenAI

ANTHROP\C

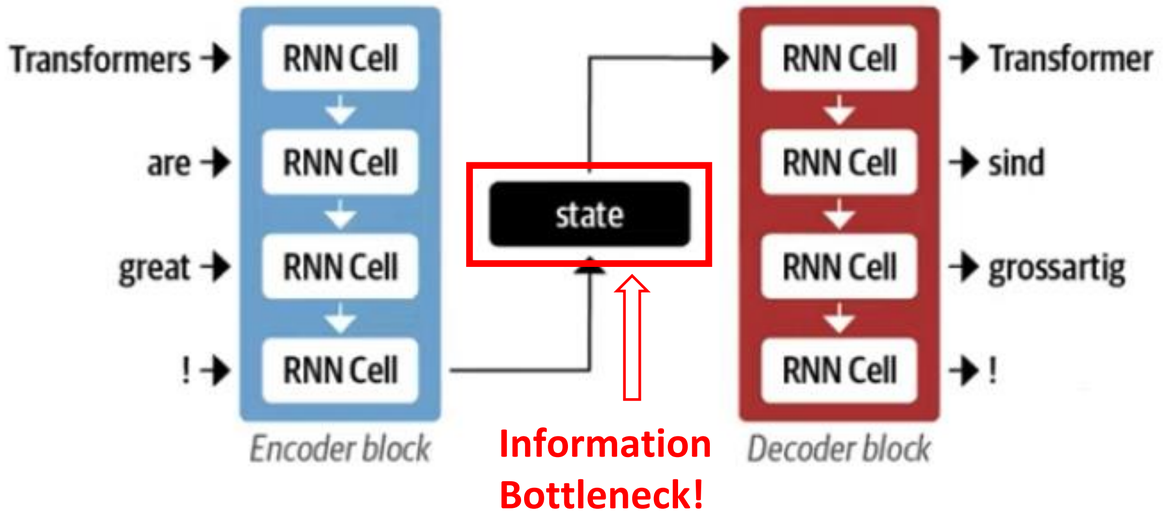
HUGGING FACE ECO SYSTEM



WHAT IS A TRANSFORMER?

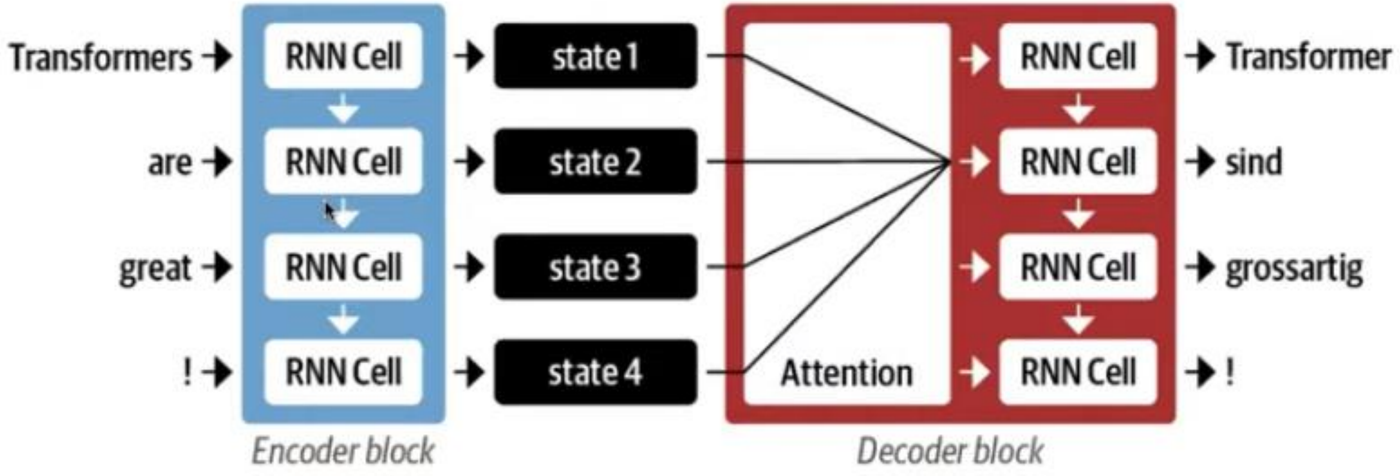


CLASSICAL SEQUENCE TO SEQUENCE APPROACH



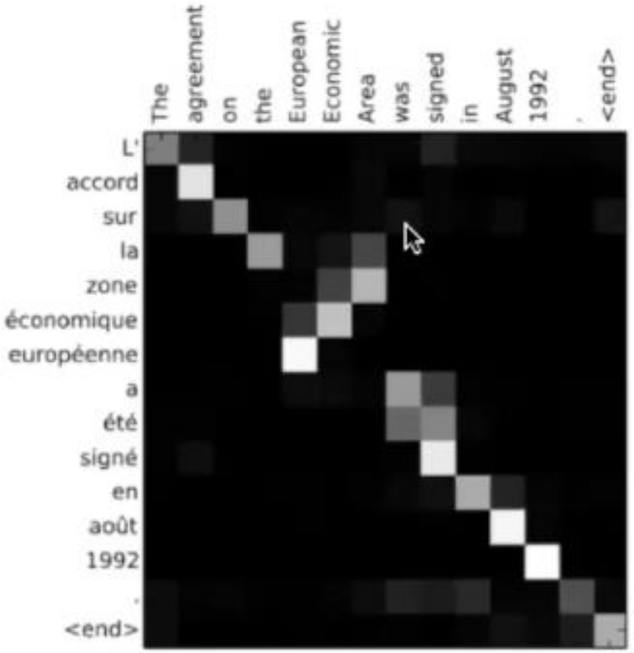
Originally developed for recurrent neural networks

ATTENTION MECHANISM



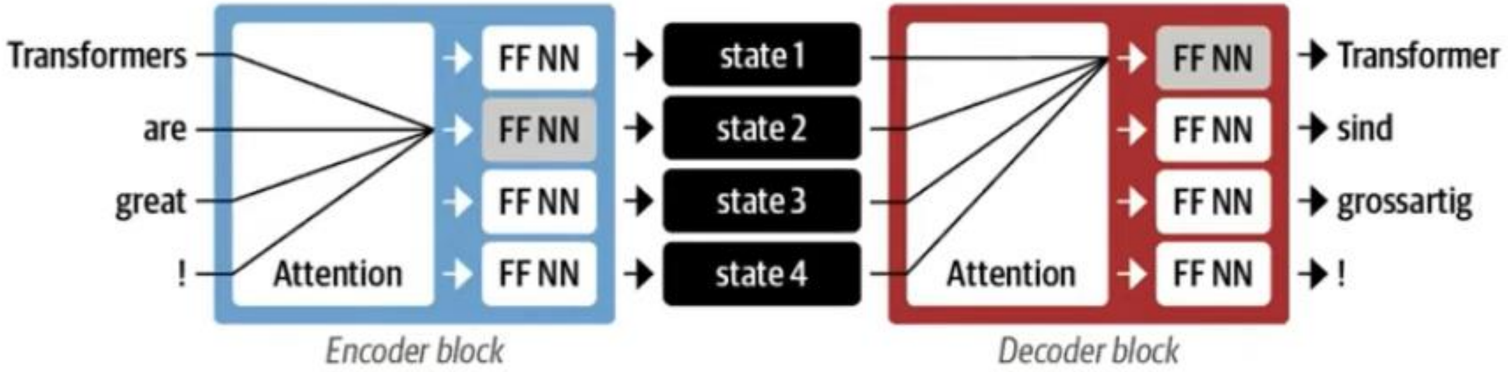
Assign a weight or "pay attention" to specific states

INTERPRETATION



Attention gives better modelling of word order

ATTENTION IS ALL YOU NEED



Transformers much easier to scale with compute & data

BENEFITS

- Solving the bottleneck problem in sequence-to-sequence tasks
- Provides some interpretability
- No Vanishing Gradient Problem
- Multimodality

TODOS FOR THE NEXT WEEK

- Watch video 1 (Self-Attention) and video 2 (Keys, Values, Queries) of the Rasa Series on Transformers & Attention.
- Note at least one question on the each of the videos!
- Do chapter 1 of the Hugging Face course.

🤗 Huggingface Datasets

- https://huggingface.co/datasets

Hugging Face | Search models, datasets, users... | Models | Datasets | Pricing | Resources | Log In | Sign Up

Task Category

conditional-text-generation, text-classification, structure-prediction, sequence-modeling, question-answering, text-scoring, + 3

Task

machine-translation, language-modeling, named-entity-recognition, sentiment-classification, dialogue-modeling, extractive-qa, + 128

Language

en, es, fr, de, ru, ar, + 164

Multilinguality

monolingual, multilingual, translation, other-language-learner

Size

10K<n<100K, 1K<n<10K, n<1K, 100K<n<1M, n>1M, 1k<10K, + 16

License

mit, cc-by-4.0, cc-by-sa-4.0, cc-by-sa-3.0, apache-2.0, cc-by-nc-4.0, + 56

Datasets 638 | Search Datasets | ↑↓ Sort: Alphabetical

acronym_identification

Acronym identification training and development sets for the acronym identification task at SDU@AAAI-21.

annotations_creators: expert-generated | language_creators: found | languages: en | licenses: mit | multilinguality: monolingual | size_categories: 10K<n<100K | source_datasets: original | task_categories: structure-prediction | task_ids: structure-prediction-other-acronym-identification

ade_corpus_v2

ADE-Corpus-V2 Dataset: Adverse Drug Reaction Data. This is a dataset for Classification if a sentence is ADE-related (True) or not (False) and Relation Extraction between Adverse Drug Event and Drug. DRUG-AE.rel provides relations between drugs and adverse effects. DRUG-DOSE.rel provides relations between drugs and dosages. ADE-NEG.txt pro...

annotations_creators: expert-generated | language_creators: found | languages: en | licenses: unknown | multilinguality: monolingual | size_categories: 10K<n<100K | size_categories: 1K<n<10K | size_categories: n<1K | source_datasets: original | task_categories: text-classification | task_categories: structure-prediction | task_categories: structure-prediction | task_ids: fact-checking | task_ids: coreference-resolution | task_ids: coreference-resolution

adversarial_qa

AdversarialQA is a Reading Comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles using an adversarial model-in-the-loop. We use three different models; BiDAF (Seo et al., 2016), BERT-Large (Devlin et al., 2018), and RoBERTa Large (Liu et al., 2019) in the annotation loop and construct three datasets;...

annotations_creators: crowdsourced | language_creators: found | languages: en | licenses: cc-by-sa-4.0 | multilinguality: monolingual | size_categories: 10K<n<100K | source_datasets: original | task_categories: question-answering | task_ids: extractive-qa | task_ids: open-domain-qa

Stanford

Paperswithcode Datasets

- https://www.paperswithcode.com/datasets?mod=texts&page=1

835 dataset results for Texts ×

Penn Treebank
The English Penn Treebank corpus, and in particular the section of the corpus corresponding to the articles of Wall Street Journal (WSJ), is one of the most known and used corpus for t...
1,545 PAPERS • 10 BENCHMARKS

SQuAD (Stanford Question Answering Dataset)
The Stanford Question Answering Dataset (SQuAD) is a collection of question-answer pairs derived from Wikipedia articles. In SQuAD, the correct answers of questions can be any se-...
1,254 PAPERS • 7 BENCHMARKS

Visual Genome
Visual Genome contains Visual Question Answering data in a multi-choice setting. It consists of 101,174 images from MSCOCO with 1.7 million QA pairs, 17 questions per image on aver-...
903 PAPERS • 11 BENCHMARKS

GLUE (General Language Understanding Evaluation benchmark)
General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language understanding tasks, including single-sentence tasks CoLA and SST-2, similarity...
847 PAPERS • 14 BENCHMARKS

SNLI (Stanford Natural Language Inference)
The SNLI dataset (Stanford Natural Language Inference) consists of 570k sentence-pairs manually labeled as entailment, contradiction, and neutral. Premises are image captions fro...
743 PAPERS • 1 BENCHMARK

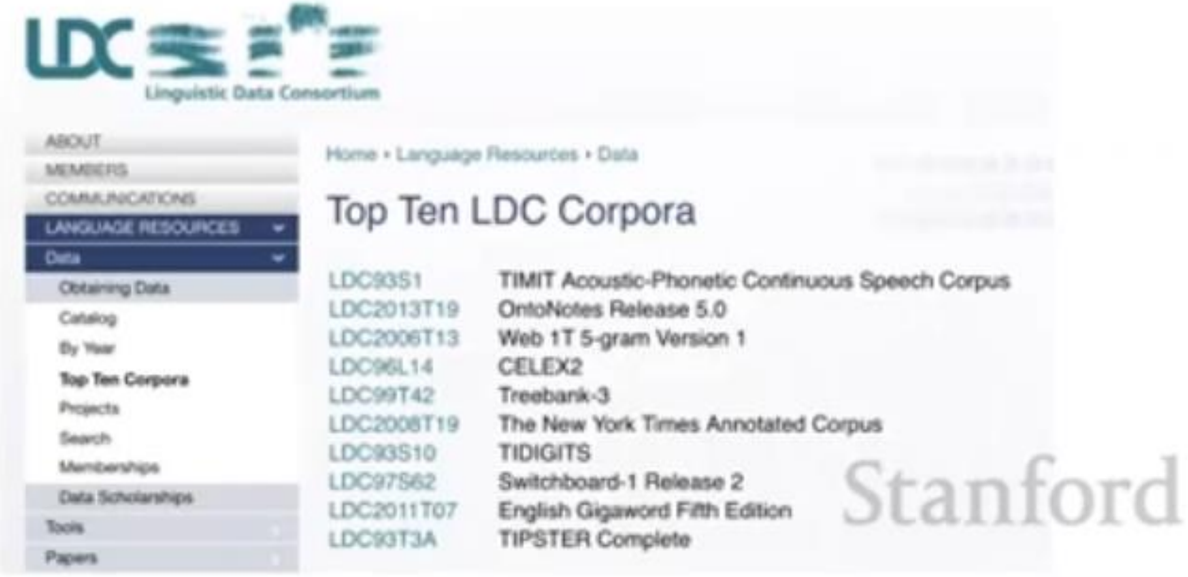
CLEVR (Compositional Language and Elementary Visual Reasoning)
CLEVR (Compositional Language and Elementary Visual Reasoning) is a synthetic Visual Question Answering dataset. It contains images of 3D-rendered objects; each image comes...
528 PAPERS • 1 BENCHMARK

Visual Question Answering (VQA)
Visual Question Answering (VQA) is a dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense...
435 PAPERS • 2 BENCHMARKS

Billion Word Benchmark
The One Billion Word dataset is a dataset for language modeling. The training/held-out data was produced from the WMT 2011 News Crawl data using a combination of Bash shell and...
417 PAPERS • 1 BENCHMARK

Linguistic Data Consortium

- https://catalog.ldc.upenn.edu/
- Stanford licenses data; you can get access by signing up at: https://linguistics.stanford.edu/resources/resources-corpora
- Treebanks, named entities, coreference data, lots of clean newswire text, lots of speech with transcription, parallel MT data, etc.
  - Look at their catalog
  - Don't use for non-Stanford purposes!



Dependency parsing: Universal Dependencies

- https://universaldependencies.org

Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Discussion on UD
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained by the Charles University in Prague
  - Kontext maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
- Download UD treebanks

If you want to receive news about Universal Dependencies, you can subscribe to the UD mailing list. If you want to discuss individual annotation questions, use the Github issue tracker.