

# Advanced Information Retrieval

Introduction 2022

## Our Online Format

- YouTube uploads of recorded lectures
  - 45min to 1 hour each
  - Additionally: PDF slides + automatic & corrected full text transcripts
- Flexible grading structure
- Bi-Weekly online office hours for exercises & lectures
- 2 exercises (no interviews)
- 24h take home exam (2 dates offered)

## Lecturers

- Sebastian Hofstätter
- Sophia Althammer

## Getting in Touch

If you have questions regarding the content of lectures or exercises open a thread on our GitHub discussion platform:

[github.com/sebastian-hofstaetter/teaching/discussions](https://github.com/sebastian-hofstaetter/teaching/discussions)

*(This is our first time trying this out, and we hope it gives a better experience than TUWEL, if not then we'll switch back next semester)*

If you spot an error in either the lectures or exercise assignments create an issue in our GitHub repository:

[github.com/sebastian-hofstaetter/teaching/issues](https://github.com/sebastian-hofstaetter/teaching/issues)

## Problems / Questions / Feedback

Write an email to [advanced-information-retrieval@ec.tuwien.ac.at](mailto:advanced-information-retrieval@ec.tuwien.ac.at)

*(Please don't write us individually, as we have varying workloads during the semester, and you might not get an answer)*

About 46.400.000 results (0,43 seconds)

## Dictionary



# information retrieval

*noun* **COMPUTING**

the tracing and recovery of specific information from stored data.  
"an information retrieval system"



Translations, word origin, and more definitions

[Feedback](#)

## Information retrieval - Wikipedia

[https://en.wikipedia.org/wiki/Information\\_retrieval](https://en.wikipedia.org/wiki/Information_retrieval) ▼

**Information retrieval** (IR) is the activity of obtaining **information** system resources relevant to an **information** need from a collection of **information** resources. Searches can be based on full-text or other content-based indexing.

[Overview](#) · [History](#) · [Model types](#) · [Timeline](#)

## Information Retrieval – Wikipedia

[https://de.wikipedia.org/wiki/Information\\_Retrieval](https://de.wikipedia.org/wiki/Information_Retrieval) ▼ [Translate this page](#)

**Information Retrieval** [ˌɪnfəˈmeɪʃən ɹɪˈtʃiːvəl] (**IR**) oder Informationsrückgewinnung, gelegentlich ungenau Informationsbeschaffung, ist ein Fachgebiet, ...

[Geschichte](#) · [Grundbegriffe](#) · [Relevanz und Pertinenz](#) · [Typologie von ...](#)

## <sup>[PDF]</sup> Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/pdf/01bool.pdf> ▼

**Information retrieval** (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an **information** need from within large collections (usually stored on computers).



# Information retrieval



Information retrieval is the activity of obtaining information system resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. [Wikipedia](#)

[Feedback](#)

# Information Retrieval

a3 size


All Images Maps News Videos More Settings Tools

About 289.000.000 results (0,73 seconds)

**21.0 x 29.7cm**

The **A3 size** print measures 29.7 x 42.0cm, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4 size** print measures 21.0 x 29.7cm, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)  
[https://www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)



mainthebest.com

About Featured Snippets Feedback

a3 size

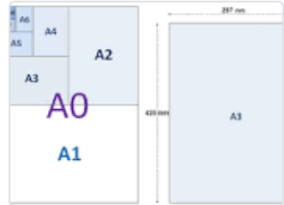
All Images Maps News Videos More Settings Tools


About 646.000.000 results (0,50 seconds)

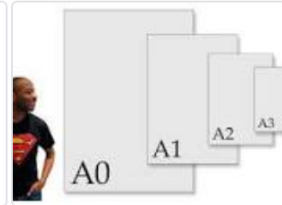
**29.7 x 42.0cm**


The **A3 size** print measures **29.7 x 42.0cm**, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4 size** print measures **21.0 x 29.7cm**, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)  
[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)









About Featured Snippets Feedback

# Information Retrieval

a3 size

AI Images Maps News Videos More Settings Tools

About 289.000.000 results (0,73 seconds)

**21.0 x 29.7cm**

The **A3 size** print measures 29.7 x 42.0cm, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4 size** print measures 21.0 x 29.7cm, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)  
[https://www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)

About Featured Snippets Feedback

Google.com – 14.10.2019

a3 size

AI Images Maps News Videos More Settings Tools

About 646.000.000 results (0,50 seconds)

**29.7 x 42.0cm**

The A3 size print measures **29.7 x 42.0cm**, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The A4 size print measures **21.0 x 29.7cm**, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)  
[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](#)

About Featured Snippets Feedback

Google.com – 3.3.2020

a3 size

AI Images Videos News Shopping More Settings Tools

About 686.000.000 results (0,70 seconds)

**297 x 420 mm**

Paper	mm	inches
A1	<b>594 x 841 mm</b>	23.4 x 33.1 inches
A2	<b>420 x 594 mm</b>	16.5 x 23.4 inches
A3	<b>297 x 420 mm</b>	11.7 x 16.5 inches
A4	<b>210 x 297 mm</b>	8.3 x 11.7 inches

[7 more rows](#)

[www.brother.co.uk](http://www.brother.co.uk) › Support › Brother Answers Articles

[Is A3 bigger than A4? | Printer Paper Sizes | Brother UK](#)

About featured snippets Feedback

Google.com – 11.2.2021

# Information Retrieval

a3 size

AI Images Maps News Videos More Settings Tools

About 289,000,000 results (0,73 seconds)

**21.0 x 29.7cm**

The **A3 size** print measures 29.7 x 42.0cm, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4 size** print measures 21.0 x 29.7cm, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)  
[https://www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)

About Featured Snippets Feedback

Google.com – 14.10.2019

a3 size

AI Images Maps News Videos More Settings Tools

About 646,000,000 results (0,50 seconds)

**29.7 x 42.0cm**

The **A3 size** print measures **29.7 x 42.0cm**, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4 size** print measures **21.0 x 29.7cm**, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)  
[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)

About Featured Snippets Feedback

Google.com – 3.3.2020

a3 size

AI Images Videos News Shopping More Settings Tools

About 686,000,000 results (0,70 seconds)

**297 x 420 mm**

Paper	mm	inches
A1	594 x 841 mm	23.4 x 33.1 inches
A2	420 x 594 mm	16.5 x 23.4 inches
A3	297 x 420 mm	11.7 x 16.5 inches
A4	210 x 297 mm	8.3 x 11.7 inches

7 more rows

[www.brother.co.uk/Support/BrotherAnswersArticles/IsA3biggerthanA4PrinterPaperSizes](https://www.brother.co.uk/Support/BrotherAnswersArticles/IsA3biggerthanA4PrinterPaperSizes) | [Brother UK](https://www.brother.co.uk/Support/BrotherAnswersArticles/IsA3biggerthanA4PrinterPaperSizes)

About featured snippets Feedback

Google.com – 11.2.2021

a3 size

AI Images Shopping News Videos More Tools

Sketchbook Wall Canvas binder Card stock

**297 x 420 mm 29.7 x 42 cm**

The most common and recognised sheet of paper, A4 paper size is 210 x 297 mm.

...

Paper	A3
mm	<b>297 x 420 mm</b>
cm	29.7 x 42 cm
inches	11.7 x 16.5 inches

10 more columns • May 22, 2020

<https://www.brother.co.uk/Support/BrotherAnswersArticles/IsA3biggerthanA4PrinterPaperSizes> | [Brother UK](https://www.brother.co.uk/Support/BrotherAnswersArticles/IsA3biggerthanA4PrinterPaperSizes)

About featured snippets Feedback

Google.com – 10.3.2022

# Machine Learning

---



**Mat Velloso** @matvelloso · 22 Nov 2018

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI



193



7.9K



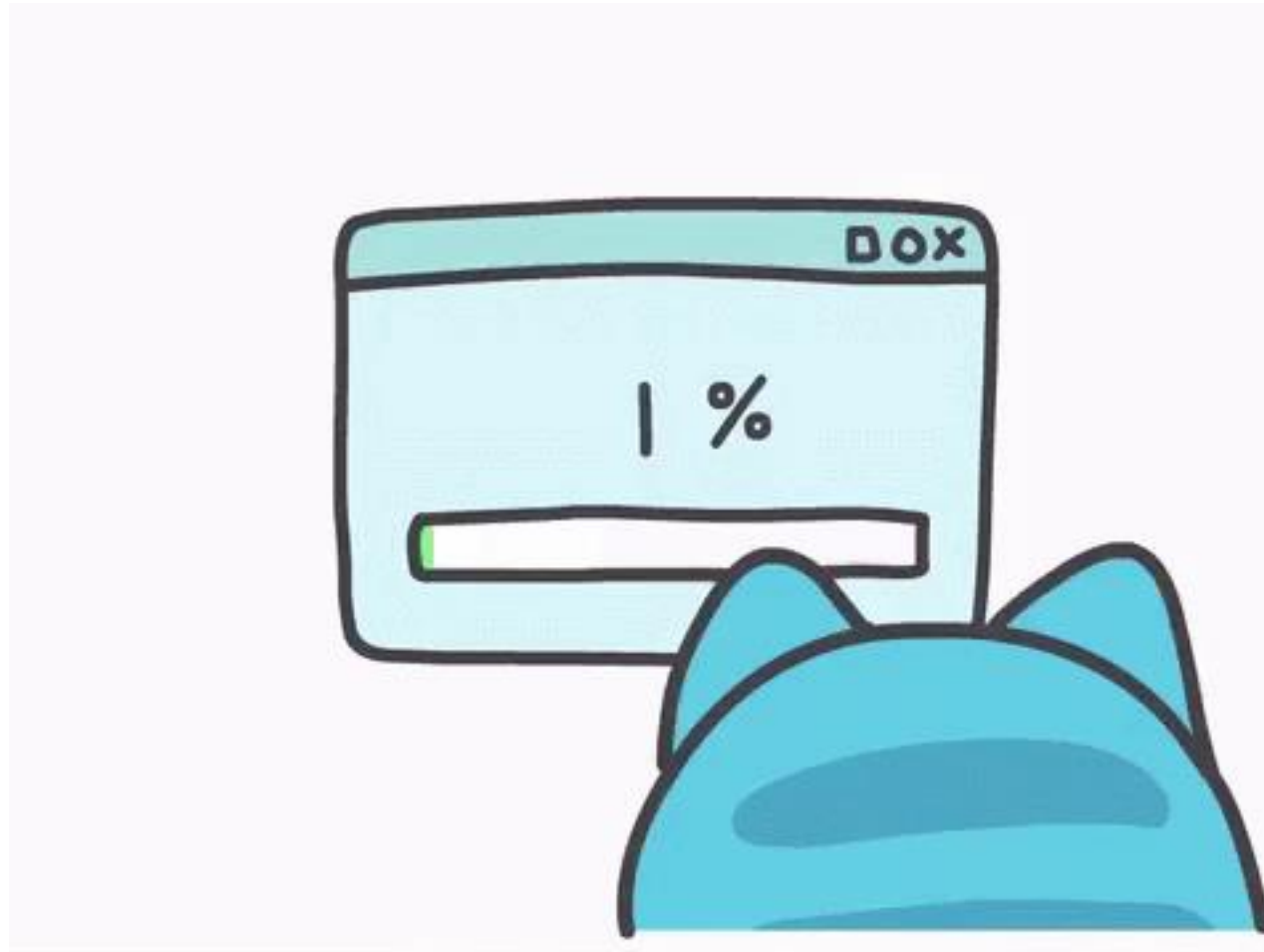
22K





# Machine Learning

---



# Recommended Prerequisites

---

- Machine Learning know how
  - Know the basic concepts
  - Experience with a neural network framework (PyTorch, TensorFlow, etc...)
  - Experience in reading academic papers
- Basic IR course
  - Always good, but we will revisit the basics
- Good programming skills
- Available Nvidia GPU (or alternatively a free GPU from Google Collab)

# Some pointers to get you started ...

---

- Neural Network Methods in Natural Language Processing by Yoav Goldberg
  - Contains a good introduction to ML as well
- Pretrained Transformers for Text Ranking: BERT and Beyond by Lin et al. <https://arxiv.org/abs/2010.06467>
  - Survey of neural IR progress starting in 2019
- Google crash course on ML <https://developers.google.com/machine-learning/crash-course/ml-intro>
- PyTorch Tutorials <https://pytorch.org/tutorials/>
- AllenNLP Tutorials <https://allennlp.org/tutorials>

# The Works

All Things: Syllabus, Lectures, Exercises, Exam, Grading

# Syllabus

---

## 1 Crash Course IR

- **Fundamentals:** Inverted index & probabilistic scoring (BM25)
- **Evaluation:** List-based measures (binary & graded relevance)
- **Test Collections:** Create and analyze IR datasets

## 2 Representation Learning (NLP)

- **Word Embeddings:** Basic building blocks & intro to vector representations
- **Sequence Representations:** Contextual vectors with: CNNs, RNNs, & (pre-trained) Transformers
- **Extractive QA:** Find answer location in text

## 3 Neural IR

- **Re-ranking:** From early beginnings of neural re-ranking to pre-trained BERT
  - **From scratch:** early IR specific re-ranking models
  - **Efficient Transformers:** Transformer-Kernel family
  - **State-of-the-art:** Large BERT-based models
- **Domain-Specific:** Caveats and task-changes between passage/document + web and legal/patent domains
- **Retrieval:** Encoding passages into single vectors; directly retrievable with embedded query rep from nearest neighbor index
- **Knowledge Distillation:** Improving the training of efficient architectures with the help of slow, but good models

# Syllabus

---

## 1 Crash Course IR

- **Fundamentals:** Inverted index & probabilistic scoring (BM25)
- **Evaluation:** List-based measures (binary & graded relevance)
- **Test Collections:** Create and analyze IR datasets

## 2 Representation Learning (NLP)

- **Word Embeddings:** Basic building blocks & intro to vector representations
- **Sequence Representations:** Contextual vectors with: CNNs, RNNs, & (pre-trained) Transformers
- **Extractive QA:** Find answer location in text

## 3 Neural IR

- **Re-ranking:** From early beginnings of neural re-ranking to pre-trained BERT
  - **From scratch:** early IR specific re-ranking models
  - **Efficient Transformers:** Transformer-Kernel family
  - **State-of-the-art:** Large BERT-based models
- **Domain-Specific:** Caveats and task-changes between passage/document + web and legal/patent domains
- **Retrieval:** Encoding passages into single vectors; directly retrievable with embedded query rep from nearest neighbor index
- **Knowledge Distillation:** Improving the training of efficient architectures with the help of slow, but good models

# Syllabus

---

## 1 Crash Course IR

- **Fundamentals:** Inverted index & probabilistic scoring (BM25)
- **Evaluation:** List-based measures (binary & graded relevance)
- **Test Collections:** Create and analyze IR datasets

## 2 Representation Learning (NLP)

- **Word Embeddings:** Basic building blocks & intro to vector representations
- **Sequence Representations:** Contextual vectors with: CNNs, RNNs, & (pre-trained) Transformers
- **Extractive QA:** Find answer location in text

## 3 Neural IR

- **Re-ranking:** From early beginnings of neural re-ranking to pre-trained BERT
  - **From scratch:** early IR specific re-ranking models
  - **Efficient Transformers:** Transformer-Kernel family
  - **State-of-the-art:** Large BERT-based models
- **Domain-Specific:** Caveats and task-changes between passage/document + web and legal/patent domains
- **Retrieval:** Encoding passages into single vectors; directly retrievable with embedded query rep from nearest neighbor index
- **Knowledge Distillation:** Improving the training of efficient architectures with the help of slow, but good models

# Lectures / Content

---

- Find extra information here:
  - Introduction to IR slides
  - Additional lecture notes

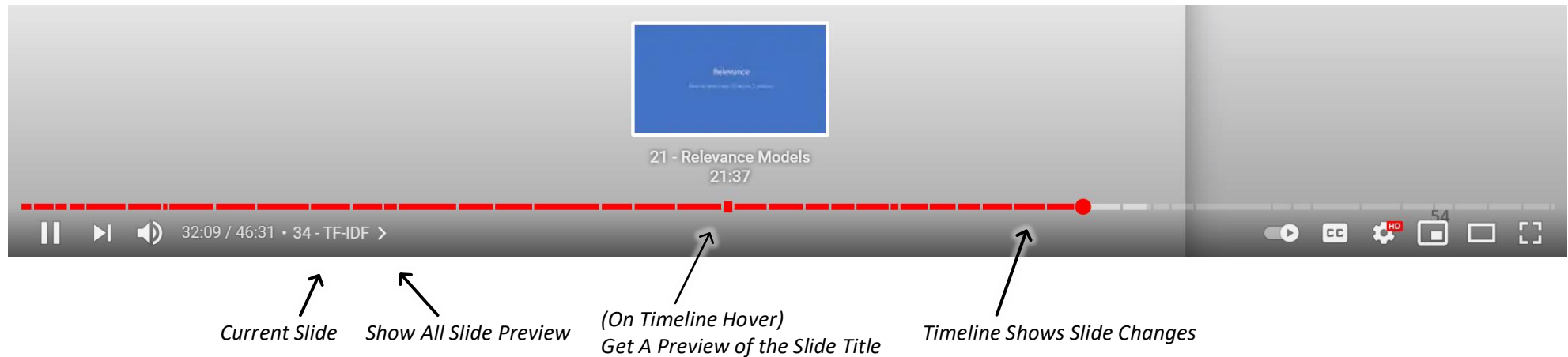
[github.com/sebastian-hofstaetter/teaching](https://github.com/sebastian-hofstaetter/teaching)

- Star it + add your content via issues & pull requests (🎁 bonus points!)
  - Fix automatic closed captions (5 bonus points for each lecture)
  - Lecture notes, summaries, examples, bugfixes ... (generous bonus points)



# Student Experience - YouTube

---



- YouTube converts timestamps into user interface improvements in the video player across devices.
  - These improvements move the YouTube player closer to specialized slide-show players, but with the benefit of YouTube's scale and ease of use + discoverability

# Student Experience - Transcript

- Our aim is to enable fine-grained navigation
  - Every resource is slide-based navigable
- Transcribing per slide allows us to produce well-formatted transcripts
  - Azure Speech outputs mostly correct capitalization & punctuation
- Text format allows for re-use

## Lecture 10 - Dense Retrieval and Knowledge Distillation

*Automatic closed captions generated with the Azure Speech API*

1

Hi and welcome everyone. Today's lecture is about dense retrieval and knowledge distillation. My name is Sebastian, and as always, if you have any questions, please feel free to contact me. Let's get started.

16.97 seconds

2 Today

*Extract Slide  
Number & Title*

*Correctly Recognized with  
Text Hints from Slide Text*

Today we're talking about a couple of different things, so this lecture is jam-packed with information. First, we're going to look at dense retrieval, what that means, which types of model, specifically the BERT DOT retrieval model, is used for that, and how approximate nearest neighbor search brings everything together, and allows for a search engine. Then the second part is about knowledge distillation and specifically about our work in the last couple of months, where we looked at cross architecture, distillation losses such as Margin-MSE or our recently proposed TAS-Balanced with dual supervision training approach for dense retrieval models. And then finally, we're going to look at how we can analyze dense retrieval models and what those analyses mean for the future development of those models.

68.03 seconds

*Automatic Punctuation Transcription*

*Fixed by a Student*

3 Neural Methods for IR Beyond Re-ranking

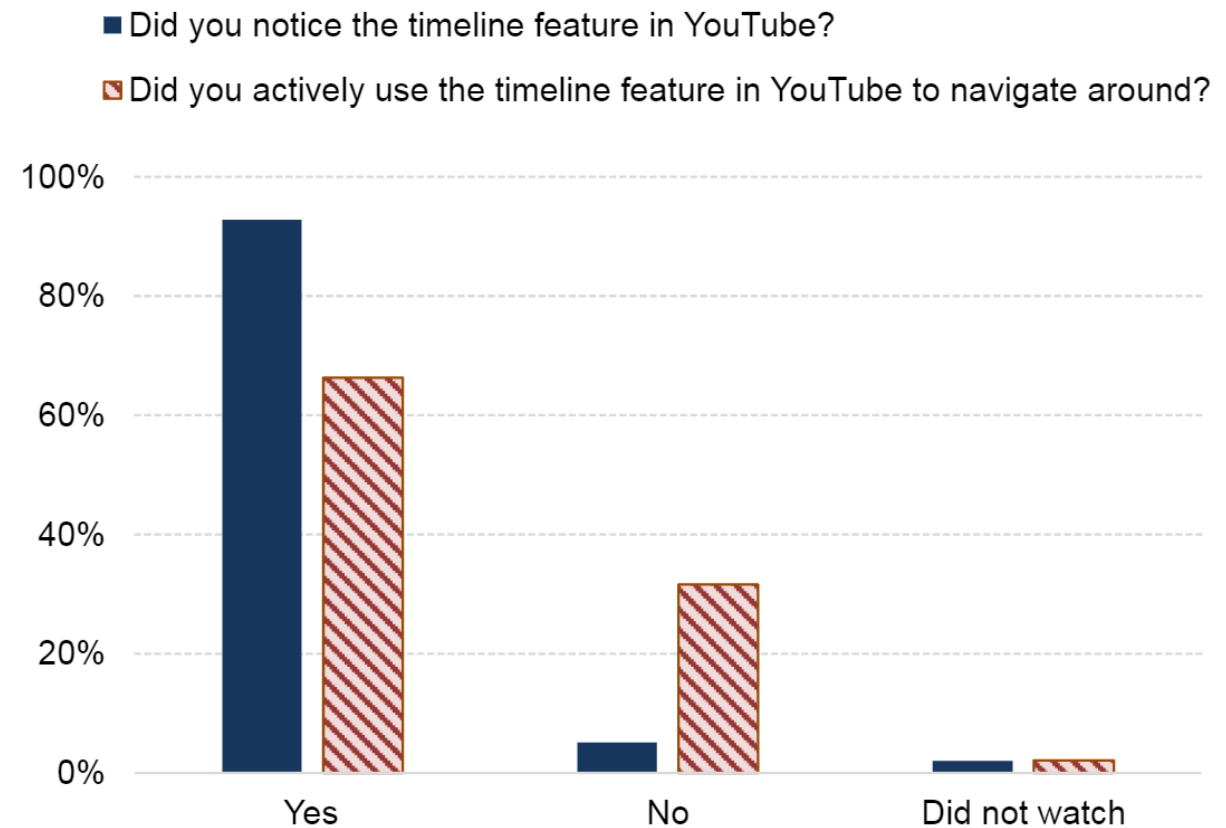
Why are we talking about dense retrieval? We want neural methods for information retrieval to work beyond re-ranking models that we saw in previous lectures. We want to bring neural advances in the first stage phase of the ranking pipeline to remove the bottleneck of an inverted index with BM25. And today we look at dense retrieval as one such alternative. Although before we start, I want to acknowledge there are a lot of other very cool techniques how neural approaches can improve first stage retrieval, and some of them are listed here. So in very popular approaches like Doc2query, which actually does document expansion with potential query texts that would semantically match a document. And here you then still use the inverted index with the expanded documents and use a traditional BM25 function. But because the document text has more words in it, it gives you better results. Another option is DeepCT, which assigns term weights based

...

# Feedback 2021 - YouTube Usage

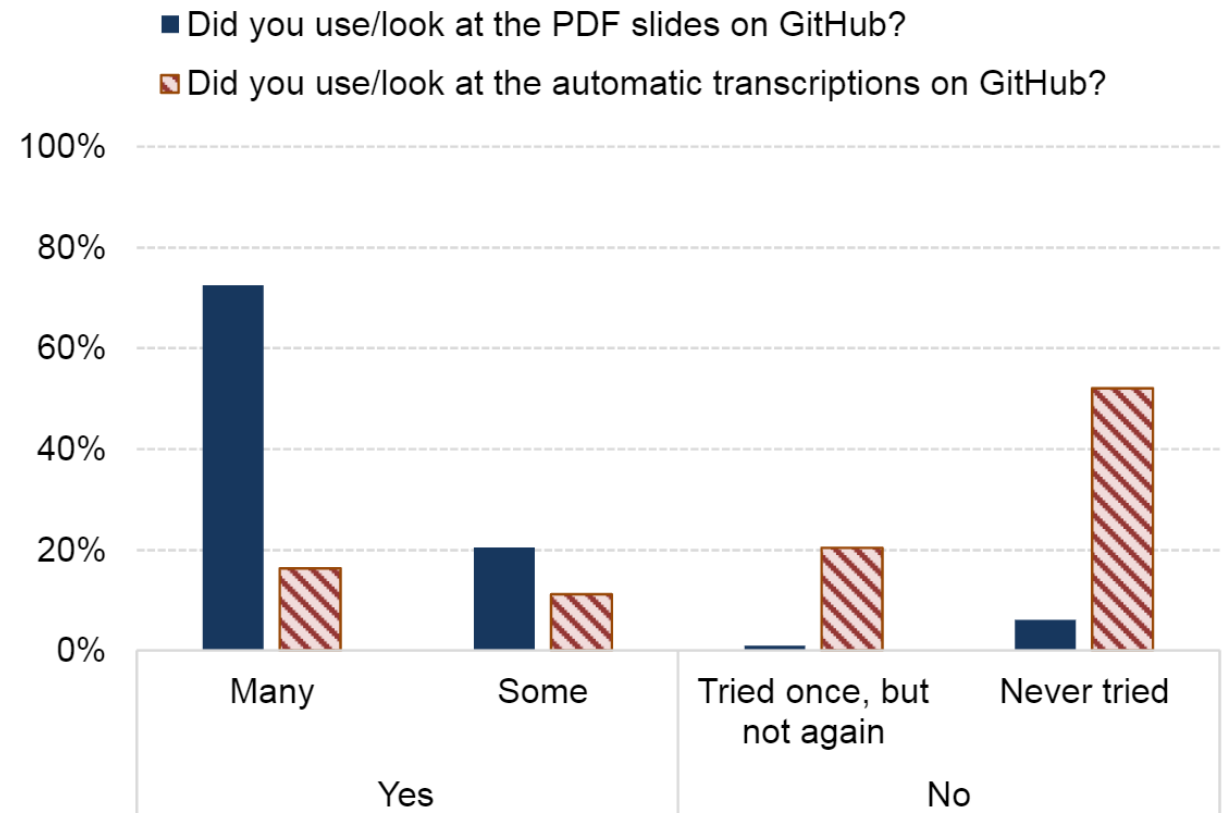
---

- 83% favor GitHub and YouTube over comparable self-hosted university services
  - And almost all students watched the videos
- 93% noticed the timeline feature UI
- 66 % used the timeline feature actively



# Feedback 2021 - GitHub & Transcript Usage

- 92% of students regularly used the PDFs on GitHub
- Transcripts were only used many times by 16% and sometimes by 11%
  - Is this bad?
  - Room for improvement: 52% of students never tried to look at the transcripts



# Exercises

---

## ① Data annotation

- Understand the task, that we want to teach the machine
- Create testing & analysis data for exercise 2

## ② Neural Re-ranking & Extractive QA

- Using Python & PyTorch
- **Part 1:**
  - Implement & train neural re-ranking models from scratch
- **Part 2:**
  - Use pre-trained models from HuggingFace to create an extractive QA system

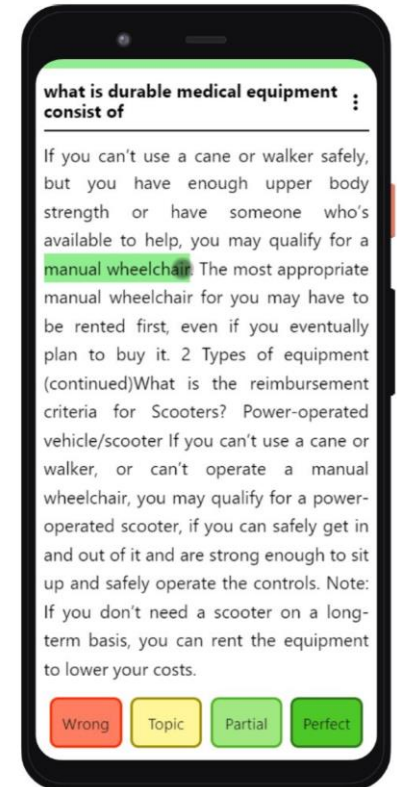
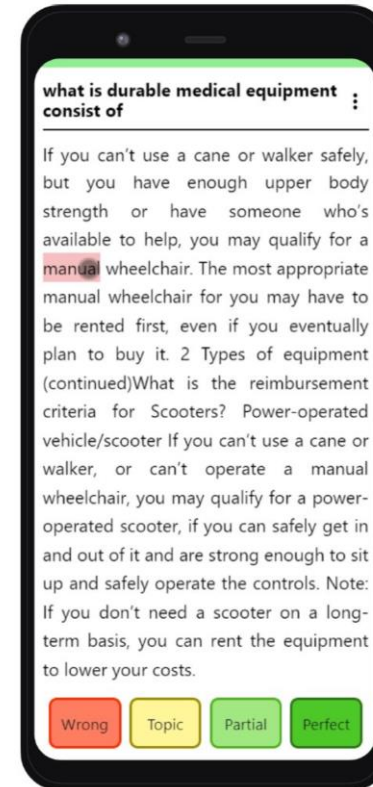
# Exercise 1

---

- Creating annotations is time consuming;  
easier to split the task among many people
  - Each student spends a few hours (500 annotations for 100%, 250 min.)
- We create a fine-grained passage retrieval and extractive QA dataset, based on MSMARCO
  - This fine-grained data over a lot of queries doesn't exist yet
- We use this dataset in Exercise 2 for evaluation and exploration
  - Potentially we also share it with the research community
  - Data is completely anonymized before publication

# Exercise 1 - FiRA

- We created a specialized & simple to use tool (FiRA) for mobile and desktop use
- Each registered student receives pre-created account information via email
- More annotations = bonus points 🏆
  - +4% of the total grade per 100 extra annotations
  - The bonus points are unlimited
  - > 1.000 annotations remove min. point requirement of exercise 2 & exam



# Exercise 2

---

- 2 Parts (both must be done):
  - ① Implement & train neural re-ranking models from scratch
    - Not state of the art, but teaches you how PyTorch works from the ground up and allows you to learn to work with training loops, loss functions, and tensor operations
  - ② Use pre-trained models from HuggingFace to create an extractive QA system
    - Now, we don't train a model, but download a pre-trained model; and put together the pipeline necessary to get google-like results (example from the beginning)



# Exercise 2

---

- Exercise 2 in groups
  - 4 persons per group (managed via TUWEL)
  - Will be evaluated together
- Work in 1 private GitHub repository (via GitHub classroom)
- Lots of bonus point opportunities 🎉 🙌 100 ✨
  - For creative extra work, if you have fun doing it, go for it and you'll get points!
  - For finding & fixing bugs in the starter code or lecture slides

# Online Exam

---

- 24-hour take-home exam
- Type: Paper reading with questions showing the understanding of it
  - You must answer both easier & more complicated questions
  - Relates to one or more lectures
  - Open-Book
- Exam on: **1.6.** & **15.6.** Starting at noon (12:00)
  - Pick **one** date (do-overs only for failed attempts or technical problems)
  - Administered completely via TUWEL Test

# Grading

---

Exercise 1 (Annotation):	10%	(min 50% to pass)
Exercise 2 (Neural IR):	50%	(min 50% to pass)*
Exam:	40%	(min 30% to pass)*
<hr/>		
<b>Total</b>	<b>100%</b>	<b>(min 50% to pass)</b>

## Grading Scheme (TUWEL defaults)

1:  $\geq 88,00$     2:  $\geq 75,00$     3:  $\geq 63,00$     4:  $\geq 50,00$

\* Can be removed with enough bonus points in Exercise 1

# See you next week – virtually 🖐️

For feedback, problems, or any other issues, please write to:  
[advanced-information-retrieval@ec.tuwien.ac.at](mailto:advanced-information-retrieval@ec.tuwien.ac.at)