Von Maschinen und Menschen, Bibliotheken und Intelligenz

Prof. Dr. Thorsten Koch

kobv

KOBV Forum, Juni 2021, virtuell

ZIB
ZUSE INSTITUTE BERLIN

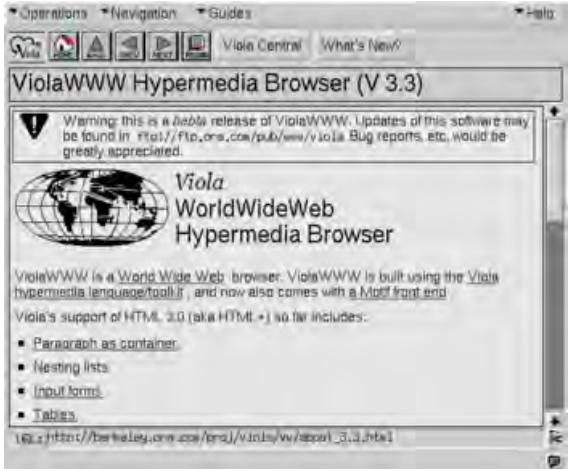
Wie Ideen sich entwickeln

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

Was war die ursprüngliche Idee hinter HTML?

(Hyper Text Mark-up Language)

```
<h1>Überschrift</h1>
<p>und hier kommt <em>jetzt der</em> Inhalt </p>
<ul>
<li>Wichtiger Punkt</li>
<li>Nicht so wichtig</li>
</ul>
```



https://en.wikipedia.org/w/index.php?curid=6098546

Eine Vision was passiert, wenn die Werber die Welt übernehmen:

F. Pohl, C.M. Kornbluth: *The Space Merchants* (1952), <dt.> *Eine Handvoll Venus und ehrbare Kaufleute* (1973)

As with many significant works of science fiction, it was lexically inventive. The novel is cited by the Oxford English Dictionary as the first recorded source for a number of new words, including “soyaburger”, “moon suit”, “tri-di” for “three-dimensional”, “R and D” for “research and development”, “sucker-trap” for a shop aimed at gullible tourists, and one of the first uses of “muzak” as a generic term. It is also cited as the first incidence of “survey” as a verb meaning to carry out a poll. — https://en.wikipedia.org/wiki/The_Space_Merchants

Die LoC hat etwa 38 Millionen Bücher

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

38.000.000 Bücher × 250 Seiten × 2 KB = 500 KB · 38 · 10^6 = 19 TB

Sie können jetzt **eine** Festplatte kaufen, auf der die Texte aller Bücher der LoC passen.

(und ihr Laptop zu Hause braucht mindestens einen Tag um sie komplett durchzulesen)

Festplatte
Exos X18
18 TB ≈ €800

60 % des gewichteten Pre-Training-Datensatzes für GPT-3 stammen aus einer gefilterten Version von Common Crawl, die aus 410 Milliarden Bytepaar-kodierten Token besteht. Weitere Quellen sind 19 Milliarden Token aus WebText2, was 22 % der gewichteten Gesamtmenge entspricht, 12 Milliarden Token aus Books1, was 8 % entspricht, 55 Milliarden Token aus Books2, was 8 % entspricht, und 3 Milliarden Token aus Wikipedia, was 3 % entspricht. GPT-3 wurde auf Hunderten von Milliarden von Wörtern trainiert [...]. Da die Trainingsdaten von GPT-3 allumfassend waren, benötigt es kein weiteres Training für bestimmte Sprachaufgaben. Die Trainingsdaten enthalten gelegentlich toxische Sprache und GPT-3 generiert gelegentlich toxische Sprache als Ergebnis der Nachahmung seiner Trainingsdaten.

In einem ersten Experiment wurden 80 US-amerikanische Probanden gebeten zu beurteilen, ob kurze ~200-Wort-Artikel von Menschen oder GPT-3 geschrieben wurden. Die Teilnehmer urteilten in 48 % der Fälle falsch und schnitten damit nur geringfügig besser ab als beim zufälligen Raten.

Aus https://en.wikipedia.org/wiki/GPT-3 übersetzt mit www.DeepL.com/Translator

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

Generative Pre-trained Transformer 3 (GPT-3)

https://en.wikipedia.org/wiki/GPT-3

Bidirectional Encoder Representations from Transformers (BERT)

https://en.wikipedia.org/wiki/BERT_(language_model)

Texte generieren mit GPT-3: Die eloquenteste KI der Welt (SPON)

https://www.spiegel.de/netzwelt/web/gpt-3-die-eloquenteste-kuenstliche-intelligenz-der-welt-a-dd3b3423-d214-4a2f-bc51-d51a2ae22074

GPT-3 und andere Sprachmodelle : Nimmt uns der Computer die Sprache ab? (FAZ)

https://www.faz.net/aktuell/feuilleton/debatten/die-informatikerin-sina-zarriess-ueber-sprachmodelle-wie-gpt-3-17070713.html

Lyrik aus der Maschine: Der Textgenerator GPT-3 schreibt Werke im Stile Oscar Wildes, übersetzt Fremdsprachen und chattet mit Menschen über Gott und die Welt. Wie geht das? (SZ)

https://www.sueddeutsche.de/wissen/gpt-3-ki-kuenstliche-intelligenz-computerlinguistik-informatik-1.5122050

KI: Sprachmodelle wie GPT-3 könnten völlig neue Suchmaschinen ermöglichen (Heise)

https://www.heise.de/hintergrund/KI-Sprachmodelle-wie-GPT-3-koennten-voellig-neue-Suchmaschinen-ermoeglichen-6048582.html

GPT-3: Forscher zeigen Vorurteile in riesigem Sprachmodell auf

https://www.golem.de/news/gpt-3-forscher-zeigen-vorurteile-in-riesigem-sprachmodell-auf-2101-153658.html

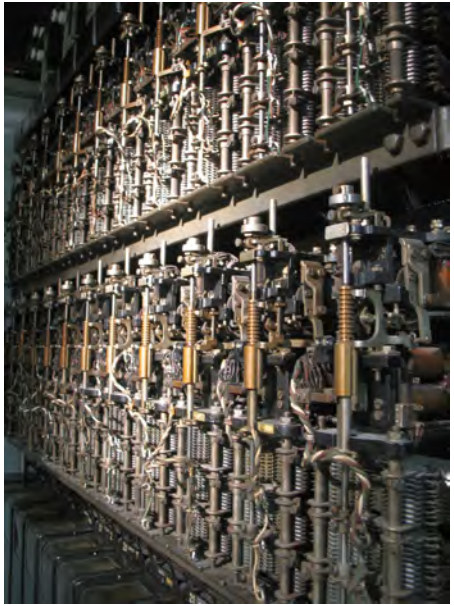
Umstrittener Artikel: Bender, Gebru, et. al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

Veränderungen in der Informationsvermittlung (dauern lange)

Neue Osnabrücker Zeitung vom 29.11.2017
Vor genau 50 Jahren wurde die „Handvermittlung“ im Fernmeldeamt Lingen durch den „Selbstwähldienst“ ersetzt. An das Verstummen des „Fräuleins vom Amt“ erinnert am 6. Dezember eine Veranstaltung in der Reihe „Mittwochs im Museum“. Foto: Privatfoto/Emslandmuseum Lingen

https://www.noz.de/lokales/lingen/artikel/985346/vor-50-jahren-verstummte-in-lingen-das-fraeulein-vom-amt



Von © Túrelio (via Wikimedia-Commons), CC BY-SA 3.0 de, https://commons.wikimedia.org/w/index.php?curid=4066629

Von Mudares 22:13, 8 August 2007 (UTC) - own work (own picture, taken in a Central Office in ~1999), CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=2536266

Handvermittlung bei Telefongesprächen gab es von ca. von 1880-1980

1908 in Hildesheim das erste automatische Ortsamt.

1967 Ende der Handvermittlung, automatische Vermittlung seit 1997 in Deutschland voll digital.

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

Wie	Wo	Tiefe Lieferkette	Entfernung
Selber machen	Zu Hause	0	Lokal
Handwerksbetrieb	Dorf	1-2	Regional
Manufaktur	Stadt	1-3	Überregional
Fabrik (classic)	Industriegebiet	groß	International
Build-to-order	Industriegebiet	größer	International

https://commons.wikimedia.org/w/index.php?curid=20214730

Individualität – Qualität – Quantität – Innovativität – Preis

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

- Was wird von Hand gemacht und was automatisch?
- Was ist individuell und was ist „Massenprodukt“?
- Wo wird eine Arbeit („Rechenleistung“/Datenverarbeitung) durchgeführt?

https://billaltmann.com/tag/ibm-3270/

CC BY-SA https://commons.wikimedia.org/w/index.php?curid=51833

Erkenntnisse

- Die Informationen liegen zunehmend frei im Internet (arXiv).
- Wissenschaftliche Veröffentlichungen hinter Paywalls haben keine Zukunft, da keinen Mehrwert!

- Ohne die großen Suchmachienen (insb. Google) würden wir schon jetzt wenig im Netz finden.
- Die Datenmenge steigt und es wird immer schwieriger, Relevantes von Irrelvantem zu unterscheiden.
- In diesem Bereich wird sich noch viel entwickeln und es wird sehr schwer werden, bei der Informationssuche mit den Big Playern zu konkurrieren.

- Die automatische Verarbeitung von Daten wird noch ganz neue Formen annehmen: Übersetzung, Zusammenfassung, Recherche, ...
- Manuelles Kuratieren von Daten wird zunehmend sinnlos. Es sind zu viele.
- Automatisches Daten-Management!

Ablauf Publikation

kobv
Kooperativer Bibliotheksverbund Berlin-Brandenburg

1. **Artikel schreiben**
   Nicht vergessen, die OrcIDs und Affiliations richtig einzutragen.
2. **Interne Durchsicht**
   - ▷ Zustimmung des Industriepartners einholen (wenn nötig)
   - ▷ Zustimmung aller Koautoren zur Veröffentlichung einholen
   - ▷ Zum Sekretariat zur Durchsicht, Grammarly benutzen
3. **Preprint hochladen, Einträge in den Publikationsdatenbanken**
   - ▷ ZIB Opus, TU DepositOnce, arXiv, Optimization Online
   - ▷ Alle Artikel von TU-Autoren ins FG Publikationsverzeichnis stellen
4. **Beim Journal einreichen**
   Bei der Journalauswahl OA und DEAL berücksichtigen
5. **Revisionen**
6. **Wenn veröffentlicht, alle Datenbankeinträge aktualisieren**
   - ▷ Email zum AG Leiter für die den ZIB-Newsletter
   - ▷ Update: ZIB OPUS, TU DepositOnce, arXiv, Linf (Mail zum Sekretariat), FG Publikationsverzeichnis

Forschungsdaten-Management?

- Gute wissenschaftliche Praxis unterstützen
  -> Forschungsdatenmanagement

- Publikationsflut
  -> Qualitätskontrolle mehr in den produzierenden Institutionen

- **Sicherstellen der dauerhaften Verfügbarkeit der Publikationen und Daten der eigenen Institution.**

- Verteilung von Artikeln:
  Warum läuft der Publikationsprozess nicht über die Bibliothek?

Vielen Dank!

Fragen?