

Textual Data Analysis

Lecture 1 - Introduction



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**

/ Practicalities

Slides, recordings, links, exam dates, etc.: course Moodle

- <https://moodle.utu.fi/course/view.php?id=33776>

Communication: Discord

- <https://discord.gg/8tA9pZasf2> channel **#textual-data-analysis-2025**

Rolling feedback: (tell us anything, anonymously)

- <https://link.webpolsurveys.com/S/6387D0610C74B692>

Attendance:

- Mark your physical in-class attendance in Moodle for a little extra boost in the exam (successfully piloted in the DL-in-HLT course in 2024)

/ Context of this course

(1) Introduction to Human Language Technology

- Basics of text data processing and NLP
- Foundation for later courses

(2) Deep Learning in Human Language Technology

- Technical course on modern DL methods used in NLP
- Introduces the fundamental models in NLP
- Less about applications, more about inner workings and training of models

(3) Textual Data Analysis (this course)

- Builds on previous two courses, going deeper into selected applications
- Geared towards learning to **apply** models on textual data
- Does **not** go into inner workings of the models

/ Context of this course

The **Introduction to HLT** and **Deep Learning in HLT** courses are strongly recommended prerequisites for this course

(Show of hands who has taken these)

We'll present some **required basic information** on this introduction lecture, you can consider this a gentle reminder if you've taken these courses

(3) Textual Data Analysis (this course)

- Builds on previous two courses, going deeper into selected applications
- Geared towards learning to **apply** models on textual data
- Does **not** go into inner workings of the models

/ Why is this course relevant?

NLP and its applications are at the forefront of AI

- ChatGPT, text-to-image and text-to-video models all build on model architectures originating within NLP
- Web search (Information Retrieval) increasingly powered by deep learning
- State-of-the art text classification and text mining applications build on language models
- Deep learning models used throughout the AI industry originate from NLP

→ **This knowledge is sought after, and presently very well paid too :)**

/ Course content

Introduction → search → analysis → generation

1. Introduction and data sources
2. Text indexing, information retrieval, and semantic search
3. Named entity recognition, sequence labeling
4. Text classification and explanation methods
5. Information extraction, relations and events
6. Large language models in practice
7. Advanced LLM applications
8. [no lecture – NLP conference]
9. Question answering and retrieval-augmented generation

/ Course projects

Compulsory!

Well-executed project including optional parts → +1 to course grade

Demos (Wed every week) will support project step by step

- You can compile all these into the project, and add a bit :)

This year's project topic: TBA

Selected tasks and applications



TURKUNLP
.ORG



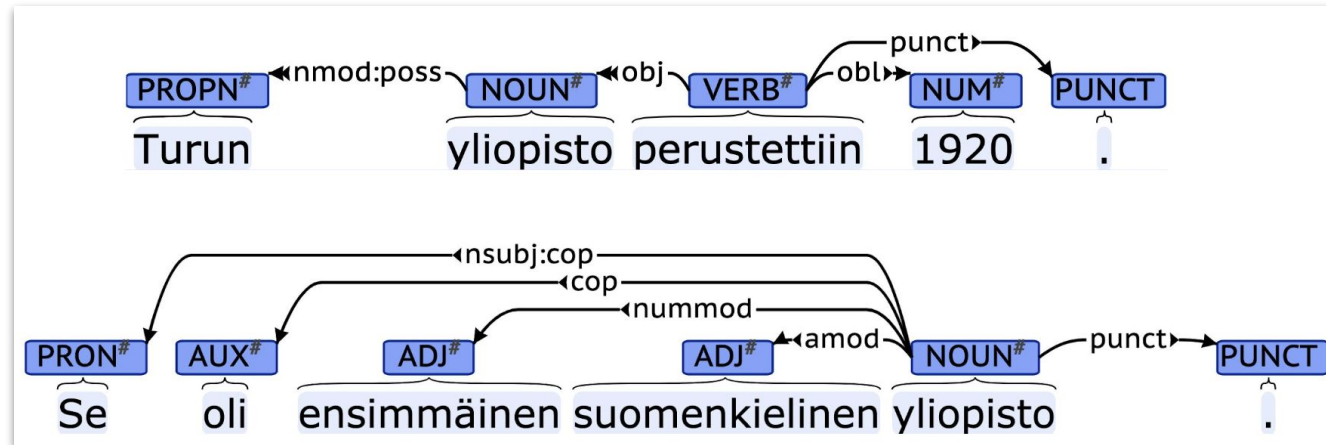
**UNIVERSITY
OF TURKU**

/ Basic text processing and IR

Text parsing and lemmatization

Error correction

Query expansion



/ Named entity recognition and IE

Legend

LANGUAGE

PERSON

GPE

LOC

ORG

EVENT

DATE

Turun yliopisto (lyhenne TY tai UTU) on ensimmäinen täysin suomenkielinen yliopisto, joka perustettiin 1920. Elokuussa 2019 yliopiston rehtorina aloitti Jukka Kola. Yliopiston viimeisenä kanslerina ennen instituution lakkauttamista toimi vuosina 2010–2013 Pekka Puska. Yliopiston Turun kampus sijaitsee kaupungin keskustan alueella Yliopistonmäen läheisyydessä. Yliopisto toimii Turun lisäksi Raumalla (lastentarha-, luokanopettaja- ja käsityöopettajakoulutus) ja Porissa (kauppakorkeakoulu, kulttuurituotanto ja maisemantutkimus). Turun Akatemia Turkuun perustettiin jo vuonna 1640 yliopisto, Turun Akatemia, joka on nykyisen Helsingin yliopiston edeltäjä. Se oli kolmas Ruotsin valtakuntaan perustettu yliopisto ja Suomen ainoa. Turun palon jälkeen vuonna 1827 Akatemia siirrettiin Helsinkiin. Akatemian siirtämisen jälkeen yliopistollinen opetus Turussa loppui vuoteen 1917 asti.

/ Question answering

mikä on maltan pääkaupunki

All Images Videos News Short videos Web Books : More

Malta › Capital :

Valletta



/ Semantic search

les femmes doivent avoir les mêmes droits et la même éducation que les hommes



72

1786 - Condorcet, Jean-antoine-nicolas De Caritat (1743-1794)

Vie de Monsieur Turgot

justes & claires & simples & claires que l'application ne l'auroit jamais qu'une opération de combinaisons & de calcul. (Z59) la femme égale en tout l'homme, doit jouir absolument des mêmes droits si on se rappelle dans quelques limites étroites le droit de la société sur la liberté des individus doit être rétabli on trouver

72

1798 - Rush, Benjamin (1745-1813)

Essays, literary, moral & philosophical

concur in all our plans of education for young men, or no laws will ever render them effectual. To qualify our women for this purpose, they should not only be instructed in the usual branches of female education, but they should be taught the principles of liberty and government; and the obligations of patriotism should be inculcated upon them. The opinions and conduct of men are often regulated by the women in



/ Text-to-text tasks (generation)

Machine translation

Summarization

Chatbots

Closed-book QA

... increasingly, any task where the input and output can be cast as text (“just ask GPT”)



Correct the OCR errors in the following passage: HIS MAJESTIES oECLARATION

'CONCERNINQ The TR E A T y and 1113 Diflike of the eflrmie: p1oceedings, (ye. Hen large pretences prove but the ihadowes of weakc perfdrmantes, then the greateft when aperiod is put to aworke of great concernment, all mens cars do (as it were) hunger till they are fatif- , ' fled 1n their expefiati'ons. Hath nor this diflra111 ed Na~ tion groaned a long time under the burden of tyranny, and oppreflion P and hath not all the blood that hath- been fpilt thefe feven yeares been cafi upon My head, who am the greatefi Suffcrer, though the leaf't guilty. 1' And was it 110t requifite to endeavotu the? flopping of that flux , which , if 1101 {'top'1.,wzllbring- an abolute dei'nuaion to this Nation. 1' And What more A 2 ' fpecdy' "1 11 111111; labouts produce the finallefi elYe€ts and'



HIS MAJESTY'S DECLARATION CONCERNING The TREATY and the DISLIKE of the ARMY:

proceedings, (ye. When large pretenses prove but the shadows of weak performances, then the greatest when a period is put to a work of great concernment, all men's hearts do (as it were) hunger till they are satisfied in their expectations. Hath not this distressed Nation groaned a long time under the burden of tyranny, and oppression? And hath not all the blood that hath been spilled these seven years been cast upon My head, who am the greatest Sufferer, though the least guilty? And was it not required to endeavor the stopping of that flux, which, if not stopped, will bring an absolute diminution to this Nation? And what more speedy and effectual labors produce the final effects?

Text sources



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**

/ Text sources

To perform text analysis, you'll need text

- Material to analyze / search / etc.
- Evaluation data for automatic approaches
- Training data for machine learning methods

We'll next look at

- Types of text data and their sources
- Acquiring, and preparing text data
- Some associated legal issues

/ Corpora

A **corpus** is a *structured* and *documented* collection of language data

Corpora can be classified e.g. by

- **Modality**: text corpora vs. speech corpora
- **Language**: monolingual (Finnish, English, etc.) vs. multilingual
- **Domain**: “general” (domain-agnostic) vs. specific domain (e.g. life sciences)
- **Producer**: human-authored vs. automatically generated
- **Annotation**: unannotated (raw) vs. annotated

On this course, primary focus on both **annotated** and **unannotated** human-authored monolingual text corpora, mostly not restricted by domain

/ Corpus annotation (examples)

Unannotated text:

When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take. He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice. Ray was among the first journalists to arrive and he confirmed those reports -- in all formats. He shot powerful video, put victims on the phone with AP Radio and played back sound to an editor who

Annotated text (morphology+syntax):

1	It	PRON	4	nsubj
2	is	AUX	4	cop
3	a	DET	4	det
4	time	NOUN	0	root
5	to	PART	6	mark
6	learn	VERB	4	acl
7	what	PRON	8	nsubj
8	happened	VERB	6	ccomp
9	and	CCONJ	13	cc
10	how	ADV	13	advmod
11	it	PRON	13	nsubj
12	may	AUX	13	aux
13	affect	VERB	8	conj
14	the	DET	15	det
15	future	NOUN	13	obj

/ Corpus annotation

Trillions (10^{12}) of words of **unannotated text** text freely available – enough for almost any TDA need (but potential challenge for small languages)

Annotations for TDA tasks typically created by humans, often reflecting substantial effort → availability of annotated corpora potential challenge

Automatic annotation possible when methods are sufficiently reliable

- Manual annotation typically gives higher **quality** at *much* higher **cost**

(Annotations can be almost anything: language, genre, segmentation, morphology, syntax, fluency, sentiment, emotion, toxicity, names, dates, relations, etc.)

PERSON
Erik Justander (DATE noin 1623 , LOCATION Turku – DATE 10. marraskuuta 1678 , LOCATION Mynämäki)

Understanding what **resources** are available and what **methods** best applicable to the task is key to developing successful approaches!

/ Annotations and methods

Different methods place different requirements on corpora

- **Rule-based methods:** mostly no need for **unannotated corpora**, **annotated corpora** only required for evaluation (cost: engineering effort)
- **Early machine learning methods:** mostly no need for **unannotated corpora**, large **annotated corpora** required for training (+evaluation)
- **Early deep learning methods:** moderate need for **unannotated corpora**, decreasing need for large **annotated corpora** for training (+evaluation)
- **Modern deep learning methods:** **extreme requirements** on unannotated corpora, **annotated corpora** often only required for evaluation

Web crawl data



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**

/ Web crawl data

The internet is by far the **largest readily accessible source of text data**

Data collected from internet by **web crawlers** (e.g. <https://nutch.apache.org/>)

Crawled data mostly HTML, PDF, etc. → **need to extract text**

Key publicly accessible collections of crawl data:

- Common Crawl (<https://commoncrawl.org/>): data from over 100 billion web pages, freely available for **downloadable in bulk**
- Internet Archive (<https://archive.org/>) history of over **900 billion web pages**, searchable and downloadable via web page (**no bulk download**)

(Many large tech companies, esp. in search/AI, have proprietary crawls)

/ Common Crawl

Web crawl data collected and freely shared by US-based nonprofit organization

First data from 2008, **over 100 crawls** performed (e.g. 10 crawls in 2024)

Over **10PB** (10 million gigabytes) of data, over **100B pages** (including duplicates), tens of trillions of words of text

By far the largest text collection readily available to anyone, but very mixed quality, containing samples of the best and worst of the web

Distributed in Web Archive ([WARC](https://archive.org/)) format from <https://commoncrawl.org/> along with metadata (WAT) and extracted text (WET) archives

/ Common Crawl data (example)

https://github.com/TurkuNLP/textual-data-analysis-course/blob/main/common_crawl_example.ipynb

```
Biblio | Distributed Systems Group
Skip to main content
Distributed Systems Group
Main menu
login
Home
People
[...]
You are here
Home
Biblio
```

← Headers

← Navigation

Usable text?



```
Found 13 results
Author Title [Type] Year Filters: Author is René Brunner [Clear All Filters]
Conference Paper
R. Brunner, Chao, I., Chacin, P., Freitag, F., Navarro, L., Ardaiz, O., Joita, L., and Rana, O. F., "Assessing a Distributed Market Infrastructure for Economics-Based Service Selection", in GADA'07 On the Move to Meaningful Internet Systems, Vilamoura, Portugal, 2007, Springer., vol. 4804, pp. 1403-1416.
P. Chacin, León, X., Brunner, R., Freitag, F., and Navarro, L., "Core Services for Grid Markets", in The CoreGRID Symposium (CGSYMP 2008), Las Palmas de Gran Canaria, Spain, 2008.
R. Brunner and Freitag, F., "Elaborating a Decentralized Market Information System", in On the Move to Meaningful Internet Systems 2007: OTM Academy Doctoral Consortium, Vilamoura, Portugal, 2007, vol. 4805, pp. 245-254.
F. Rodriguez-Haro, Freitag, F., Navarro, L., [...]
```



/ Challenges working with web data

- Complexity of WARC and HTML parsing
- Headers, footers, navigation and other non-content elements
- Repeated boilerplate text (e.g. “about us”)
- Ads, spam, SEO keyword lists, etc.
- Dynamic content (JavaScript)
- Multilingual content, typically mostly not in target language(s)
- Format errors, incomplete or missing metadata
- Encoding errors, erroneous encoding information
- Duplication of text across multiple pages / crawls
- Toxic / sexually explicit / illegal content
- Personally identifying information
- ...

/ From crawl to text corpus

Some steps to create clean unannotated text corpora from crawl data

- Text extraction from HTML, excluding boilerplate (e.g. [Trafilatura](#))
- Exact duplicate removal (e.g. [hashlib](#))
- Near-duplicate removal (e.g. [Onion](#), [minhash](#))
- Heuristic quality filtering (e.g. remove docs with low token-type ratio)
- Language model-based quality filtering (e.g. [kenlm](#))
- URL-based filtering (e.g. [UT1 blacklists](#))
- Content-based toxicity filtering (e.g. [detoxify](#))
- Content quality filtering (e.g. [FineWeb-edu classifier](#))
- Personal information masking (e.g. [pii-process](#))

Pipelines: [datatrove](#) (Hugging Face), [ungoliant](#) (OSCAR)

/ Some crawl-derived corpora

C4 (English) and mC4 (multilingual): derived from **Common Crawl** (Google, Ai2)

CulturaX: derived from **Common Crawl** (UONLP)

Dolma (English): derived in part from **Common Crawl** (Ai2)

FineWeb (English) and FineWeb 2 (multilingual): derived from **Common Crawl** (HF)

HPLT: multilingual corpora derived from **Common Crawl** and Internet Archive (HPLT)

RedPajama: corpora in five language derived from Common Crawl (together.ai)

OSCAR: multilingual corpora derived from **Common Crawl** (OSCAR, Common Crawl)

(Sizes range from low trillions to low tens of trillions of words)

/ Selecting a crawl-derived corpus

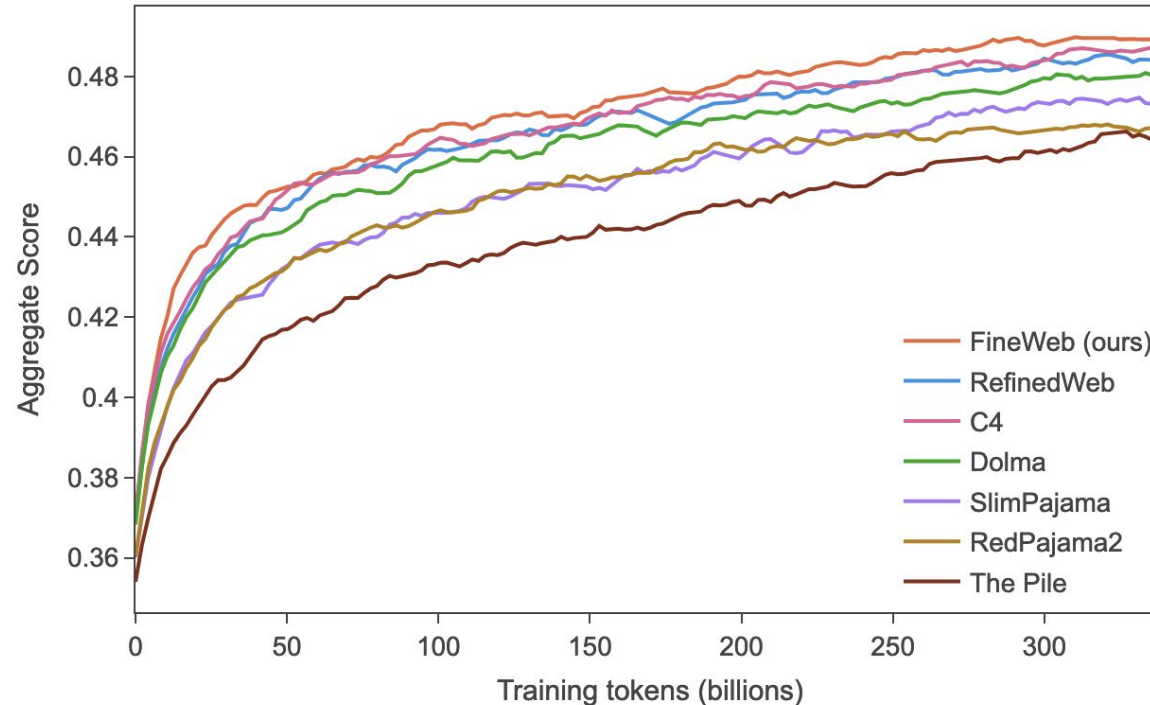
There is no single best crawl-derived corpus, and multiple criteria for selecting one. One approach is demonstrated by the Hugging Face [FineWeb](#) effort:

- **Identify tasks** that reliably measure large language model (LLM) capabilities
- **Train LLMs** on different corpora, measure task performance
- Corpora that produce LLMs with higher task performance are “better”

Limitations:

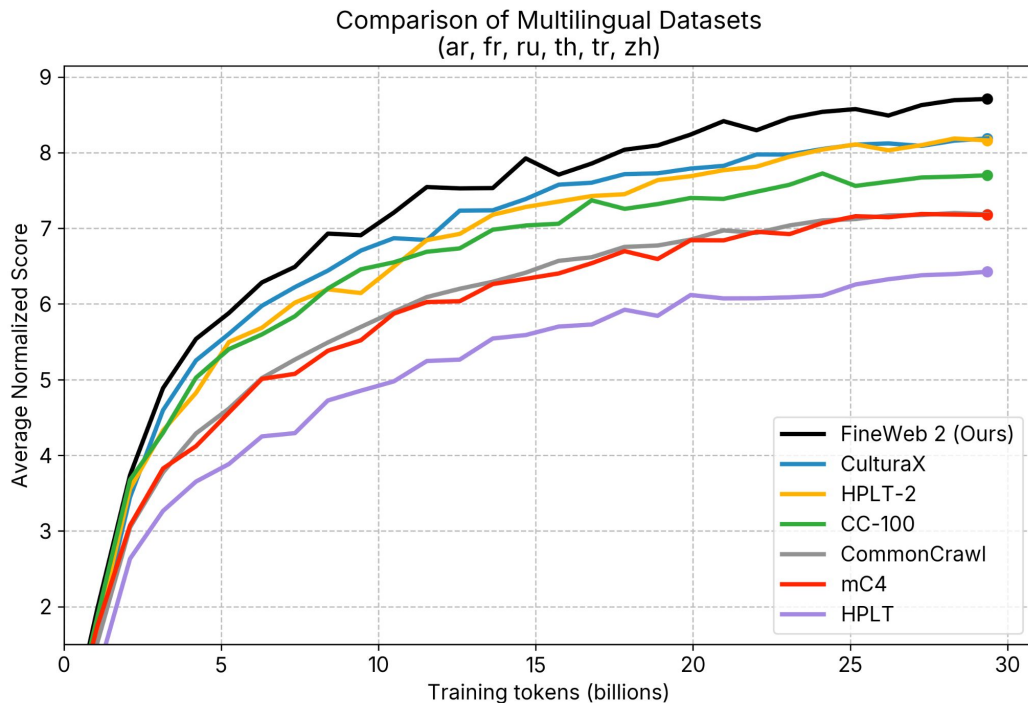
- Compute-heavy: LLMs of >1B parameters trained on > 10B tokens
- Only focuses on LLM pre-training

Selecting a crawl-derived corpus



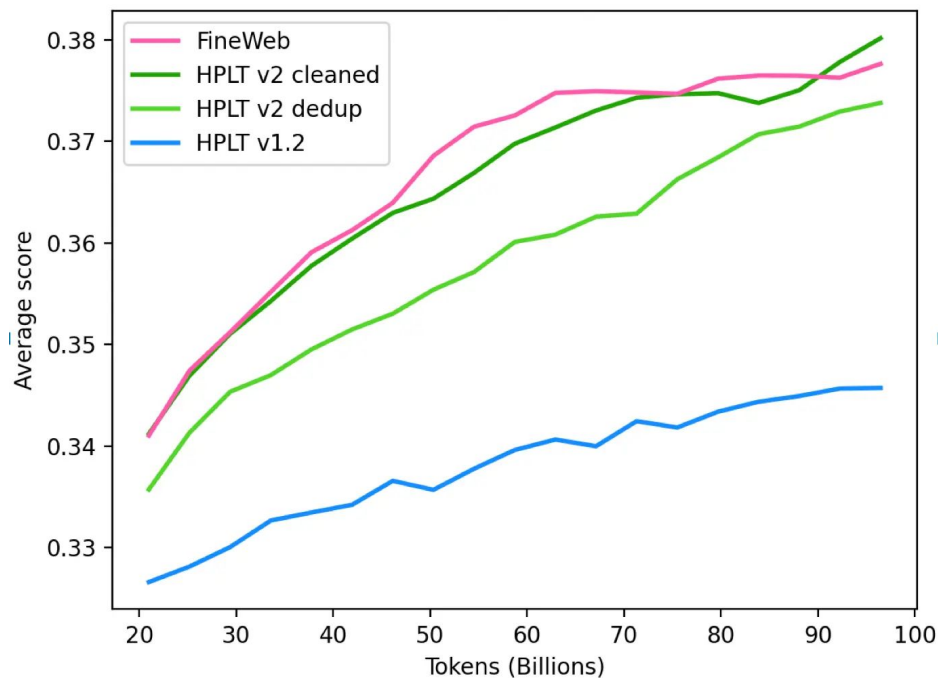
Hugging Face
evaluation results
from FineWeb v1
work (English only)

Selecting a crawl-derived corpus



Hugging Face
evaluation results
from FineWeb v2
work (multilingual)

Selecting a crawl-derived corpus



HPLT evaluation
results (English only)

Other text data sources



TURKUNLP
.ORG



UNIVERSITY
OF TURKU

/ Finnish-focused corpora: Kielipankki

<https://www.kielipankki.fi/>

The **Language Bank of Finland** (Kielipankki) is a service for researchers and students that provides access to a variety of text and speech corpora and tools

Nearly 300 resources, including more than a decade of the Finnish Broadcasting Company (**Yle**) news

Some of the largest collections of **high-quality Finnish text** available for research

Many resources (including news) require registration and have academic-use-only licenses

/ Social media

Text published via social media can serve as examples of **informal language use** and can be used to study e.g. **public sentiment**

Social media sites generally **disallow** systematic crawling of their content and increasingly only provide large-scale access for a price (e.g. Reddit, X/Twitter)

Examples of accessible social media resources:

- **Suomi24**: largest online Finnish discussion site, 2001-2017 archive (over a billion words) available for academic use from Kielipankki
- **Ylilauta**: anonymous Finnish discussion site, 2012-2014 archive (approx. 25 million words) publicly available from Kielipankki



/Wikipedia

Wikipedia is available under **open licenses** in over 300 languages

- Sizes vary a lot: e.g. **English ~3B words**, **Finnish ~100M words** (~3%)

Mostly well-edited and factual (by web standards)

Partly **structured** (infoboxes, internal links, WikiData links)

Full database freely available for download from <https://dumps.wikimedia.org/>

- Comes in hard-to-parse wikitext format but off-the-shelf text extractors exist, e.g. [wikiextractor](#), [mwparserfromhell](#)

Widely used source for corpora and other NLP work

/ Gutenberg and Lönnrot

Project Gutenberg and its Finnish sister project **Projekti Lönnrot** collect and distribute copyright-free books

- Project Gutenberg: 70,000 books, multilingual
- Projekti Lönnrot: ~2,500 books, mostly Finnish

High quality and substantial size, but **low coverage of recent language** use due to primarily consisting of out-of-copyright works (death of author + 70 years)

<https://www.gutenberg.org>

<http://www.lonnrot.net>

/ Other out-of-copyright text collections

Examples of Historical datasets

- **ECCO Eighteenth Century Collections Online**: all titles printed in England in 18th century. Challenges: moderate-to-sever OCR noise, restrictive license on the data by GALE who invested millions into digitizing the books and needs to recover the cost, over 200,000 titles
- **Gallica** [\[HF Dataset\]](#): 17-20th Century French literature, nearly 300,000 titles. Challenges: OCR noise (not as severe as in ECCO)
- **Finnish National Library newspaper and magazine collection** (1771-1920). Challenges: moderate-to-severe OCR noise caused by challenging font and newspaper layout. Accessible via kielipankki.fi

/ Scientific literature

Scientific publishing is increasingly moving to an **Open Access** model where papers are freely available to anyone

For example, in the biomedical domain, millions of full-text articles and tens of millions of publication abstracts available (**tens of billions of words**)

- PubMed (abstracts): <https://pubmed.ncbi.nlm.nih.gov/>
- PubMed Central (full texts): <https://www.ncbi.nlm.nih.gov/pmc/>

Similar repositories exist for many other fields of science, e.g. <https://arxiv.org/> for physics, math, CS and many other areas

(These text sources may only be highly relevant for NLP targeting scientific text)

/ Synthetic corpora

Until recently, effectively all corpora contained only language produced by humans

Synthetic corpora produced by generative AI methods are an emerging class of resource for NLP

A possible concern is **model collapse**: models trained on data generated by models get progressively worse (see e.g. Shumailov et al. 2023)

- Possible resolution by combining synthetic with human-authored data: Gerstgrasser et al. 2024

Example synthetic corpus: [Cosmopedia](#) (HF): 30 million English documents (25 billion tokens) of textbooks (etc.) generated by [Mixtral-8x7B-Instruct](#)

Annotated corpora



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**

/ Selected annotated corpora

This is a quick look at some collections of corpora (and a few individual ones)

The total number of resources introduced in NLP research is vast, and this does not attempt to be a comprehensive overview

We do not expect you to remember particular details of these resources, consider this more of a potential reference than study material for the exam

(... but if you consider becoming an NLP practitioner, you should at least be aware of most of these)

/ CoNLL corpora

The **Conference on Computational Natural Language Learning** (CoNLL) has held yearly shared tasks (community competitions) yearly since 2000

The CoNLL shared tasks have introduced many of the longest-standing **benchmark corpora** in NLP, frequently emphasizing multilinguality

Some tasks covered by corpora introduced in the CoNLL shared tasks:

- [2002](#) / [2003](#): Named entity recognition
- [2006](#) / [2007](#): Dependency parsing
- [2011](#) / [2012](#): Coreference resolution
- [2013](#) / [2014](#): Grammatical error correction
- [2017](#) / [2018](#): Dependency parsing (Universal Dependencies)
- [2023](#) / [2024](#): Resource-limited LM training

/ SemEval corpora

Semantic Evaluation (SemEval) is a series of NLP system evaluations similar to CoNLL shared tasks

SemEval has emphasized tasks focusing on **meaning** (as opposed to e.g. syntax), running multiple tasks each year and recently introducing resources for e.g.

- Question answering
- Sentiment analysis
- Semantic parsing
- Multilingual textual similarity

SemEval resources can be found via <https://semeval.github.io/>

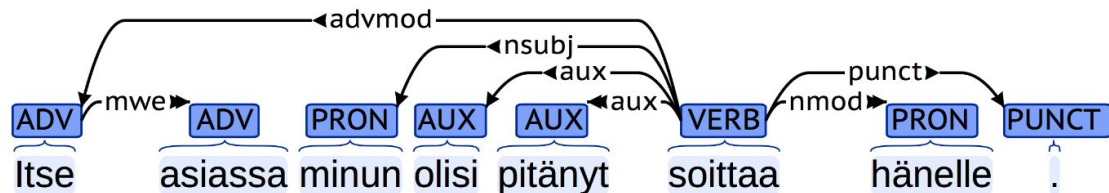
/ Universal Dependencies

Collaboratively created collection of corpora with cross-linguistically consistent annotation of **tokenization**, **morphology** and **dependency grammar**

Over 600 contributors, over 200 corpora in over 150 languages

Homepage: <https://universaldependencies.org/>

UD parsers: <https://ufal.mff.cuni.cz/udpipe>, <https://github.com/Hyperparticle/udify>,
<https://stanfordnlp.github.io/stanza/>, <http://turkunlp.org/Turku-neural-parser-pipeline/>



/ LLM benchmarks

In recent years, there has been intense interest in the development of corpora specifically for the purpose of **evaluating large language models**

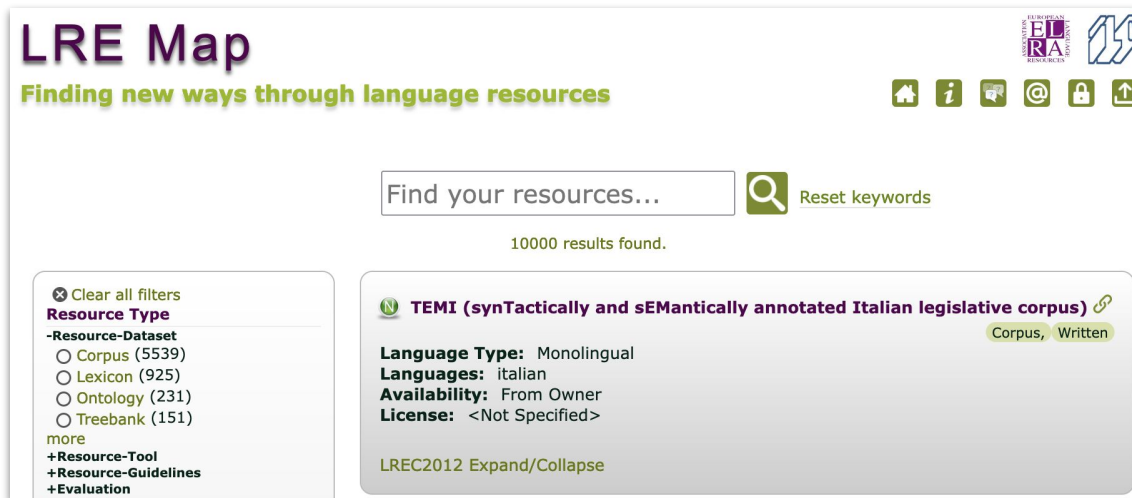
Thousands of individual tasks have been introduced; some notable collections and tools

- [BIG-bench](#): collaborative Google initiative to create benchmark tasks, 200+ tasks in multiple languages (most in English)
- [LM eval harness](#): evaluation framework developed by Eleuther AI, integrates over 60 different benchmarks totalling hundreds of tasks
- [Lighteval](#): evaluation framework developed by Hugging Face, extends on LM eval harness and [HELM](#) (Stanford)


/LRE Map

The Language Resources and Evaluation (LRE) Map (<https://lremap.elra.info/>) is a repository of corpora, tools, guidelines and other NLP resources


As of early 2025, over 5000 corpora have been registered



LRE Map
Finding new ways through language resources



Find your resources...  Reset keywords

10000 results found.

 Clear all filters

Resource Type

- Resource-Dataset
 - ☐ Corpus (5539)
 - ☐ Lexicon (925)
 - ☐ Ontology (231)
 - ☐ Treebank (151)
- more
- +Resource-Tool
- +Resource-Guidelines
- +Evaluation

 **TEMI (synTactically and sEMantically annotated Italian legislative corpus)** 

Corpus, Written

Language Type: Monolingual
Languages: italian
Availability: From Owner
License: <Not Specified>

LREC2012 Expand/Collapse



/ Hugging Face datasets

The AI startup Hugging Face has created a dataset repository that has quickly expanded to have broad coverage of NLP corpora (among other resources)

The repository now contains **over 28,000 datasets**, including resources such as many CoNLL corpora and large unannotated corpora derived from internet crawls

A key benefit of the resource is that all corpora are usable via a **uniform Python library**, reducing the need to do resource-specific format processing

- Datasets repository: <https://huggingface.co/datasets>
- Datasets library documentation: <https://huggingface.co/docs/datasets/index>

(We use this repository and the `datasets` library extensively on our courses)

/ datasets and FineWeb-edu example

https://github.com/TurkuNLP/textual-data-analysis-course/blob/main/datasets_fineweb_edu_example.ipynb

The Independent Jane

For all the love, romance and scandal in Jane Austen's books, what they are really about is freedom and independence. Independence of thought and the freedom to choose.

Elizabeth's refusal of Mr. Collins offer of marriage showed an independence seldom seen in heroines of the day. Her refusal of Mr. Darcy while triggered by anger showed a level of independence that left him shocked and stunned.

The freedom she exhibited in finally accepting him in direct defiance of Lady Catherine and knowing her father would disapprove was unusual even for Austen. In her last book Anne Elliot is persuaded to refuse Captain Wentworth at Lady Russel's insistence.

Although Jane played by the rules of the day, all of her writing is infused with how she wanted life to be. She 'screams' her outrage at the limitations for women in Emma.

When accosted by Mrs. Elton, Jane Fairfax says,

"Excuse me, ma'am, but this is by no means my intention; I make no inquiry myself, and should be sorry to have any made by my friends.

When I am quite determined as to the time, I am not at all afraid of being long unemployed.

Code:

```
dataset = load_dataset(  
    'HuggingFaceFW/fineweb-edu',  
    'sample-10BT',  
    split='train',  
    streaming=True  
)  
  
first = next(iter(dataset))  
print(first['text'])
```



Legal issues



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**

/ Legal questions (EU perspective)

Text is born copyrighted: “If you create literary, scientific and artistic work, you automatically have copyright protection” [1]

Copyright lasts **70 from death of author** in EU (>50 years in any Berne country)

Most textual data analysis involves creating at least a technical copies, and thus typically require either

- Texts not under copyright (out of copyright duration or [public domain](#))
- Explicit permission (license), e.g. a [Creative Commons license](#)
- An exception, esp. [Text and Data Mining \(TDM\) exceptions](#) (**NB:** no “fair use”!)

(In practice, students rarely affected, research and commercial efforts are)

/ Text and Data Mining (TDM) exceptions

TDM exceptions in EU [Digital Single Market](#) directive:

- **TDM for research** (Article 3): TDM copyright exception for “research organisations ... for the purposes of scientific research ... of works ... to which they have lawful access” (**mandatory**)
- **TDM for other purposes** (Article 4): TDM copyright exception for anyone to works to which they have lawful access (**opt-out possible**)

Implemented in **finnish copyright law [13 b §](#)**: “Tutkimusorganisaatiot ... joilla on laillinen pääsy teokseen, saavat valmistaa siitä kappaleita tieteellisessä tutkimuksessa tapahtuvaa tekstin- ja tiedonlouhintaa varten ...”

On paper, very broad protections for TDA for research in particular

/ TDM challenges

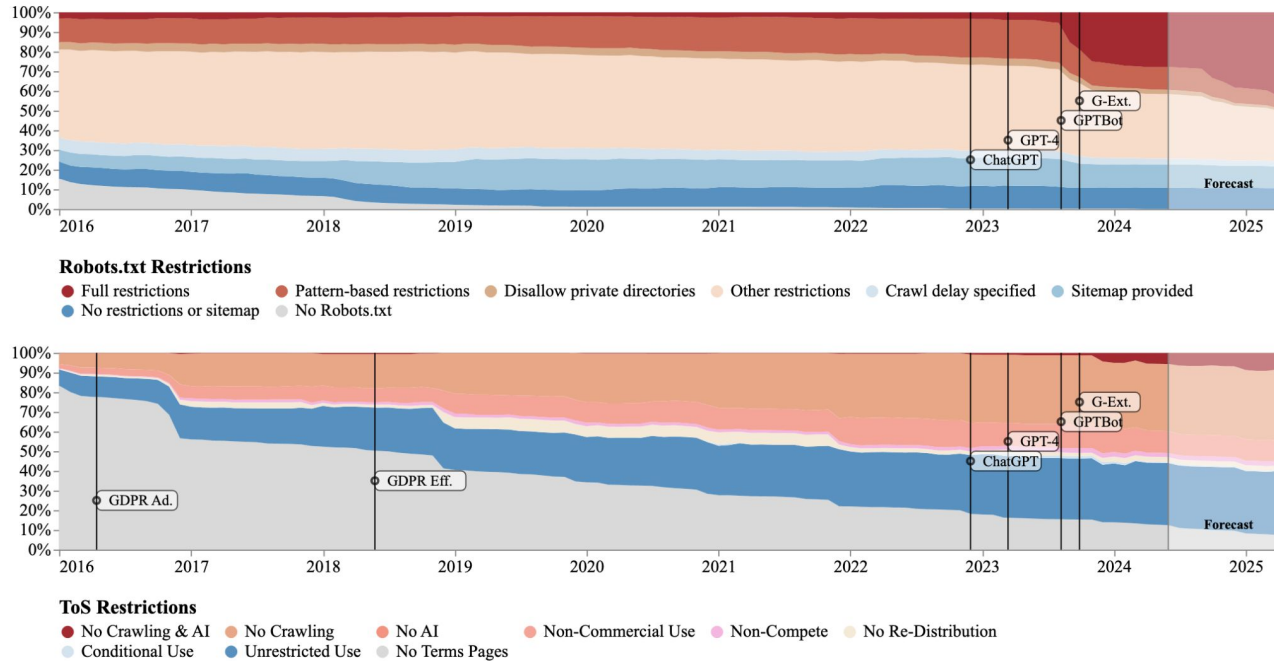
A number of copyright holders have recently **challenged the TDM exceptions**, in particular as they relate to creating large language models

EU legislation appears to broadly support an interpretation where **LLMs are covered by the exceptions**, but the situation remains unclear and national laws may go various ways

A recent view from Finland: [Tekoäly, tekijänoikeus ja tutkimus – näkökulmia ajankohtaiseen tilanteeseen ja tekoälytutkimuksen tulevaisuuteen](#) (Lilja 2024)

Related (US): [New York Times sues Microsoft and OpenAI for ‘billions’](#) (2023), contesting fair use arguments

/ Increasing restrictions on web data



Longpre et al. (2024) report a rapid increase in limitations on the use of web data, with approx. 50% of C4 domains limiting crawling either by robots.txt or terms of service

/ Personal data and GDPR

Several corpora derived from web crawls have been demonstrated to contain **personal data**, defined **very broadly** in the [GDPR](#):

All data related to an identified or identifiable person are personal data

(The related term in US legislation is Personally identifiable information, or PII)

Current state-of-the-art tools for masking personal data/PII typically work on comparatively **easy-to-identify** classes of identifiers, e.g.

- Phone numbers, credit card numbers, social security numbers, email addresses

No broader personal data/PII challenges have been raised against web-based corpus providers, many of which provide a takedown option as a partial answer