

Making a Very Large Pre-Training Dataset: Social and Technical Considerations

Yacine Jernite, 🙌
@Stanford NLP Seminar
5/6/2021

Introduction: Who's Talking?



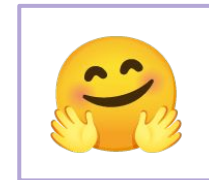
- NYU, PhD, 2012-2018
 - Language Modeling
 - Medical Applications



- FAIR NY, Postdoc, 2018-2020
 - Minecraft AI bot
 - Conditional LM



- Hugging Face, Researcher, 2020-now
 - 🔍: More Long Form QA
 - 📖: Language Datasets, Social Context



Introduction:

Recent work [@huggingface](#)



- Hugging Face, Researcher, 2020-now
 - [ELI5: Long Form Question Answering](#), Fan et al., 2019
 - From Subreddit “Explain Like I’m Five”, complex questions!

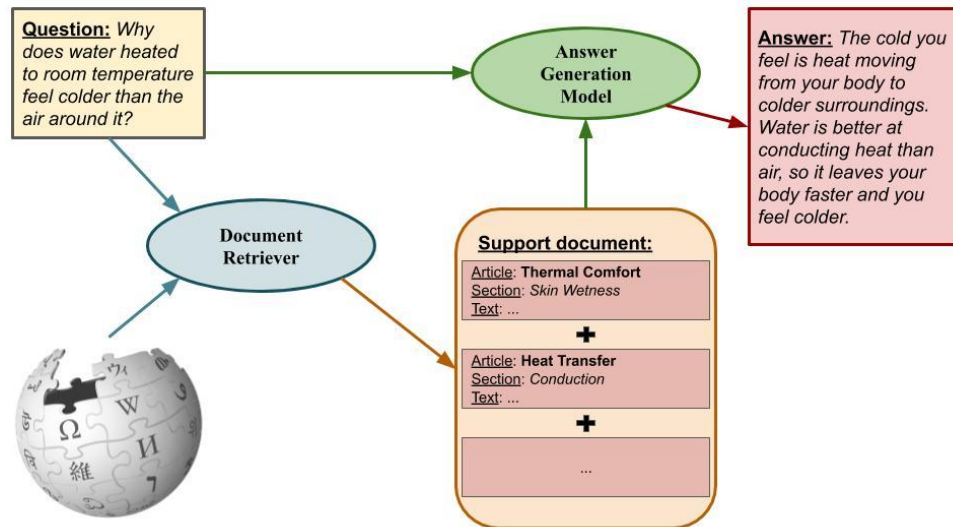
HOW	How do unions work and what gives them authority? How do ISP Internet Service Providers work? How does my car engine work? How exactly does a massive sewer system work in a large city?	IS	WHERE
WHY	Why do we get munchies? Why did the federal government take over student loans? Why is this video blocked in your country necessary? Why can't we just print money to pay off our debt? Why did Blu ray beat HD DVD in their format war? Why was there a rivalry between Tesla and Edison?	WHEN	IN
		When is y considered a vowel?	CAN
		IF	If I don't have stock, how would a recession affect me personally?
		WHAT	What is a Turing machine and why is it so important? What exactly is hiccuping and why do we do it? What's the difference between 32 and 64bit operating systems?

Introduction:

Recent work [@huggingface](#)



- Hugging Face, Researcher, 2020-now
 - ELI5 and Dense Retrieval from “scratch” (*read: from BERT*)
 - First version from CommonCrawl/ElasticSearch. Wikipedia?



Introduction: Recent work [@huggingface](#)



- Hugging Face, Researcher, 2020-now
 - ELI5 and Dense Retrieval from “scratch” (*read: from BERT*)
 - There’s a [tutorial](#) and a [demo](#)
 - [Dataset on the HF hub](#)

The screenshot displays the ELI5 demo interface, which is divided into three main sections:

- Left Panel:** Titled "Long Form Question Answering with ELI5". It includes a dropdown menu for selecting a question, an "OR:" section for entering a custom question (e.g., "How do people make chocolate?"), and a "Go!" button. Below this, it shows the "The model generated answer is:" section with a paragraph of text explaining the process of making chocolate.
- Middle Panel:** Titled "The model is drawing information from the following Wikipedia passages:". It lists ten numbered passages, each with an article title and a section title. For example, "01 - Article: The Licktators" with section "Foundation".
- Right Panel:** Titled "The model generated answer is:". It contains a paragraph of text: "There are a lot of different ways, but the most common way is to grind up the cocoa beans and put it in a bowl with a little bit of butter. Then you heat the butter and let it melt. The melted butter is used to make the cocoa butter, and the melted butter makes the chocolate." Below this, it shows "The most similar question in the ELI5 training set was:" followed by "How is white chocolate made?". At the bottom, it lists "Its answers were:" followed by two numbered items describing white chocolate.

Introduction:

Recent work [@huggingface](https://huggingface.co)



- Hugging Face, Researcher, 2020-now
 - Lots of **datasets!**
 - Launched in May 2020
 - Unified API for diverse OSS datasets
 - Memory-mapped for convenience
 - Community effort for [multilingualism](#)
 - ***Sprint*** with >200 contributors!
>1000 dataset, >100 langs
 - Centralized hub and **documentation**

English	en	312
Spanish	es	72
German	de	68
French	fr	68
Polish	pl	56
Portuguese	pt	56
Arabic	ar	54
Russian	ru	50
Italian	it	48
Dutch	nl	48
Turkish	tr	47
Chinese	zh	45
Swedish	sv	44
Romanian	ro	43
Finnish	fi	42
Czech	cs	40
Greek	el	39
Thai	th	39
Hungarian	hu	38
Korean	ko	37
Bulgarian	bg	36
Japanese	ja	36

Introduction:

Recent work [@huggingface](#)



- Hugging Face, Researcher, 2020-now
 - Dataset documentation for everyone (w/ Angie McMillan-Major)
 - Motivated by recent works:
 - [Model cards for Model Reporting](#), Mitchell et al., 2018
 - [Datasheets for Datasets](#), Gebru et al., 2018
 - [Data Statements for NLP](#), Bender & Friedman, 2018
 - Especially important for central hub, and great opportunity!
 - Come back next week for more detail 😊

Introduction:

Recent work [@huggingface](#)



- Hugging Face, Researcher, 2020-now
 - Dataset documentation for everyone (w/ Angie McMillan-Major)
 - Example: the [ELI5 Data Card](#)
 - **Social Biases**: contrasting known issues with Reddit as a whole to specific choices made when constituting the ML dataset
 - [Post from Reddit founder on the failures of moderation](#)
 - [2019 Wired article on misogyny on Reddit](#)
 - [r/AskHistorians subreddit specific moderation rules](#)
 - **Known Limitations**:
 - Recent work outlines limitations of the split: [Krishna et al. 2021](#)
 - [Subjective answers!](#)

Introduction:

Recent work [@huggingface](https://github.com/huggingface)



- Hugging Face, Researcher, 2020-now
 - Dataset documentation for everyone (w/ Angie McMillan-Major)
 - Example: the [ELI5 Data Card](#)
 - **New Limitations:**
 - Living document!
 - 1st step - team playing around
 - Camille Saint-Saëns - hehe it got the gender wrong :)
 - Pineapple pizza - get the answer you want
 - 2nd step - Twitter feedback
 - Hmmmm what about all of India though?
 - Religion, Tagore mistakes

Introduction:

Recent work [@huggingface](https://twitter.com/huggingface)



- Hugging Face, Researcher, 2020-now
 - Since January 2021: [BigScience](https://twitter.com/BigScience)



Introduction:

Recent work [@huggingface](https://huggingface.co)



- Hugging Face, Researcher, 2020-now
 - Since January 2021: [🌸 BigScience 🌸](#)
 - Very Large LMs are here to stay
 - Also, pose a number of scientific and society questions
 - [Stochastic Parrots 🦜](#): Angie next week!
 - [Datasets as Infrastructure](#), Hutchinson et al. 2021
 - Very Large LMs are extremely resource intensive
 - What we do at 😊: community!
 - Thom: “*We need a Large Hadron Collider of LMs!*”

The BigScience Project: Summer of Language Models 2021



The Large Hadron Collider is a particle physics research tool which

- has involved **10.000 researchers**
- from **100 countries**
- lead to the discovery of **59 hadrons**
- publication of more than **2.800 papers** (🤖)

In many scientific fields (epidemiology, space, fusion...), **large-scale and worldwide research collaborations** create **tools** useful for the entire research community, e.g. LHC, ITER, ISS...

*Isn't it time to build similar **large, diverse, open research collaborations** in AI/NLP as well?*



The BigScience Project: Summer of Language Models 2021



Recent developments in NLP from training **larger language models** on **larger dataset**. But the compute resources are typically found **in industry, which means:**

- **Research**

- Models **not designed as general research tools**
- Difficult **involvement of academic** researchers
- Lack of **fields diversity** of the research teams building them

- **Environmental**

- Training parallel models in private setting => **duplication** of energy requirements
- **Carbon footprint** not documented/taken into account

- **Ethical and societal**

- **Shortcomings in the text corpora** used to train these models
- Ethical/bias/usage question are usually asked **a-posteriori**



The BigScience Project: Summer of Language Models 2021



- What do we need: Infrastructure
 - [Jean Zay supercomputer](#) at IDRIS
 - Managed by French Department of Research and Education
 - 5M GPU hours to support the project!

Accelerated partition (or GPU partition)

- 261 four-GPU accelerated compute nodes with:
 - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
 - 192 GB of memory per node
 - 4 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- 31 eight-GPU accelerated compute nodes, currently dedicated to the AI community with:
 - 2 Intel Cascade Lake 6226 processors (12 cores at 2.7 GHz), namely 24 cores per node
 - 20 nodes with 384 GB of memory and 11 nodes with 768 GB of memory
 - 8 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- Extension in the summer of 2020, 351 four-GPU accelerated compute nodes with:
 - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
 - 192 GB of memory per node
 - 4 Nvidia Tesla V100 SXM2 GPUs (16 GB)



The BigScience Project: Summer of Language Models 2021



- What do we need: Organization
 - Year-long Workshop open to the research community and beyond
 - May'21 to May '22 (Kick-off session last Thursday 28/4)
 - 330+ currently registered
 - Driving thread: train a very large multilingual LM together
 - Gathering into working groups on modeling, evaluation, tokenization, environmental impact, and many others!
 - Social impact work happens IN the working groups
 - LHC parallel: release the trained model, enable replication of the effort

A Large Multilingual Dataset for a Large Multilingual Model

- Training a very large multilingual model:
 - Collaborative effort
 - Research and evaluation
- The model is specified by its:
 - Architecture, training objective and algorithm
 - **Training Data**
 - *What are the important questions?*



Dataset Background: Project Philosophy




Starting points for the data work:

-  - Respecting the **rights of data subjects**

Dataset Background: Data Creators and Data Subjects



-  - Respecting the **rights of data subjects**
 - “Data Are People!”
 - [Human Data Science](#), Oberski, 2020

Dataset Background: Data Creators and Data Subjects



- **Data creators:** language writers or speakers
- **Data subjects:** who the language data gives information about
- Existing regulation defines the **rights** of data subjects
- This project also aims to more broadly consider their **interest**


Dataset Background: Data Creators and Data Subjects



-  - Respecting the **rights of data subjects**
 - **Data creators:** commercial ownership
 - Copyright, Intellectual Property
 - In the US: fair use, DMCA provisions
 - [Fair learning](#), Lemley & Casey, 2021
 - In the EU: [Directive on Copyright in the Digital Single Market](#)
 - Elsewhere?


Dataset Background: Data Creators and Data Subjects



-  - Respecting the **rights of data subjects**
 - **Data subjects:** GDPR, CCPA, etc.
 - GDPR:
 - broad definition includes e.g. political opinions
 - Different regime for “public interest” and archival institution
 - Managed at the national level (e.g. CNIL)
 - [Integrating the Management of Personal Data Protection and Open Science with Research Ethics](#), Lewis et al. 2016

Dataset Background: Data Creators and Data Subjects



-  - Respecting the **rights of data subjects**
 - **Data communities:** the Mapuche and their language
 - [Language Ownership and Language Ideologies](#), M Speas, 2013
 - h/t A. Paullada [@ResistanceAI Workshop](#)
 - Advocacy groups
 - [Our Data Bodies](#)
 - [Participatory Methods in ML/Data](#)

Dataset Background: Data Creators and Data Subjects





-  - Respecting the **rights of data subjects**
 - **In other fields:** human subjects and Internal Review Boards
 - [Federal Policy for the Protection of Human Subjects](#) (Common Rule)
 - [Where are human subjects in Big Data research? The emerging ethics divide](#), Metcalf and Crawford 2016

Dataset Background: Project Philosophy



Starting points for the data work:

-  - Respecting the **rights of data subjects**
-  - Language is a **social object**

Language as a Social Object



- NLP/ML datasets pre ~2012
 - A dataset presents a sampled realization of a task
 - Underlying truth is found in the common features between realizations
 - Generality tied to capacity constraints, robustness, domain shift
 - Multi-task *a posteriori* from combining tasks
- NLP/ML datasets post Pre-Training Paradigm
 - Popularized: ImageNet for vision (2012), ELMo/BERT for NLP (2018)
 - Multi-task, general world understanding *a priori* from pre-training
 - So what does general mean?


Language as a Social Object



- GPT
 - BookCorpus, 5GB
- BERT/BART
 - BookCorpus + Wikipedia (~10GB)
- GPT-2
 - OpenWebText, 36GB - Reddit links
- Roberta
 - +CC-NEWS (76GB) - News RSS sites
- XLM-Roberta
 - CC-100 (1TB/100GB) - CommonCrawl filtered by “Wikipedia-like”
- GPT-3
 - CommonCrawl filter by “OWT-like”
- T5/Switch-C
 - C4 - “Cleaned” Common Crawl - heuristics + banned words - See [Dodge et al.](#)

Language as a Social Object



-  - Language is a **social object**
 - Different people say things differently - sociolinguistics



*How do you
do?*

*How are you
doing?*

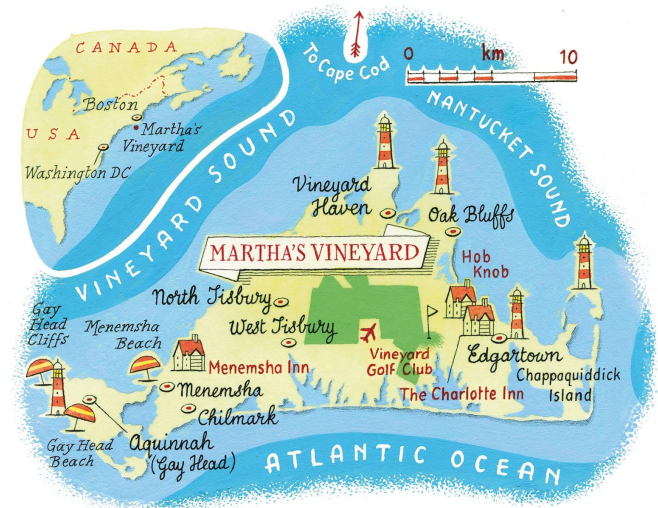


Language as a Social Object: Social Variations of Linguistic Form



A view to sociolinguistics

- “Free variation” in linguistics
 - You say ([/aɪ/])-either, and I say ([/i:/])-either
 - Changes with no bearing on meaning
 - Learned/learnt, colour/color
- Sociolinguistics: “free” but not random or unconstrained
 - The Social Motivation of a Sound Change, W. Labov, 1963



Language as a Social Object: Social Variations of Linguistic Form



A view to sociolinguistics

- [Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation](#), Penelope Eckert, 2012
 - Initial studies place speaker on a socioeconomic ladder of desirability
 - [The Social Stratification of \(r\) in New York Department Stores](#), Labov, 1972
 - Speakers have agency - positive function of proper vernacular
 - “The principal move in the third wave then was from a view of variation as a reflection of social identities and categories to the linguistic practice in which speakers place themselves in the social landscape through stylistic practice”

Language as a Social Object: Social Variations of Linguistic Form



A view to sociolinguistics

- Switching to proper vernacular conveys meaning
 - [From Sociolinguistic Variation to Socially Strategic Stylisation](#), J. Snell, 2010
 - Me/my as a possessive in NE England schoolchildren (“me pencil”)
 - Uses [mi] when stepping outside “the routine flow of unexceptional business”
 - More sophistication, not less
- Interactions between institutional and community linguistic forms
 - [Semantic Variation: Meaning in society and in sociolinguistics](#), R. Hasan, 2009
 - Studies meaning-making patterns of middle- vs working-class LA families
 - “it showed that the usual mode of teachers’ talk with these children was if anything an exaggerated version of the typical middle-class ways of meaning”

Language as a Social Object: Social Variations of Linguistic Form




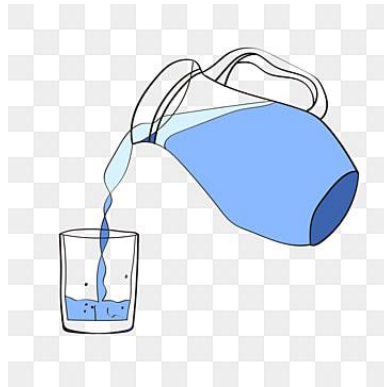
Currently in NLP

- Hierarchies of text or Domain adaptation
 - Standard vs non-standard English
 - “Wikipedia-like”, “OWT-like” filtering for RoBERTa, GPT-3
 - Twitter as “noisy” text
- [Language \(Technology\) is Power: A Critical Survey of “Bias” in NLP](#), Blodgett et al., 2020
 - Section 5, case study of AAE and how these assumptions harm
 - Annotation & data - e.g. toxicity detection, what speech is allowed?
 - [The Risk of Racial Bias in Hate Speech Detection](#), Sap et al., 2019

Language as a Social Object



-  - Language is a **social object**
 - Different people say things differently - sociolinguistics
 - Different people talk about different things
 - Interpretation in context - pragmatics



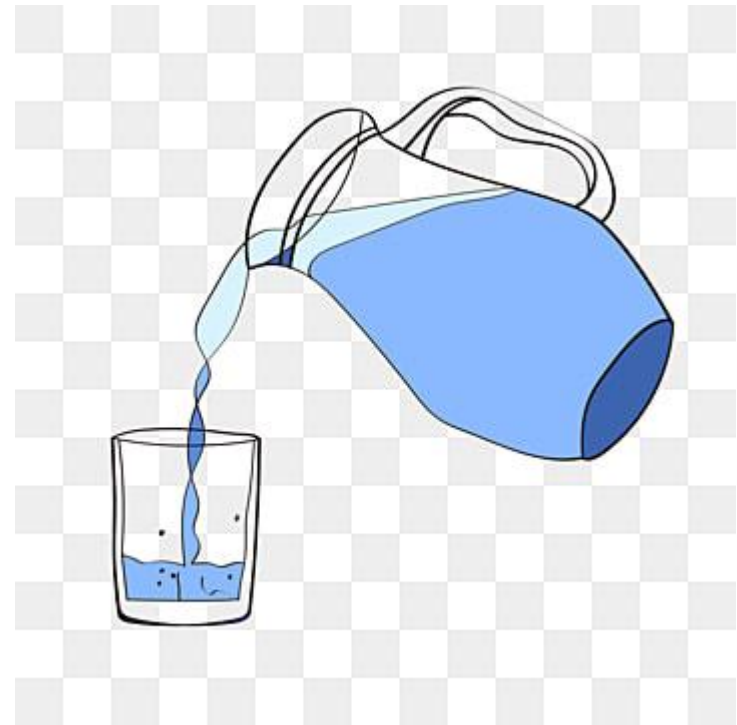
Language as a Social Object: Social Variations of Meaning



Personal anecdote:

- Partner: “Tell me when?”
- Me: “Thank you!”

Assumptions finally aligned, making
meaning together!



Language as a Social Object: Social Variations of Meaning



Pragmatics:

- “subfield of linguistics that studies how context contributes to meaning”
- Example: Grice’s maxim of quantity
 - “I ate some of the cookies” - implicates, not all - assumption about the speaker
- [Variational Pragmatics](#), Barron & Schneider, 2009
 - “Variational pragmatics can be conceptualized as the intersection of pragmatics with sociolinguistics, [...]. It is assumed that the social factors analyzed in sociolinguistics have a systematic impact not only on pronunciation, vocabulary and grammar, but also on language use in interaction”

Language as a Social Object: Social Variations of Meaning




Examples of relevant social information:

1. [Im/politeness across Englishes](#), Haugh and Schneider, 2012
 - Important if you're doing Dialogue Modeling!
2. [Some thoughts on pragmatics, sociolinguistic variation, and intercultural communication](#), Clyne, 2006
 - Australian English: "Bring a plate!"
3. "ELI5: How much power does the president actually have to make a difference in our daily lives? Is the presidency mostly a figurehead?"
 - **(Disparate) model failures** (which president? US by default?)

Language as a Social Object



-  - Language is a **social object**
 - Different people say things differently - sociolinguistics
 - Different people talk about different things
 - Interpretation in context - pragmatics
 - In-group and Out-group speech

Language as a Social Object: In-Group and Out-Group Speech



1. [Language Use in Intergroup Contexts: The Linguistic Intergroup Bias](#), Maass et al, 1989
 - LIB: positive and negative in-group vs out-group behaviors are described using different language
2. [In Consideration of Social Context: Re-examining the Linguistic Intergroup Bias Paradigm](#), Shulman et al, 2011
 - Phenomenon supported but not uniform, appears differently for minority vs majority language in a bilingual setting

Language as a Social Object: In-Group and Out-Group Speech



1. Normative expectation of prioritizing first-hand experience
 - *“In order to reduce poverty, people with lower income should...”*
 - *“In my personal experience [disabled/lgbtq/muslim/...] people do...”*
 - **Representation bias** (data subjects represented by out-group only)
2. Case of reappropriating slurs
 - In-group use of slurs can lead to re-appropriation and language change. For example, past vs current use of the word “queer”
 - [The Reappropriation of Stigmatizing Labels: Implications for Social Identity](#), Galiskiny et al., 2003

Language as a Social Object: In-Group and Out-Group Speech



Currently in NLP

- [Social Biases in NLP Models as Barriers for Persons with Disabilities](#), Hutchinson et al., 2020
 - Mostly negative mentions of people with disabilities in training corpora lead to higher incidence of toxicity labels
- [Interpreting Social Respect: A Normative Lens for ML Models](#), Hutchinson et al., 2020
 - Strongly reduces perspectiveAPI bias against LGBTQ words by soliciting some amount of group self-description

Language as a Social Object






- GPT
 - BookCorpus, 5GB
- BERT/BART
 - BookCorpus + Wikipedia (~10GB)
- GPT-2
 - OpenWebText, 36GB - Reddit links
- Roberta
 - +CC-NEWS (76GB) - News RSS sites
- XLM-Roberta
 - CC-100 (1TB/100GB) - CommonCrawl filtered by “Wikipedia-like”
- GPT-3
 - CommonCrawl filter by “OWT-like”
- T5/Switch-C
 - C4 - “Cleaned” Common Crawl - heuristics + banned words - See [Dodge et al.](#)

Dataset Background: Project Philosophy



Starting points for the data work:

-  - Respecting the **rights of data subjects**
-  - Language is a **social object**
-  - **Documentation** is paramount

Dataset Background: The Role of Documentation



- **Measuring representativeness** and diversity of social contexts
- **Implementing data subjects' rights** (e.g. to be forgotten)
- **Understanding model behaviors** in light of the training dataset

A Large Multilingual Dataset, 3 Working Groups

- Present a clear positioning with respect to data regulations
- Motivate choices with respect to their impact on data subjects
- Gather a large quantity of language data
- Support the model's ambitions of generality and multilingualism
- Provide clear, useful documentation of its components and design
- Develop appropriate infrastructure for using the training dataset

Data Working Groups:

A. Data Governance and Archival Strategies

1. *Governance of Aggregated Language Data*, 2. *Legal Frameworks*,
3. *Metadata and Infrastructure Needs*

B. Data Sourcing and Representativeness

1. *Frameworks for Representativeness*, 2. *Modes of Data Gathering*
3. *Defining Language Groups*, 4.x. *Localized Language Data Sourcing*

C. Data Tooling: from Sources to Training Dataset

1. *Multimedia & Multiformat processing*, 2. *Web Crawling and Documenting*
3. *Private Information Protection*, 4. *Indexing and Storage Infrastructure*

BigScience Data Section



- If you have any questions, see the documentation available at the [website: https://bigscience.huggingface.co/](https://bigscience.huggingface.co/)
- Fill out the [form](#). Looking forward to working with you!



A. Data Governance and Archival Strategies

Purpose of this Working Group

Data choices affect a **range of stakeholders**, so:

- What are our **legal and ethical responsibilities**?
- **Who** are we **directly and indirectly responsible to**?
- **What mechanisms** support these responsibilities?

A.1: Governance Models for Aggregated Language Data

First collaborative task:

- **Evaluate** existing modes of data governance that involve stakeholders
- **Assess** relevance to the project, and specific ethical and social implications of using speech and text data
- **Adapt** recommendations to a heteroclitic aggregated language training dataset

A.2: Legal Frameworks for Data Use around the World

Second collaborative task:

- **Identify** relevant legal frameworks for data use across regions
- **Position** the project with respect to these frameworks, especially in terms of private information and intellectual property
- **Collaborate** with governmental and nongovernmental data custodians to establish guidelines for research use
- **Internal guidelines** for other working groups to act coherently with the proposed positioning

A.3: Infrastructure Needs, Indexing Strategies, and Metadata Choices

Third collaborative task:

- How do we index such a heteroclitic dataset?
- Governance requirement: evaluating representation of stakeholders
- Legal requirement: ability to remove data
- Scientific requirement: understanding the dataset at a glance

B. Data Sourcing and Representativeness

Purpose of this Working Group

- Find language sources that are as diverse as we can make them
- Axes of representativeness to be defined for each **intersection of region and language** 🌍
- Get data sources **including but not limited to (web) text**
- Language groups: Arabic, Bantu languages, Chinese, English, French, Hindi and Urdu, Portuguese, Spanish

B.1: Frameworks and Tools for Representativeness across Regions

- What does existing literature say about the **relevant cultural categories** in each region? 🌍
- How are these categories and especially under-represented groups **presented in existing language resources and media**?
- What are the **consequences of under-representation**?

B.2: Modes of Data Gathering and Creation

Representativeness through various data collection methodologies

- What are the ways other projects and other fields have gone about **collecting data**, in various regions around the world?
- What are some **examples and best practices**?
- How will we **collaborate with data custodians** in our data collection?
- How will we **collaborate with volunteers and participatory efforts** in our data collection?

B.3: Languages, Language Varieties, and Multilinguality

8 groups: Arabic, Bantu languages, Chinese, English, French, Hindi and Urdu, Portuguese, and Spanish

Selected for geographical representativeness 🌍

- How do we **operationalize ‘language’** in our tools, modeling and evaluation?
- What **other language varieties** should we add?

B.language: Localized Data Sourcing

Data sourcing working groups by language, with subgroups for identified regions

- What are the **available resources** for each language and region?
- What are the **concerns and challenges** for specific subgroups?
- What **organizations** are already engaging in similar work?

C. Data Tooling: from Sources to Training Dataset

Purpose of this Working Group

How do we turn the language data from all of the language varieties in various media into one text format?

- Process visual and spoken data into text
- Process and document text collected from the web
- Detect private information and de-identify data
- Construct efficient database infrastructure

C.1: Multimedia and Multiformat Processing Tools

For each language how do we:

- Retrieve text from PDFs and other formats?
- Transcribe spoken language?
- Identify text in that language?
- Recognize characters for that language in images (OCR)?

C.2: Web Crawling and Documenting Tools

For data that comes from the **web** and for **each language variety** how do we:

- Collect the text through shallow crawling?
- Automatically document the text?
 - Identify the style and register of the text
 - Identify the language variety
 - Extract the Terms of Service

C.3: Automatic Detection of Private Information

For data that comes from the **web** and for **each language variety** how do we:

- Identify private information
- De-identify across the various domains

We want to collect and **maintain** the dataset with respect for data subjects' privacy and rights

C.4: Indexing, Storage, Document Representation Infrastructure

For all the data **across** all language varieties, we need:

- A common format for storing the data
 - Efficient processing without removing relevant context in the original format
- A system for indexing the data
 - Easy to navigate
- Tools to identify duplicated content

BigScience Data Section

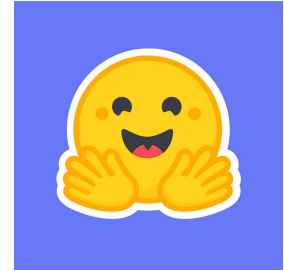


- If you have any questions, see the documentation available at the [website: https://bigscience.huggingface.co/](https://bigscience.huggingface.co/)
- Fill out the [form](#). Looking forward to working with you!



Introduction:

Short story of [@huggingface](#)



- 2016 - Started in New York/Paris
 - 2017 - Widely used chatbot
- 11/2018 - “*Let’s port BERT to PyTorch!*” - `pytorch-transformers`
 - Turns out that was useful to quite a few people!
- 2018-now - making NLP easier to use 🤗
 - `transformers` unified API for popular transformer models in TF/PT
 - 12/2019 - `tokenizers`
 - 05/2020 - `datasets`

Introduction:

Short story of [@huggingface](#)



- Code you can use - at any level of abstraction/optimization
- [Model](#) (10,000) and [Dataset](#) (1000) hubs
- Inference API to deploy models, [AutoNLP](#) as a service
- 2021 - Started [Summer of Language Models 2021](#) 🌸 (BigScience)

Tasks

Fill-Mask Question Answering Summarization

Table Question Answering Text Classification

Text Generation Text2Text Generation

Token Classification Translation

Zero-Shot Classification +5

Libraries

PyTorch TensorFlow +9

Datasets

common_voice wikipedia dcep europarl jrc-acquis squad

Models 10,727 Search Models

bert-base-uncased
Fill-Mask · Updated Apr 23 · 25,693k

jplu/tf-xlm-roberta-base
Fill-Mask · Updated Dec 11, 2020 · 7,4

bert-base-cased
Fill-Mask · Updated Apr 23 · 4,104k

bert-base-chinese
Fill-Mask · Updated Dec 11, 2020 · 2,4

Task Category

text-classification conditional-text-generation

structure-prediction question-answering sequence-modeling

other +8

Task

machine-translation named-entity-recognition

sentiment-classification language-modeling extractive-qa

multi-class-classification +150

Language

en es fr de pl pt +197

Datasets 1131 Search Datasets

acronym_identification
Acronym identification training and development task at SDU@AAAI-21.

adversarial_qa
AdversarialQA is a Reading Comprehension crowdworkers on a set of Wikipedia articles

afrikaans_ner_corpus
Named entity annotated data from the NCH Project, annotated with PERSON, LOCATION