Software projects: data or not data

Karën Fort and Fanny Ducel

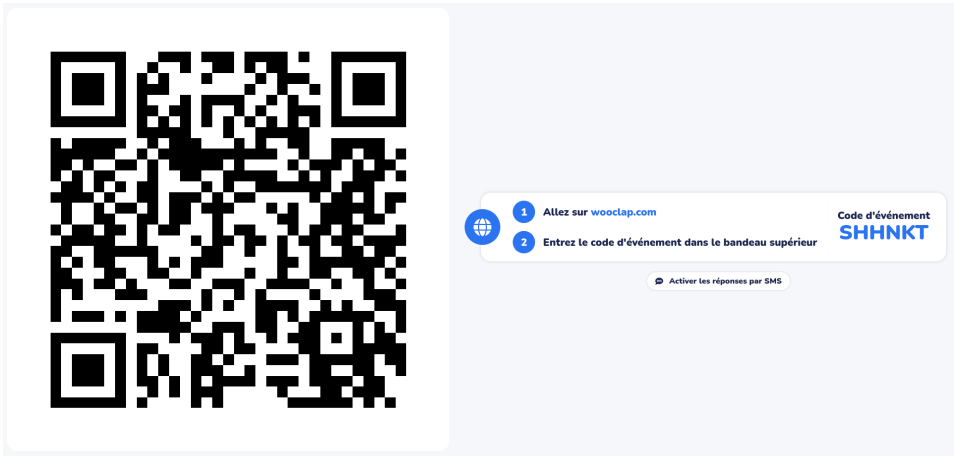
karen.fort@loria.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Wooclap

where are the data in an NLP project?



Where are the data in an NLP project?

- in the pre-training

Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)

Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)
- in the benchmark / gold

Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)
- in the benchmark / gold

- in the input

Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)
- in the benchmark / gold

- in the input
- in the output

Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)
- in the benchmark / gold

- in the input
- in the output
- in the variables

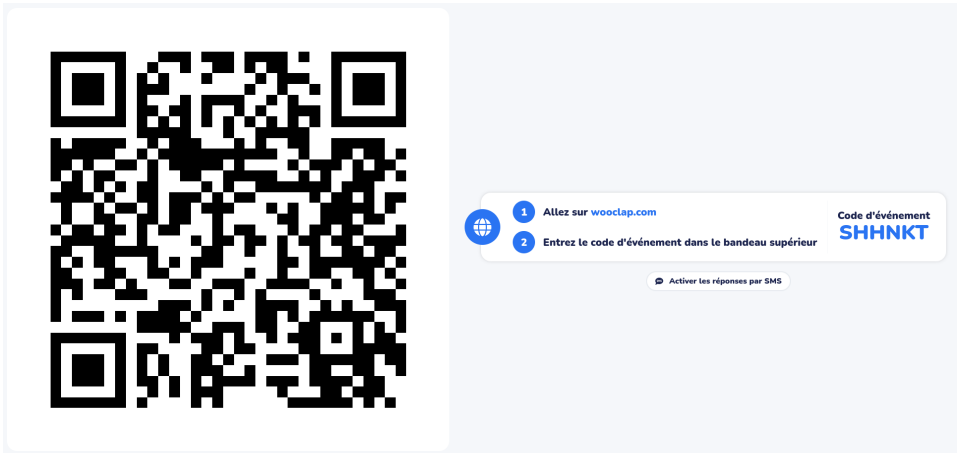
Where are the data in an NLP project?

- in the pre-training
- in the post-training (fine-tuning, prompting, CoT, etc)
- in the benchmark / gold

- in the input
- in the output
- in the variables
- in the documentation (examples)

Wooclap

How often do you look at the data? Be honest.



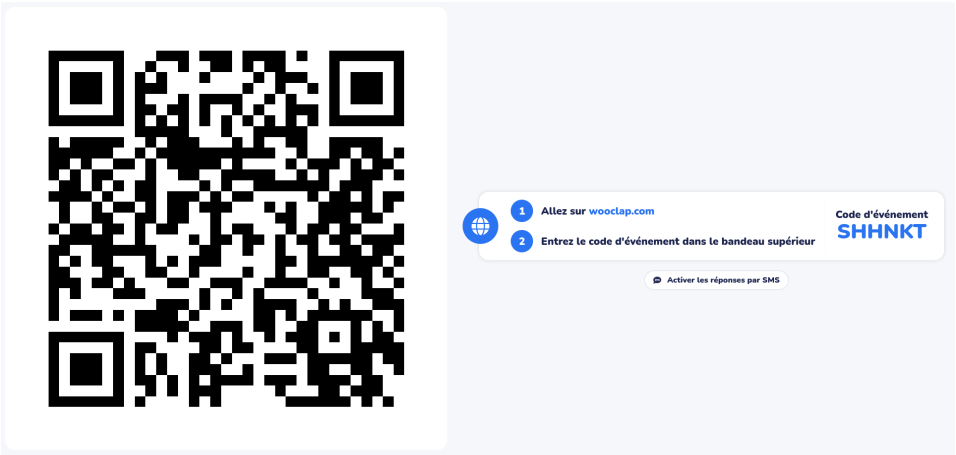
Why looking at the data?

- allows to confront the reality:
  $\rightarrow$ does it correspond to what you expected (question your own biases)?

- a lot like doing regular `print()` when you code

Wooclap

What does it mean to "look at" the data?



Why it's important?

Ben Hamner @benhamner · Oct 9

Programming: 10% writing code. 90% figuring out why it doesn't work

Analyzing data and ML: 1% writing code. 9% figuring out why code doesn't work. 90% figuring out what's wrong with the data

89 1.9K 8.7K

Why it's important? in real life

"Manual inspection of errors in the CIC-ES and CIC-CAT datasets revealed that a substantial portion of misclassifications were due to dataset quality issues rather than model shortcomings. In our annotated sample, approximately 60% of errors in both languages were attributable to incorrect labels or tweets that were unrelated to the topic of Catalan independence. [...]"

"These findings highlight a critical point: much of the perceived model underperformance may be due to label inconsistencies and annotation errors. In fact, the evaluated LLMs may be classifying better than the raw scores indicate. However, confirming this hypothesis would require systematic re-annotation or cleaning of the CIC dataset. " [Masters' thesis of Summer Devlin, M2 student in 2025]

Wooclap

Is a corpus the same as a dataset?

Reminder

Corpus

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic [and/or extra-linguistic] criteria in order to be used as a sample of the language [Sinclair, 1996]

Corpus or dataset?

- The Penn TreeBank?

Corpus or dataset?

- The Penn TreeBank?
- MMLU?

Corpus or dataset?

- The Penn TreeBank?
- MMLU?
- texts used to pretrain ChatGPT?

Definition

Dataset

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files [Wikipedia, consulted on August 20th, 2025]

Sinclair, J. (1996).
Preliminary recommendations on corpus typology.
Technical report, Eagles.