UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

How to Do AI Research?

Technology and Innovation for Culture, Education, Language, and Society
Research Group
Y. Sigit Purnomo W.P., Ph.D.

Onsite Training Riset Berbasis Artificial Intelligence
27 Agustus 2025

1

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Outline

- There Is No AI Without Data
- Everyone Wants To Do The Model Work, Not The Data Work
- AI Modelling
- Research Example

2

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

There Is No AI Without Data

3

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

There Is No AI Without Data

- **In general, it is nothing new that data preparation and data quality are key for AI and data analytics, as there is no AI without data**
  - *There Is No AI Without Data: Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises*
    - *Christoph Gröger. 2021. **There is no AI without data**. Commun. ACM 64, 11 (November 2021), 98–108. https://doi.org/10.1145/3448247*

4

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

SO WHAT?

5

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

FIND THE DATA!

6

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

HOW?

7

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

TAKE THE TIME
TO
LOOK AROUND US

8

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- Culture
  - Ancient transcripts
  - Inscriptions
  - Various types of fabrics and motifs
  - Cultural data related to manners
  - Audio data of regional musical instruments
  - Cultural heritage data

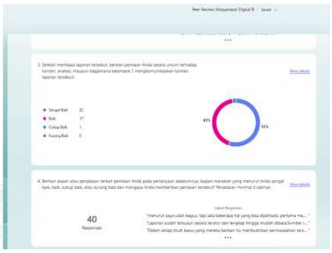
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
9

9

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- Education
  - Student exam/assignment results data
  - Peer review data
  - Teaching evaluation data
  - Student research (TA) results data
  - Student design/code data
  - LMS site log data
  - Labs data

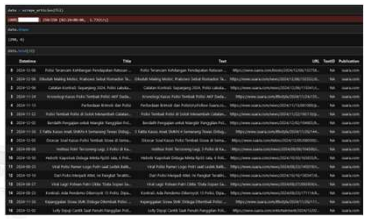


10

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- Language
  - Text in regional languages
  - Audio/Video in regional languages
  - Text data from various news sites
  - Text data from various social media
  - Opensource application UI/UX data

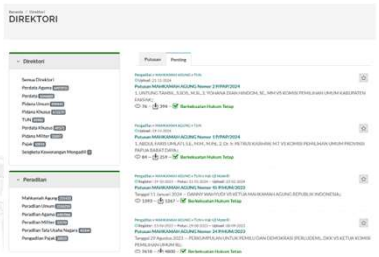


11

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- Society
  - Data on rules or policies
  - Data related to disasters
  - Data related to climate change
  - Poverty data



12

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- **Banking & Finance**
  - financial and fiscal documents, risk assessment, and market analysis

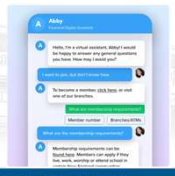
- **Media & Content Creation**
  - multi-media data, archives and user-generated content.

13

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- **Conversation & Insights**
  - conversational or dialog applications and user interactions and intents



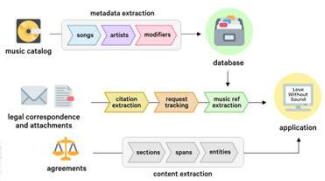
- **Healthcare & Biomedical**
  - medical documents, clinical and patient notes, pharmaceutical data and scientific research

14

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Source

- **Legal & Insurance**
  - legal and regulatory documents, legislative text, compliance process or domain-specific contracts.



15

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Everyone Wants To Do The Model Work, Not The Data Work

16

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Everyone wants to do the model work, not the data work

- Paradoxically, data is the most under-valued and de-glamorised aspect of AI
  - "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI
    - *Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021.* ***"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI****. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15. https://doi.org/10.1145/3411764.3445518*

17

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

SO WHAT?

18

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

START WITH THE DATA WORK!

19

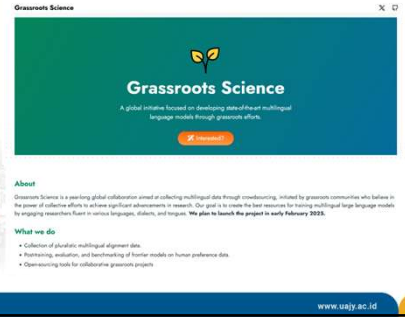
UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

HOW?

20

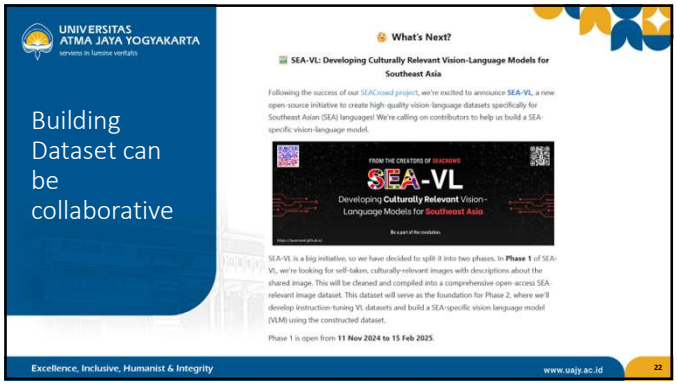
UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Building Dataset can be collaborative

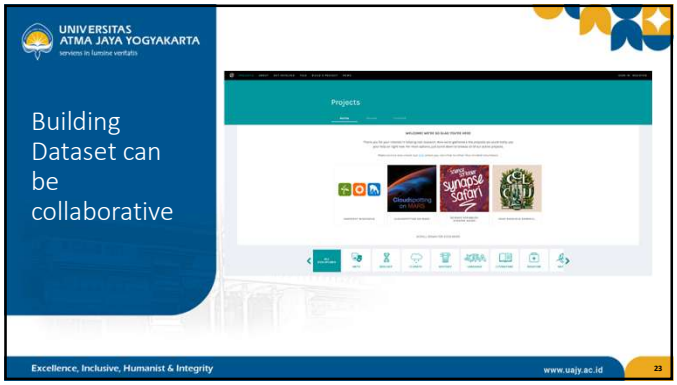


Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 21

21


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
Building Dataset can be collaborative
What's Next?
SEA-VL: Developing Culturally Relevant Vision-Language Models for Southeast Asia
Following the success of our SEACrowd project, we're excited to announce SEA-VL, a new open-source initiative to create high-quality vision-language datasets specifically for Southeast Asian (SEA) languages! We're calling on contributors to help us build a SEA-specific vision-language model.
FROM THE CREATORS OF SEACROWD
SEA-VL
Developing Culturally Relevant Vision-Language Models for Southeast Asia
SEA-VL is a big initiative, so we have decided to split it into two phases. In Phase 1 of SEA-VL, we're looking for self-taken, culturally-relevant images with descriptions about the shared image. This will be cleaned and compiled into a comprehensive open-access SEA-relevant image dataset. This dataset will serve as the foundation for Phase 2, where we'll develop instruction-tuning VL datasets and build a SEA-specific vision language model (VLM) using the constructed dataset.
Phase 1 is open from 11 Nov 2024 to 15 Feb 2025.
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
22

22


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
Building Dataset can be collaborative
Projects
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
23

23


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
NOW, WE ALREADY HAVE THE DATA, THEN?

24

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

ANOTATE THE DATA

25

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

HOW?

26

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Data Annotation Software

Data Annotation Service for University Students & Teachers

Enhance your research and learning with our easy-to-use Prodigy-based data annotation tool.

Why Choose Our Service?

Our platform provides an efficient, AI-powered data annotation tool designed specifically for students and teachers in academic settings.

Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 27

27

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Prodigy Features

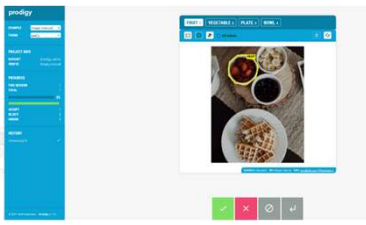
- **Information Extraction**
  - Extracting structured data like names, key phrases or relations is a crucial task for many business applications.

28

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Prodigy Features

- **Computer Vision**
  - Prodigy makes it easy to bring structure to your images and build custom AI systems for classification, segmentation and object detection.

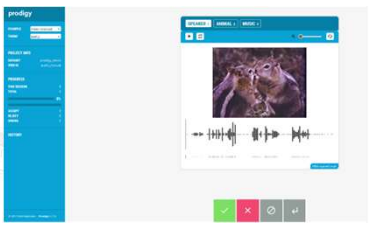


29

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Prodigy Features

- **Audio & Video Annotation**
  - Prodigy makes it easy to bring structure to your audio and video data, and build custom AI systems for classification, segmentation and transcription.



30

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Prodigy Features

- **Prompt Engineering**
  - Prodigy's prompt engineering workflows help you find and evaluate the best Large Language Model prompts for any custom use case.

31

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Prodigy Features

- **Language Model Training**
  - Distill knowledge from large models into smaller, faster and fully private pipelines for your use case that you can run cheaply and efficiently in-house.

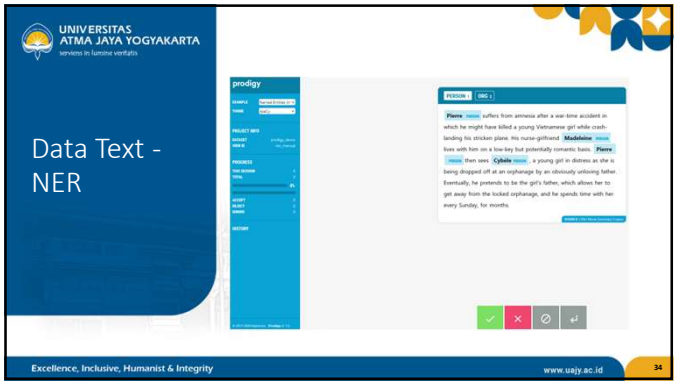
32

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

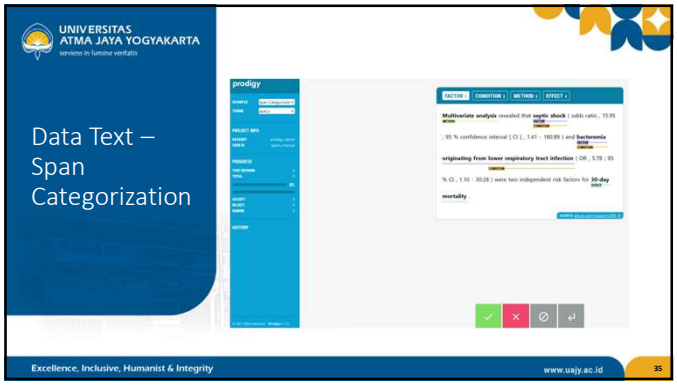
Prodigy Features

- **Custom Workflow**
  - Customize Prodigy for your specific use case and implement powerful automated workflows, interfaces and integrations.

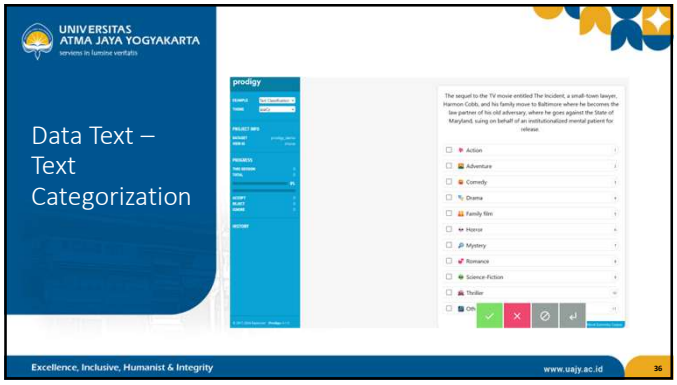
33


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text - NER
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
34

34


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text – Span Categorization
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
35

35


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text – Text Categorization
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
36

36


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text –
POS Tag
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
37

37


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text –
Dependency
Parsing
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
38

38


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text –
Relations
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
39

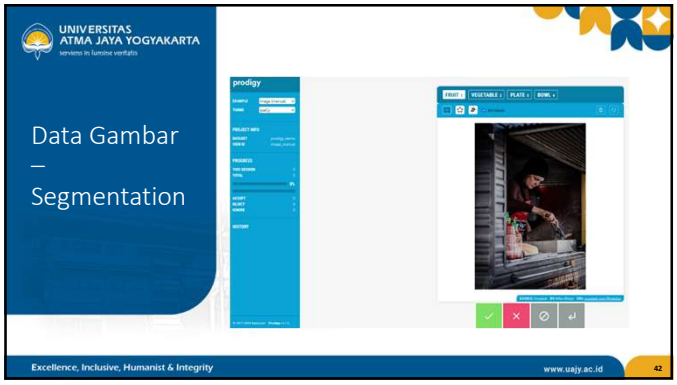
39


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text – Text Comparation
prodigy
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
40

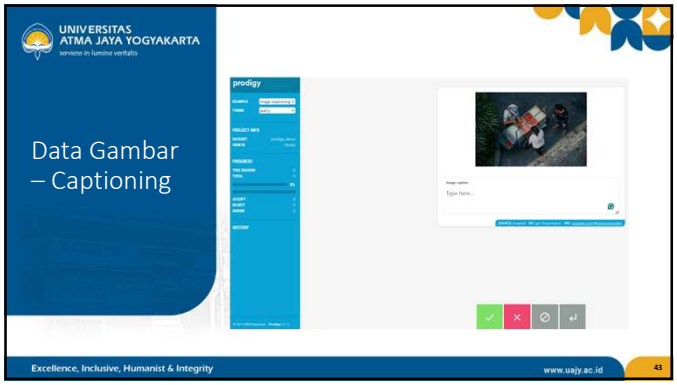
40


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Text – Question Answer
prodigy
What are the best education portals in India?
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
41

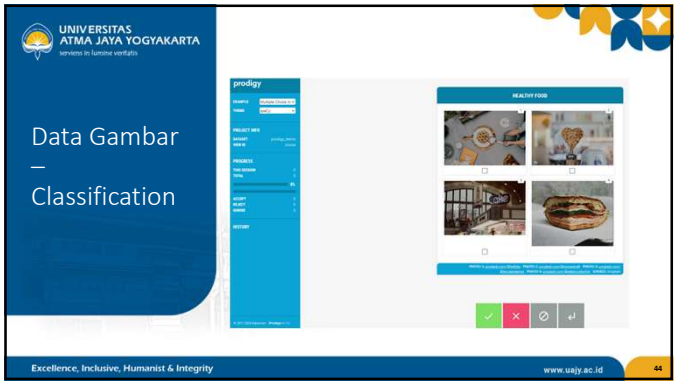
41


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Gambar – Segmentation
prodigy
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
42

42


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Gambar – Captioning
prodigy
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
43

43


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Gambar – Classification
prodigy
HEALTHY FOOD
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
44

44


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Audio – Classification
prodigy
SPEAKER
ANIMAL
MUSIC
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
45

45


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Audio –
Transcription
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
46

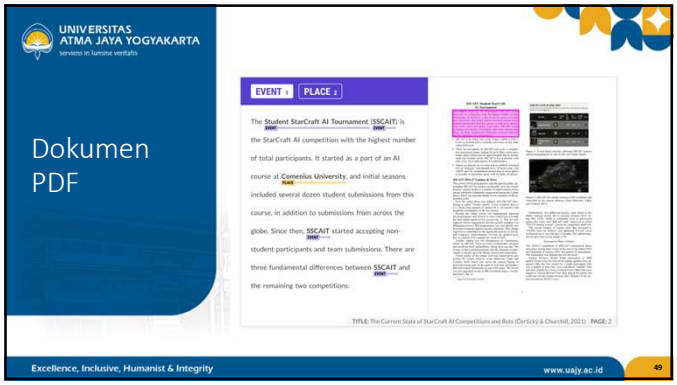
46


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Data Video –
Classification
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
47

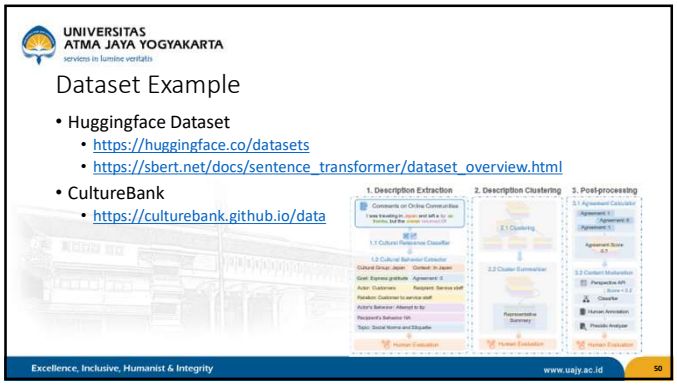
47


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Dokumen
PDF
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
48

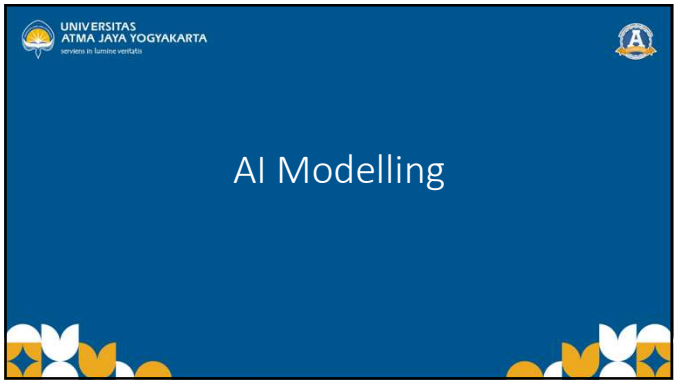
48


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
Dokumen
PDF
EVENT 1
PLACE 2
The Student StarCraft AI Tournament (SSCAIT) is
the StarCraft AI competition with the highest number
of total participants. It started as a part of an AI
course at Comenius University, and initial seasons
included several dozen student submissions from this
course, in addition to submissions from across the
globe. Since then, SSCAIT started accepting non-
student participants and team submissions. There are
three fundamental differences between SSCAIT and
the remaining two competitions:
TITLE: The Current State of StarCraft AI Competitions and Bots (Certicky & Churchill, 2021) PAGE: 2
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
49

49


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
Dataset Example
• Huggingface Dataset
• https://huggingface.co/datasets
• https://sbert.net/docs/sentence_transformer/dataset_overview.html
• CultureBank
• https://culturebank.github.io/data
1. Description Extraction
2. Description Clustering
3. Post-processing
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
50

50


UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis
AI Modelling

51

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

HOW?

52

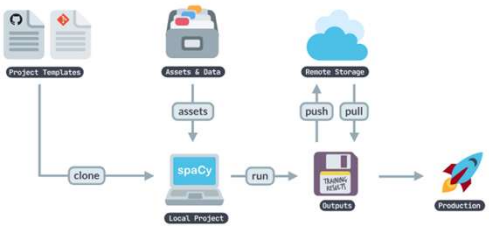
UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

I DON'T KNOW
HOW TO CODE

53

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

PRODIGY and SPACY



Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id
54

54

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

PRODIGY and SPACY

- spaCy projects let you manage and share end-to-end spaCy workflows for different use cases and domains, and orchestrate training, packaging and serving your custom pipelines.
  - You can start off by cloning a pre-defined project template, adjust it to fit your needs, load in your data, train a pipeline, export it as a Python package, upload your outputs to a remote storage and share your results with your team
    - https://github.com/explosion/projects?tab=readme-ov-file

55

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

PRODIGY and SPACY

- Clone the project template you want to use
  - `python -m weasel clone tutorials/ner_fashion_brands`
- Install any project requirements.
  - `cd ner_fashion_brands`
  - `python -m pip install -r requirements.txt`
- Fetch assets (data, weights) defined in the project.yml
  - `python -m weasel assets`

56

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

PRODIGY and SPACY

- Run a command defined in the project.yml.
  - `python -m weasel run preprocess`
- Run a workflow of multiple steps in order.
  - `python -m weasel run all`
- Run the visualization
  - `python -m spacy project run visualize`
- Adjust the template for your specific use case, load in your own data, adjust the settings and model and share the result with your team.

57

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

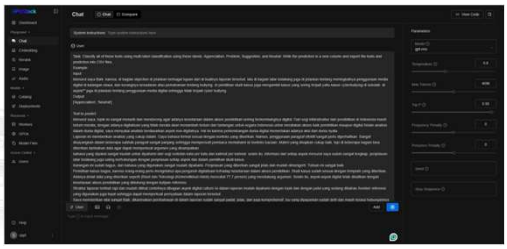
Zero/Few-shot Prompting

- Zero-shot prompting asks a large language model (LLM) to perform a task using only an instruction (no examples)
  - It relies on the model's prior, instruction-following ability
- Few-shot prompting adds a handful of in-context examples (input→output pairs) inside the prompt, so the model infers the task pattern on the fly: no parameter updates, just smarter conditioning

58

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

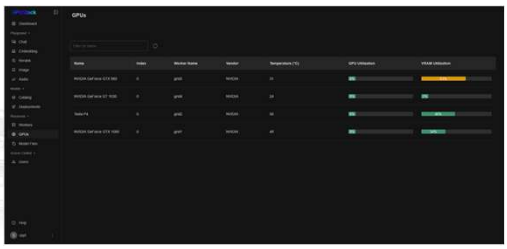
Zero/Few-shot Prompting



59

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Zero/Few-shot Prompting



60

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

I KNOW
HOW TO CODE

61

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

WE HAVE OUR HPC(High Performance Computing)

Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 62

62

UNIVERSITAS
ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

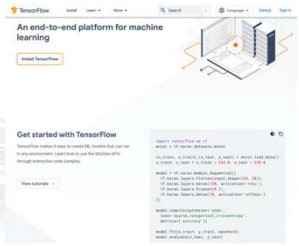
WE HAVE MANY PYTHON LIBRARY

Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 63

63

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

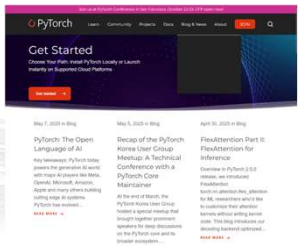
WE HAVE MANY PYTHON LIBRARY



64

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

WE HAVE MANY PYTHON LIBRARY



65

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

DON'T WORRY, MANY RESOURCES AVAILABLE

- Scikit Learn
  - https://scikit-learn.org/stable/auto_examples/index.html
- Keras
  - https://keras.io/examples/
- Tensorflow
  - https://www.tensorflow.org/tutorials
- PyTorch
  - https://docs.pytorch.org/tutorials/
- Open-Source AI Cookbook
  - https://huggingface.co/learn/cookbook/index

66

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

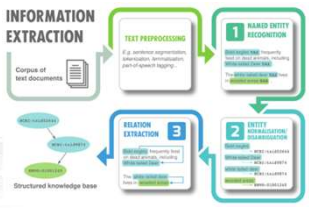
Research Example

67

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Biology

- Case Study: TaxoNERD that provides deep neural network (DNN) models to recognise taxon mentions in ecological documents
  - To achieve high performance, these models usually need to be trained on a large corpus of manually annotated text. Creating such a corpus is a laborious and costly process, with the result that manually annotated corpora in the ecological domain tend to be too small to learn an accurate DNN model from scratch. To address this issue, we leverage existing models pretrained on large biomedical corpora using transfer learning.
    - https://doi.org/10.1111/2041-210X.13778

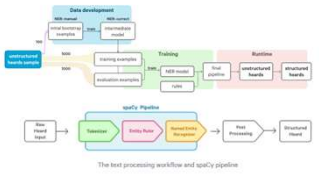


68

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Banking and Finance

- Case Study: How S&P Global is making markets more transparent with NLP
  - This case study shows how the S&P Global Commodity Insights team built and shipped impressively efficient information extraction pipelines for real-time commodities trading insights in a high-security environment, and how they were able to achieve a 10× speed-up of their data collection and annotation workflows using human-in-the-loop distillations with LLMs, spaCy and Prodigy.
    - https://explosion.ai/blog/sp-global-commodities



The text processing workflow and spaCy pipeline

	Global Carbon Credits	Americas Crude Oil	Asia Steel Rebar
Accuracy (F-score)	0.95	0.96	0.99
Speed (words/second)	15,730	13,908	16,015
Model Size	6 MB	6 MB	6 MB
Training Examples	1,598	1,695	1,368
Evaluation Examples	211	200	345
Data Development Time	~15h	~15h	~15h

69

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Media & Content Creation

- Case Study: How the Guardian approaches quote extraction with NLP
  - To facilitate trust, human-in-the-loop workflows are widespread in media applications as stakeholders require the ability to teach and to evaluate models through human-AI interfaces.
    - https://explosion.ai/blog/guardian

70

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Conversation & Insights

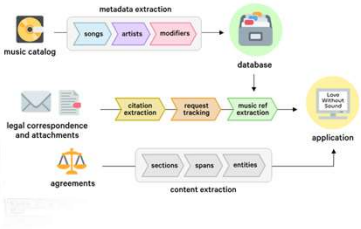
- Case Study: Deploying a Prodigy cloud service for Posh's financial chatbots
  - Posh focuses on developing custom NLP models trained on real-world banking conversations and custom models for each client's unique customer base and product offering.
    - https://explosion.ai/blog/posh-prodigy-financial-chatbots

71

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Legal & Insurance

- Case Study: How Love Without Sound helps music industry law firms recover millions in revenue for artists
  - This case study shows how Love Without Sound built innovative AI-powered tools for the music industry and law firms specializing in royalty negotiations, and helped publishers recover hundreds of millions of dollars in lost revenue for artists using spaCy and Prodigy.
    - https://explosion.ai/blog/love-without-sound-nlp-music-industry

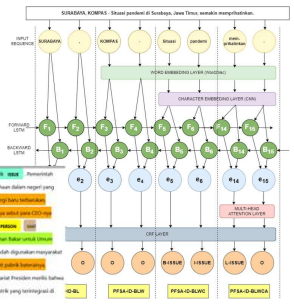


72

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Computational Journalism

- Case Study: Extraction and attribution of public figures statements for journalism in Indonesia using deep learning



Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 73

73

UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Reference

- *Christoph Gröger. 2021.* ***There is no AI without data****. Commun. ACM 64, 11 (November 2021), 98–108. DOI: https://doi.org/10.1145/3448247*
- *Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021.* ***"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI****. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15. https://doi.org/10.1145/3411764.3445518*
- https://prodi.gy

Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 74

74

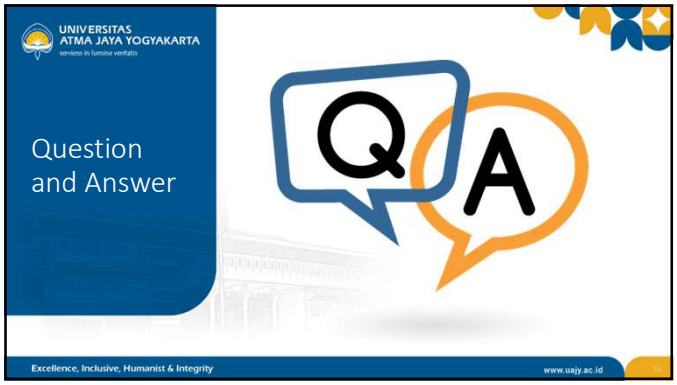
UNIVERSITAS ATMA JAYA YOGYAKARTA
serviens in lumine veritatis

Reference

- Yohanes Sigit Purnomo W.P., Yogan Jaya Kumar, Nur Zareen Zulkarnain, Basit Raza. 2024. **Extraction and attribution of public figures statements for journalism in Indonesia using deep learning**. Knowledge-Based System, Volume 289, 8 April 2024, 111558 - DOI: https://doi.org/10.1016/j.knosys.2024.111558
- Yohanes Sigit Purnomo W.P., Yogan Jaya Kumar, Nur Zareen Zulkarnain. 2022. **PFSA-ID: An Annotated Indonesian Corpus and Baseline Model of Public Figures Statements Attributions**. Global Knowledge, Memory and Communication - DOI: https://10.1108/GKMC-04-2022-0091

Excellence, Inclusive, Humanist & Integrity www.uajy.ac.id 75

75


UNIVERSITAS
ATMA JAYA YOGYAKARTA
Question
and Answer
Q
A
Excellence, Inclusive, Humanist & Integrity
www.uajy.ac.id

76