ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
CAMPUS DI FORLÌ

91258 / B0385
Natural Language Processing

Lesson 1. Introduction

Alberto Barrón-Cedeño
a.barron@unibo.it

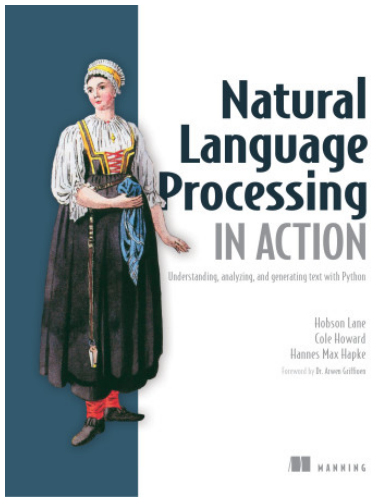
30/09/2024

Table of Contents

Materials

Core Bibliography

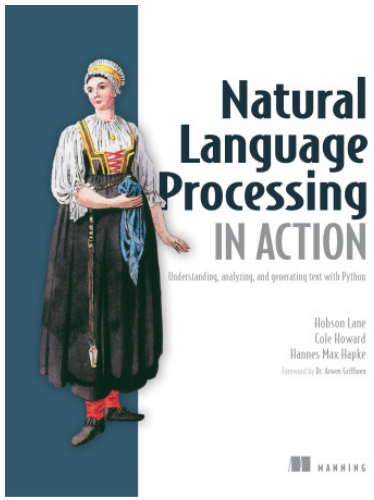
1. Lane et al. (2019)'s Natural Language Processing in Action[1]
2. Numerous Wikipedia articles on relevant topics
3. Multiple online forums



[1] https://www.manning.com/books/natural-language-processing-in-action

Core Bibliography

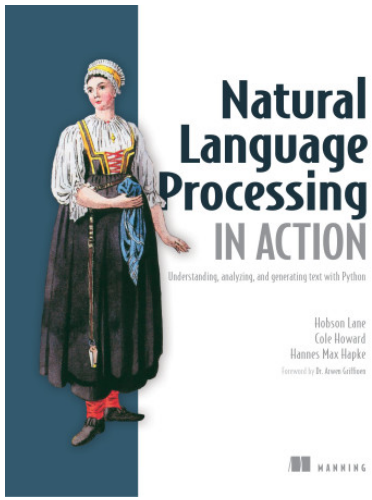
1. Lane et al. (2019)'s Natural Language Processing in Action[1]
2. Numerous Wikipedia articles on relevant topics
3. Multiple online forums

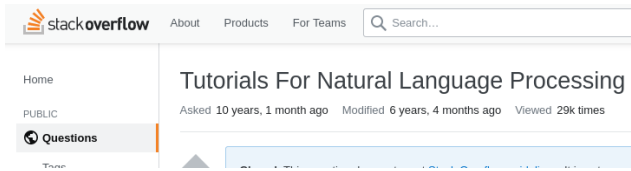


[1] https://www.manning.com/books/natural-language-processing-in-action

Core Bibliography

1. Lane et al. (2019)'s Natural Language Processing in Action[1]
2. Numerous Wikipedia articles on relevant topics
3. Multiple online forums





[1] https://www.manning.com/books/natural-language-processing-in-action

Complementary Bibliography

1. Intro to computing for text
   - K.W. Church's Unix for poets[2]
2. For social media analysis
   - Hovy (2021)'s Text Analysis in Python for Social Scientists*[3]
3. A basic intro in Italian
   - Nissim and Pannitto (2022)'s Che cos'è la linguistica computazionale
4. From linguistics
   - Bender (2013)'s Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax[4]
5. Advanced
   - Koenigstein (2024)'s Transformers in Action*[5]

[2] https://web.stanford.edu/class/cs124/kwc-unix-for-poets.pdf
[3] https://doi.org/10.1017/9781108873352
[4] https://doi.org/10.2200/S00493ED1V01Y201303HLT020
[5] https://www.manning.com/books/transformers-in-action

The NLP environment this year

1. **Tutorato**
   K. Korre (3rd year PhD student) will offer 10 lessons of tutorato.
   `https://moodle.dipintra.it`

2. **Selected Topics in NLP** (3 cfu)
   Optional lesson covering complementary (non mandatory) topics
   `https://www.unibo.it/it/studiare/dottorati-master-specializzazioni-e-altra-formazione/insegnamenti/insegnamento/2024/508811`

[6] For all three initiatives.

The NLP environment this year

1. **Tutorato**
   K. Korre (3rd year PhD student) will offer 10 lessons of tutorato.
   https://moodle.dipintra.it

2. **Selected Topics in NLP** (3 cfu)
   Optional lesson covering complementary (non mandatory) topics
   https://www.unibo.it/it/studiare/dottorati-master-specializzazioni-e-altra-formazione/insegnamenti/insegnamento/2024/508811

Lesson coordinates

Slides, code, calendar[6] and more are all available at:
albarron.github.io/teaching/natural-language-processing

[6] For all three initiatives.

Tools

Essential

Python 3 development framework on any modern OS

1. Command line or
2. Integrated development Environment; e.g., Pycharm[7], Eclipse[8] or
3. Jupyter notebook; e.g., Google's colab[9], local Jupyter[10]

[7] https://www.jetbrains.com/pycharm/
[8] https://www.eclipse.org
[9] https://colab.research.google.com
[10] https://jupyter.org
[11] Could be part of Selected topics/tutorato
[12] https://gitlab.com
[13] https://github.com

Tools

Essential

Python 3 development framework on any modern OS

1. Command line or
2. Integrated development Environment; e.g., Pycharm[7], Eclipse[8] or
3. Jupyter notebook; e.g., Google's colab[9], local Jupyter[10]

Desirable[11]

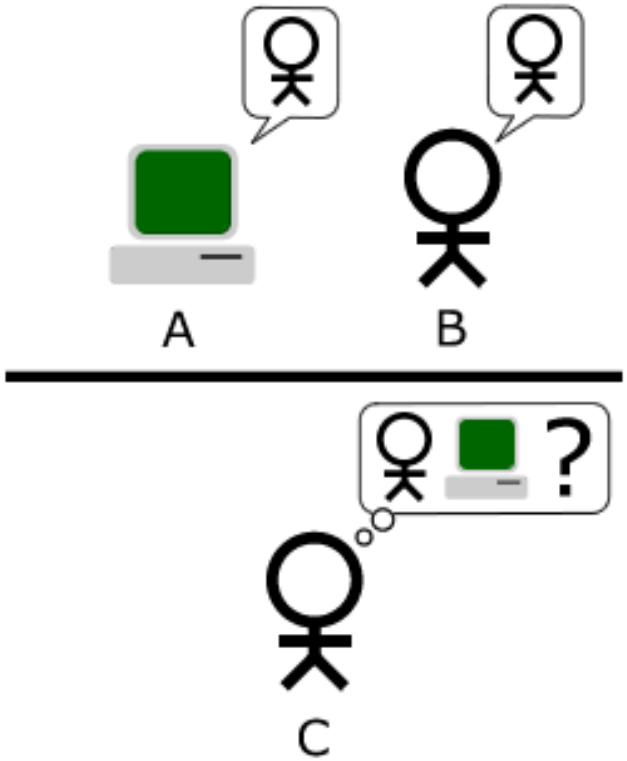
1. Git Version control system; e.g., Gitlab[12] or Github[13]
2. LaTeX system for document preparation

[7] https://www.jetbrains.com/pycharm/
[8] https://www.eclipse.org
[9] https://colab.research.google.com
[10] https://jupyter.org
[11] Could be part of Selected topics/tutorato
[12] https://gitlab.com
[13] https://github.com

Introduction

Introduction

Natural language as a measure of intelligence



Turing (1950). "Computing machinery and intelligence". Mind. 59(236)

Picture from
`upload.wikimedia.org/wikipedia/commons/e/e4/Turing_Test_version_3.png`

Computational linguistics[14]

- Interdisciplinary field concerned with the computational (it used to say "statistical or rule-based"!) modeling of natural language

[14] https://en.wikipedia.org/wiki/Computational_linguistics
[15] https://en.wikipedia.org/wiki/Natural_language_processing

Computational linguistics[14]

- Interdisciplinary field concerned with the computational (it used to say "statistical or rule-based"!) modeling of natural language
- Study of appropriate computational approaches to linguistic questions

[14] https://en.wikipedia.org/wiki/Computational_linguistics
[15] https://en.wikipedia.org/wiki/Natural_language_processing

Computational linguistics[14]

- Interdisciplinary field concerned with the computational (it used to say "statistical or rule-based"!) modeling of natural language
- Study of appropriate computational approaches to linguistic questions

Natural Language Processing[15]

[14] https://en.wikipedia.org/wiki/Computational_linguistics
[15] https://en.wikipedia.org/wiki/Natural_language_processing

Computational linguistics[14]

- Interdisciplinary field concerned with the computational (it used to say "statistical or rule-based"!) modeling of natural language
- Study of appropriate computational approaches to linguistic questions

Natural Language Processing[15]

- Interdisciplinary subfield of computer science and artificial intelligence (it used to say "linguistics"!)[...] concerned with providing computers the ability to process data encoded in natural language

[14] https://en.wikipedia.org/wiki/Computational_linguistics
[15] https://en.wikipedia.org/wiki/Natural_language_processing

Computational linguistics[14]

- Interdisciplinary field concerned with the computational (it used to say "statistical or rule-based"!) modeling of natural language
- Study of appropriate computational approaches to linguistic questions

Natural Language Processing[15]

- Interdisciplinary subfield of computer science and artificial intelligence (it used to say "linguistics"!)[...] concerned with providing computers the ability to process data encoded in natural language
- Data is collected in text corpora, using either rule-based, statistical or neural-based approaches in machine learning and deep learning

[14] https://en.wikipedia.org/wiki/Computational_linguistics
[15] https://en.wikipedia.org/wiki/Natural_language_processing

Natural Language Processing (Lane et al., 2019, p. 4)

- Area of research in computer science and artificial intelligence concerned with processing natural languages
- This processing generally involves translating natural language into data (numbers) that a computer can use to learn about the world

[16] https://en.wikipedia.org/wiki/Computational_linguistics

Natural Language Processing (Lane et al., 2019, p. 4)

- Area of research in computer science and artificial intelligence concerned with processing natural languages
- This processing generally involves translating natural language into data (numbers) that a computer can use to learn about the world

The term natural language processing is nowadays considered to be a near-synonym of computational linguistics and (human) language technology.[16]

[16] https://en.wikipedia.org/wiki/Computational_linguistics

Rule-based vs Statistical NLP

Models are based on a number of hand-crafted rules or grammars

Models are based on a number of hand-crafted rules or grammars

Diagram borrowed from L. Moroney's Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning

Introduction

Rule-based NLP

Models are based on a number of hand-crafted rules or grammars

```
greeting_inputs = ("hey", "morning", "evening", "hi",
                   "whatsup", "hello")
greeting_responses = ["hey", "hey hows you?", "*nods*",
                   "hello, how you doing", "hello",
                   "Welcome, I am good and you"]

def generate_greeting_response(input):
    for token in input.split():
        if token.lower() in greeting_inputs:
            return random.choice(greeting_responses)
```

Derived from

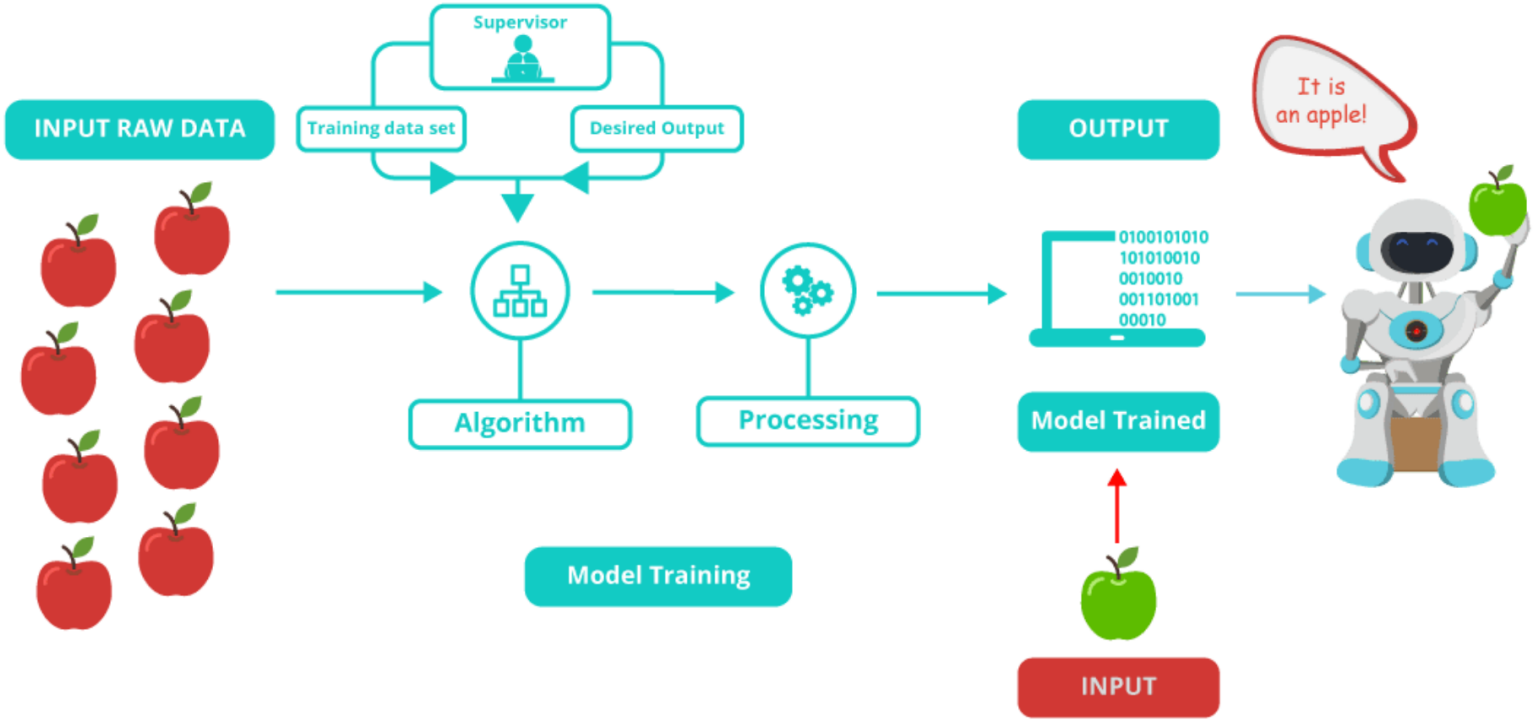
https://stackabuse.com/python-for-nlp-creating-a-rule-based-chatbot/

Models are tuned on *annotated* data

Models are tuned on *annotated* data

Diagram borrowed from L. Moroney's Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning

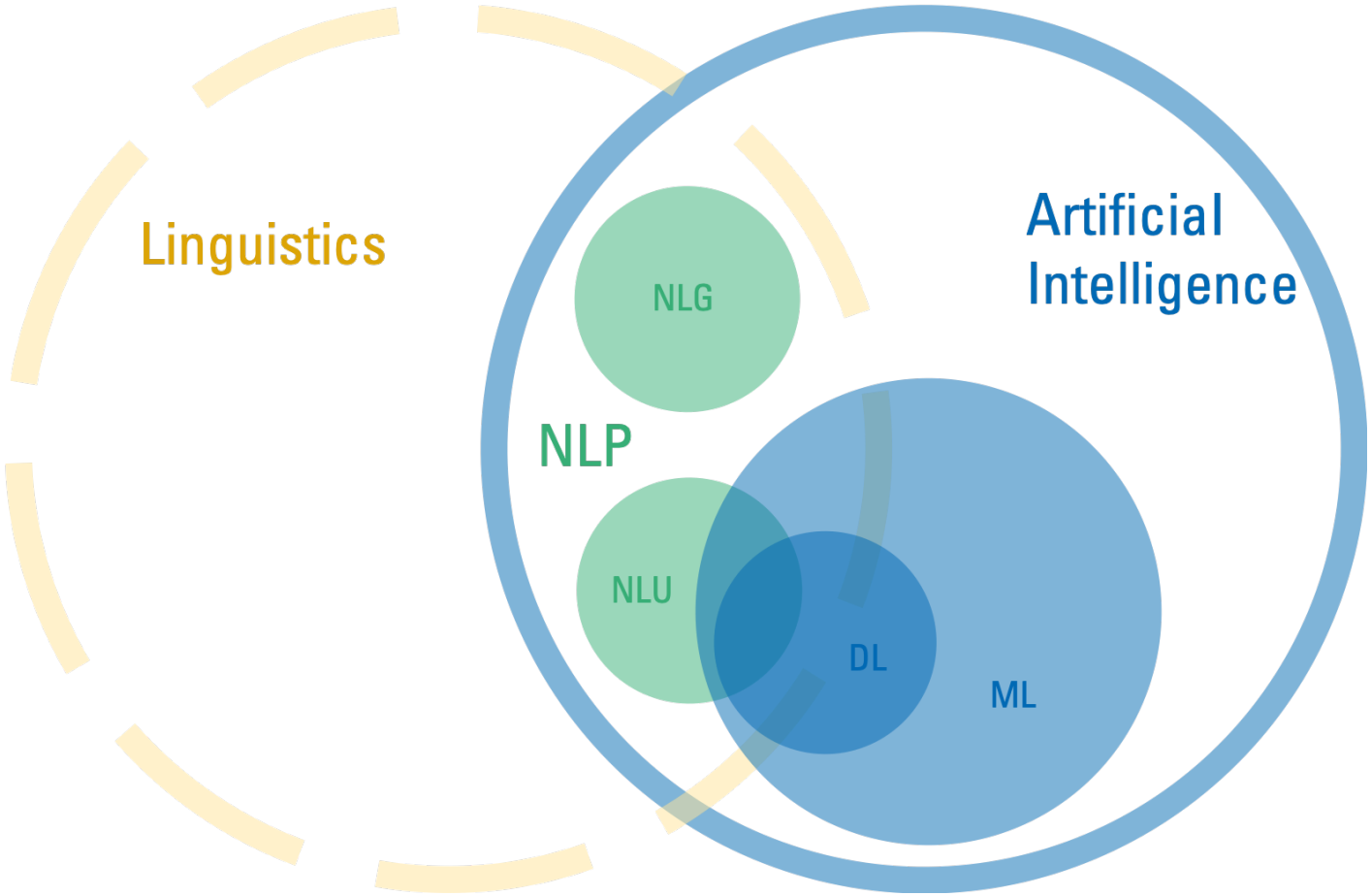
Models are tuned on annotated data



Borrowed from `https://www.edureka.co/blog/machine-learning-tutorial`

Introduction

The NLP neighborhood



Borrowed from
`https://www.retresco.de/en/how-to-ai-natural-language-processing/`

Introduction

Non-exhaustive list of NLP applications with examples

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation
News analysis	event identification · fact checking

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation
News analysis	event identification · fact checking
Forensics	plagiarism detection · authorship attribution

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation
News analysis	event identification · fact checking
Forensics	plagiarism detection · authorship attribution
Sentiment analysis	product review ranking · opinion mining

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation
News analysis	event identification · fact checking
Forensics	plagiarism detection · authorship attribution
Sentiment analysis	product review ranking · opinion mining
Creative writing	text generation with a narrative and style

Partially derived from (Lane et al., 2019, p. 8)

Introduction

Non-exhaustive list of NLP applications with examples

Search	web search engines · text autocompletion
Editing	grammar issues identification
Dialogue	chatbot creation
Email	spam filtering · message classification
Text mining	(multi-)document summarisation
News analysis	event identification · fact checking
Forensics	plagiarism detection · authorship attribution
Sentiment analysis	product review ranking · opinion mining
Creative writing	text generation with a narrative and style
Translation	translation · quality estimation

Partially derived from (Lane et al., 2019, p. 8)

The philosophy of this lesson

1. Concept understanding
2. Know what you are doing and why
3. In position to go forward

Requirements & Evaluation

Requirements

Necessary

- Linguistics
- Algebra
- Programming in Python

[17] Part of the tutorato

Requirements

Necessary

- Linguistics
- Algebra
- Programming in Python

Desirable

- Intermediate programming (e.g., object-oriented, testing)
- High-performance computing (e.g., slurm)[17]

[17] Part of the tutorato

Evaluation

Final project: 80%

You will address a relevant problem...

- within the range of your own (research) interests or
- participating (formally) in a shared task or
- proposed by me, if you prefer

Homework: 20%

- mostly programs addressing relatively small problems

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...
   IF you plan for a publication[18]
   - We meet regularly to see the advances and shape the experiments, submissions, and/or paper towards the submission deadline

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...
   IF you plan for a publication[18]
   - We meet regularly to see the advances and shape the experiments, submissions, and/or paper towards the submission deadline

   ELSE
   - We could meet sporadically, if you need it

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...
   IF you plan for a publication[18]
   - We meet regularly to see the advances and shape the experiments, submissions, and/or paper towards the submission deadline

   ELSE
   - We could meet sporadically, if you need it
3. You submit a written report ($\sim$ 7 pages),[19] your implementation, results 1 week before the *appello*

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Typical final project pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...
   IF you plan for a publication[18]
   - We meet regularly to see the advances and shape the experiments, submissions, and/or paper towards the submission deadline

   ELSE
   - We could meet sporadically, if you need it
3. You submit a written report ($\sim$ 7 pages),[19] your implementation, results 1 week before the *appello*
4. We meet on the date of the appello to discuss about your project, in the context of the lecture

[18] Talk to me well in advance; it would require my heavy involvement

[19] I do not like (student) novels

Evaluation

Final mark

(Beside the 20% for homework) a combination of the quality of the experiments, report, code, and oral discussion

Evaluation

Final mark

(Beside the 20% for homework) a combination of the quality of the experiments, report, code, and oral discussion

Targeting 30L?

If I let you submit a paper, it is very likely. But it is not the only way. . .

Evaluation

Final mark

(Beside the 20% for homework) a combination of the quality of the experiments, report, code, and oral discussion

Targeting 30L?

If I let you submit a paper, it is very likely. But it is not the only way...

$$p(30L \mid \text{paper submitted} == \mathit{True}) \approx 0.85 \quad (1)$$

$$p(30L \mid \text{paper submitted} == \mathit{False}) \approx 0.15 \quad (2)$$

Evaluation

Previous final projects

2023–2024

- 📰 Multilingual Persuasion Technique Detection in News *♠
- 𝕎 Tracing source languages in Wikipedia articles

2022–2023

- 🎮 Sentiment analysis of video game reviews
- ⚖ Authorship attribution: machine vs human

2021–2022

- ✖ Hate Speech Detection in Incel Online Spaces *•♠
- ⚥ Fishing for catfishes: predicting the author gender in Reddit

* student with previous programming skills

• turned into (part of a) thesis ♠ turned into a publication

Evaluation

Previous final projects

2020–2021

- Semantic similarity between originals and machine translations •
- Definition extraction on food-related Wikipedia articles •
- Identifying Characters' Lines in Original and Translated Plays •
- Classifying an Imbalanced Dataset with CNN, RNN, and LSTM

2019–2020

- AriEmozione: Identifying Emotions in Opera Verses *•♠
- UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AIBERTo *•♠

* students with previous programming skills

• turned into (part of a) thesis ♠ turned into a publication

Visit the projects section of the class website for details, reports and papers

References

Bender, E. M.
2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers.

Hovy, D.
2021. *Text Analysis in Python for Social Scientists: Discovery and Exploration*, Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Koenigstein, N.
2024. *Transformers in Action* . Shelter Island, NY: Manning Publication Co.

Lane, H., C. Howard, and H. Hapkem
2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.

Nissim, M. and L. Pannitto
2022. *Che cos'è la linguistica computazionale*. Carocci editore.