UNIVERSITÄT BONN

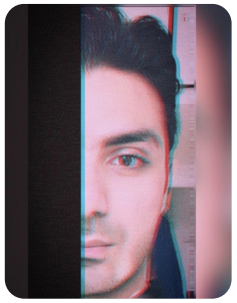
CAISA Lab
https://caisa-lab.github.io/

b-it Bonn-Aachen International Center for Information Technology

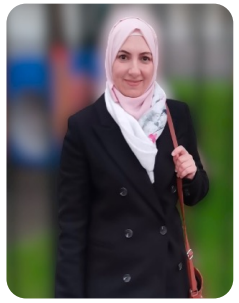
INTRODUCTION TO NATURAL LANGUAGE PROCESSING

INSTRUCTORS & TEACHING ASSISTANTS

Prof. Dr. Lucie Flek
Head of CAISA Lab
flek@bit.uni-bonn.de



Vahid Sadiri Javadi
Course Coordinator
vahidsj@bit.uni-bonn.de



Farizeh Aldabbas
Teaching Assistant
farizeh@uni-bonn.de

Ulvi Shukurzade
Teaching Assistant
ulvi@uni-bonn.de

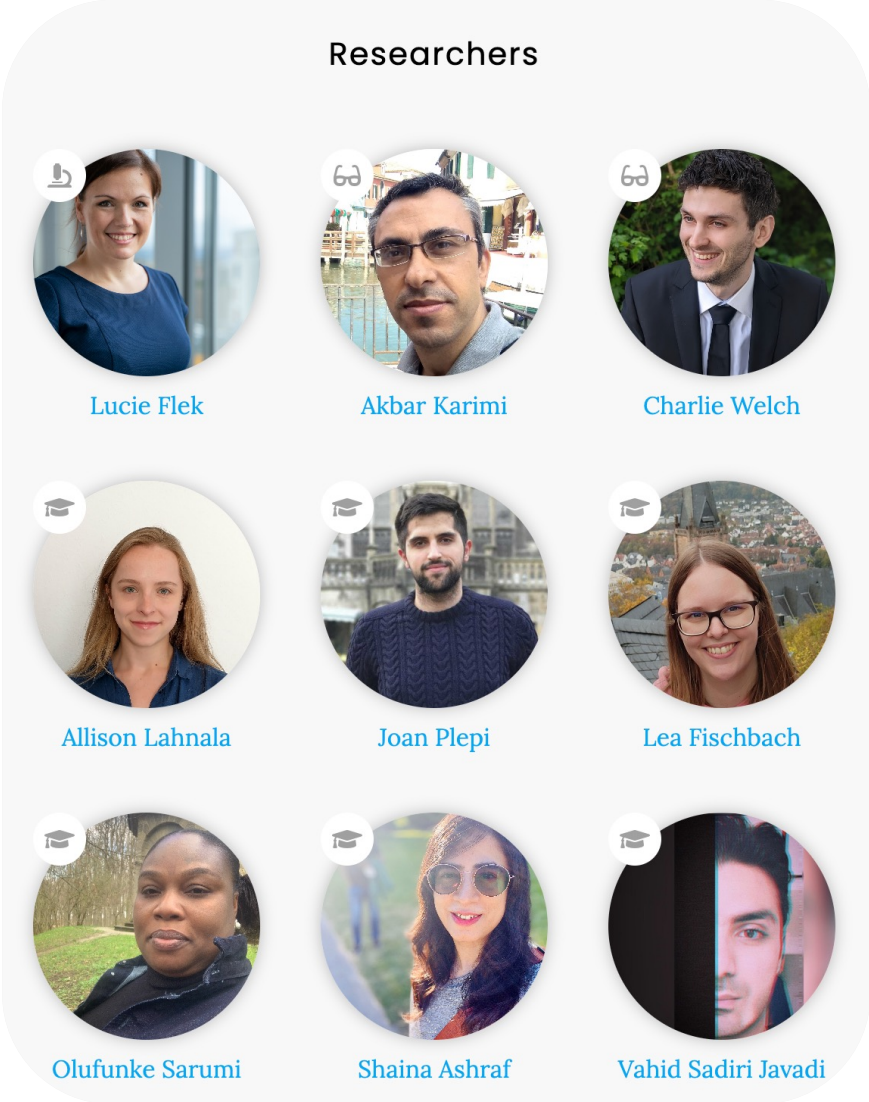
Lectures: **Thursday** 10:30 – 12.00 (B-IT-Max 0.109) (Zoom Link)

Exercises: **Wednesday** **- Group 1 (Vahid):** 14:15 - 15:45 (B-IT-Max 0.109) (Zoom Link)

- Group 2 (Ulvi): 16:00 - 17:30 (B-IT-Max 0.109) (Zoom Link)

eCampus Course

ABOUT OUR GROUP



Conversational AI and Social Analytics (CAISA) Lab

- Natural Language Processing
- **Conversational AI**
- User Modeling and Computational Social Science
- Machine Learning for NLP
- Privacy and Fairness in Natural Language Applications

https://caisa-lab.github.io

THESIS TOPICS

Theses Topics

(Master/ Bachelor)

Data Quality Improvement; Evaluate Arguments for Generating Opinionated Sales Conversations

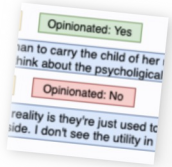
Argumentation and debating are the process of forming reasons that humans engage in. Sales negotiation is one of the conversational activities in which a buyer and a seller communicate reasons to arrive at a satisfactory selection

Implicit Hate Speech Detection

Implicit hate speech is defined by coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity. Compared to explicit hate speech detection, implicit hate

Exploring user context for detecting intended and perceived sarcasm

Sarcasm is a form of irony that occurs when there is a discrepancy between the literal meaning of an utterance and its intended meaning. Existing sarcasm detection systems focus on exploiting linguistic markers, context, or userlevel



A poor man's opinion detection tool, training a model with a handful of data

Opinion detection aims to detect an author's view towards a certain topic and has become a key component in everyday applications such as fake news detection and argumentation. While state-of-the-art deep learning models are

Investigating the level of stubbornness regarding sociopolitical views in social media

While social media platforms help to connect people worldwide and give access to enormous amounts of diverse information, they also foster an environment that promotes polarization. This occurs due to the fact that users

https://caisa-lab.github.io/theses/

AGENDA

Today, we will talk about:

- **Survey: Link**
- **Organization**
- **Introduction to Python**
- **Software Setup**
- **Jupyter Notebook**
- **Recap & Questions**

COURSE OUTLINE

Content of Course:

⚠ **Holiday:** 01.11 ⚠

Week 0: **25.10.2023 | Introduction & Python basics**

Feature Engineering:

Week 1: 08.11.2023 | Word operations & Feature extraction using Pandas, Sklearn

Week 2: 15.11.2023 | Linear classification using TF - IDF

Language Processing:

Week 3: 22.11.2023 | Word embeddings using spaCy

Week 4: **29.11.2023 | Q & A: PF + PS**

Week 5: 06.12.2023 | Transformers and Generative Models I

Week 6: 13.12.2023 | Transformers and Generative Models II

Week 7: 20.12.2023 | POS tagging & HMMs

Week 8: 10.01.2023 | Project development (supervision by appointment)

Week 9: 17.01.2023 | Project development (supervision by appointment)

Week 10: 24.01.2023 | Project development (supervision by appointment)

Week 11: **31.01.2023 | PROJECT PRESENTATIONS (PP)**

COURSEWORK

1. Assignments (Prerequisite for the exam)

- You need to achieve **at least 50%** of the credits to be allowed to take the exam.

2. Final Project (40%)

- It offers you the chance to apply your newly acquired skills to an in-depth application using different frameworks such as PyTorch and spaCy.

3. Exam (60%)

- A written exam containing both lecture and exercise topics.
- Answers in German will be also accepted.

ASSIGNMENTS

- All assignments will use Python as the programming language.

- **Collaboration:** Assignments will be shared on eCampus and should be submitted via eCampus either **individually** or **in a group of two students!**

- **Deadlines:** All assignments are due **on Tuesday before the exercise class at 11:59 PM.**

- **Grade/ Feedback:** You will receive your graded assignment one week after the submission deadline on eCampus.

- Please comment your code.

- Please name your submitted file with your name/ both student names.
  File name: <FirstName_LastName>

Prerequisite for the exam (<u>at least 50%</u>):

Assignment 1 (10%): Word Operations

Assignment 2 (20%): Text Classification (Scikit-Learn)

Assignment 3 (20%): Word Vectors (SpaCy)

Assignment 4 (30%): Fine-tuning with LLMs (Hugging Face)

Assignment 5 (20%): Hidden Markov Model

FINAL PROJECT

Project Components:

- Problem Formulation (PF) (10%)
- Problem Solving (PS) (15%)
- Project Poster (PP) (5%)
- Project Report (PR) (10%)

Project Types:

- Default Project
- Resource Creation Project
- Robustness and Reproducibility Project

- Submission of team members (a group of 3 up to 5):

- Wednesday, November 8th, 23:59

*Subject: ITNLP - WS2023 - **<Matr. Nr.>***
*Team Speaker: <Name>, **<Matr. Nr.>,** <Mail Addr.>*
Team Members: <Name>, <Matr. Nr.>, <Mail Addr.>
<Name>, <Matr. Nr.>, <Mail Addr.>

* In case, you need a teammate:

Subject: ITNLP - WS2023 - Looking for a team
*<Name>, **<Matr. Nr.>,** <Mail Addr.>*

E-Mail: itnlp.uni.bonn@gmail.com

- Submission of Problem Formulation (PF): (PDF)

- Tuesday, November 28th, 23:59

- Submission of Project Presentation (PP): (Poster)

- Thursday, January 25th, 23:59 *

- Submission of Project Report (PR) + Problem Solving (PS)

- Sunday, February 11th, 23:59

eCampus: INTRODUCTION TO NLP

NOW, LET'S TALK ABOUT PYTHON!

A BIT OF CONTEXT …

"Programming languages are how programmers express and communicate ideas — and the audience for those ideas is other programmers, not computers."
-Guido van Rossum

- Created by Guido van Rossum
  - Dutch Programmer
  - Since November 2020 at Microsoft
    - Check his blog or this interview if you want to know more
- Described by its creator as "a programming language created by a community"

- The first version of Google was written in Python and Java

PYTHON INSTALLATION


```
def install_python:
    if not Python already installed:
        Install Python following instructions from
        python.org
    return “Installed Python!”
```

WE NEED VIRTUALENV…

Option 1	Option 2
pip install virtualenv	easy_install virtualenv

Follow these steps in your terminal:

> mkdir intro2nlp

> cd intro2nlp

> virtualenv --python python<3.7*> nlpenv
(*Check the installed Python version and change it accordingly.)

> source nlpenv/bin/activate

> pip install jupyterlab

> pip install -r requirements.txt

> jupyter lab

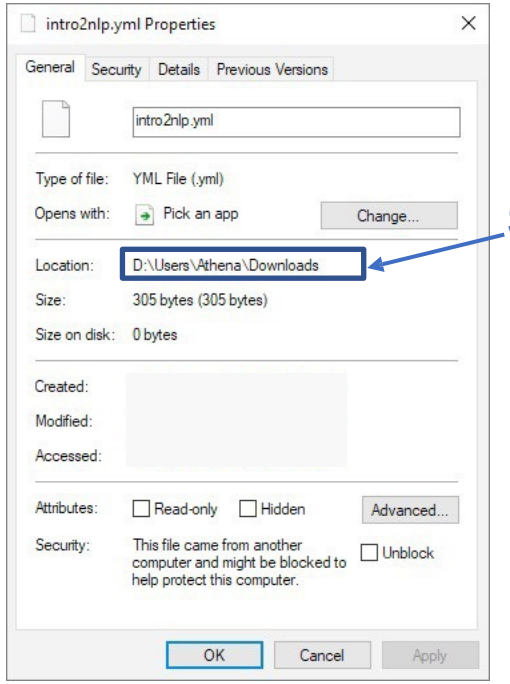
LINKS:

Virtual Python Environment (Windows)

Jupyter Lab Installation

Jupyter Notebook Installation

OR CONDA IN WINDOWS



Step 4.

Step 5.

1. Download Miniconda in this link
2. Find our "*intro2nlp.yml*" file on eCampus
3. Open the Anaconda Prompt on your computer
4. Navigate to the directory where you saved the yml-file (eg: *cd /Downloads/*)
5. Type the following line:
   "conda env create -f intro2nlp.yml"
6. Now you're ready to activate your environment with:
   "conda activate intronlp"
7. Last step navigate to your class directory and type:
   "jupyter notebook"

IF YOU STILL HAVE PROBLEMS WITH INSTALLATION

- You might consider using the notebook directly from Colaboratory.
- It is a Programming Platform offered by Google. In this case, you would need a Google Account.

See you next Wednesday!