UNIVERSITÄT BONN

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

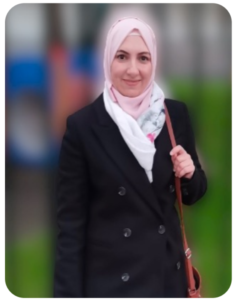
b-it Bonn-Aachen International Center for Information Technology

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** 10:15 – 11:45 (B-IT-Max 0.109) ([Zoom Link]())

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

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

[eCampus Course]()

ANNOUNCEMENT

Announcements:

- Submission of Team Members ✅

- You can find the list of teams HERE (Forum)
- This semester, we have 14 Teams (62)

- Assignment #1 ✅

- Received 10 submissions until Monday 11:59 PM
- Received 40 submissions until Tuesday 11:59 PM
- **JN File** (.ipynb), not any other file formats like .zip
- You will receive your graded assignment by next week

- Submission of Problem Formulation (PF):

- **Deadline:** Sunday, Nov 28th, 23:59
- **Guideline:** eCampus >> Project >> Problem Formulation (PF) - Guidelines
- **Submission:**
  - **What:** PDF
  - **Where:** eCampus >> Project >> Problem Formulation (PF) - Submissions >> File name: **Team_<num.>**

ANNOUNCEMENT

Announcements:

- **Datasets for <u>Default Project</u>:**
- You need to fill out this form to get access to datasets: LINK
- Add your team number to the dataset table if you choose a dataset.
- New "**INTERESTING**" datasets are welcome!
  - but you need to contact me beforehand!

- **Exercise**
  - The exercise (Group 1) on 22.11 will be held **IN-PERSON & ONLINE!**

COURSE OUTLINE

Content of Course:

Week 1: 25.10.2023 | Introduction & Python basics

Feature Engineering:

Week 3: 08.11.2023 | Word operations & Feature extraction using Pandas, Sklearn

Week 4: **15.11.2023 | Linear classification using TF - IDF**

Language Processing:

Week 5: 22.11.2023 | Word embeddings using spaCy

Week 6: **29.11.2023 | Q & A: PF + PS**

Week 7: 06.12.2023 | Transformers and Generative Models I

Week 8: 13.12.2023 | Transformers and Generative Models II

Week 9: 20.12.2023 | POS tagging & HMMs

Week 10: 10.01.2024 | Project development (supervision by appointment)

Week 11: 17.01.2024 | Project development (supervision by appointment)

Week 12: 24.01.2024 | Project development (supervision by appointment)

Week 13: **31.01.2024 | PROJECT PRESENTATIONS (PP)**

AGENDA

Today, we will talk about:

- Assignment #1
- Linear Classification

LINEAR CLASSIFICATION

using TF - IDF

TF - IDF

What is TF-IDF?

- **TF - IDF** is an information retrieval or information extraction subtask that aims to express the importance of a word to a document which is part of a collection of documents (corpus).

- **TF – IDF** is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithms for prediction.

- **Term Frequency,** which measures how frequently a term occurs in a document.

- **Inverse Document Frequency,** which measures how important a term is.

CALCULATE TF - IDF

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

$$TF\text{-}IDF = TF * IDF$$

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term} + 1})$$

We weight the frequency of each term in a document, with its relevance in the corpus:

$$tf_{t,d} = \log(count(t,d) + 1)$$

$$idf_t = \log \frac{N}{df_t}$$

$$\text{tf-idf} = tf_{t,d} \cdot idf_t$$

Example:

- We are in an NLP exercise class.
- We have many students here.
- We want to learn how to represent documents.

PRACTICAL SESSION ON JUPYTER NOTEBOOK

See you next Wednesday!