UNIVERSITÄT BONN

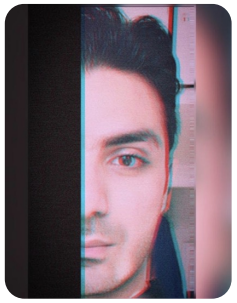
CAISA Lab
https://caisa-lab.github.io/

b-it
Bonn-Aachen
International Center for
Information Technology

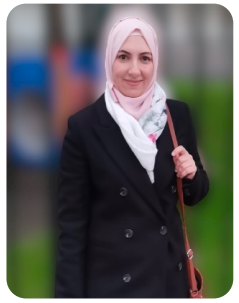
INTRODUCTION TO NATURAL LANGUAGE PROCESSING

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:

- **Zoom Links**
  - Posted on the course page and slides
- **Updating the Survey!**
  - Survey (Forum >> Survey)
  - For new students in the course (LINK)
- **Submission of Team Members:**
  - Received some team members
  - **Deadline:** Wednesday (Tonight), **Nov 8th**, 23:59
  - You will find the list of teams on our course page soon
  - The team speaker is our contact person
- **Our Forum:**
  - Introduction to Natural Language Processing/ Discussion Forum
- **Teams:**
  - Team for final project (3 - 5)
  - Team for assignment submission (1 - 2)

ANNOUNCEMENT

Announcements:

- Assignments

- Will be uploaded on eCampus after the exercise.
- We will discuss the assignment every week.
- For submission, name your file as follows:

"Assignment_1_<Your_Name>.ipynb"

"Assignment_1_<Your_Name>__<Your_Name>.ipynb"

Ex. Assignment_1_FirstName_LastName.ipynb

- Where?

eCampus >> ITNLP >> Student Submissions

- Exercise

- The exercise (Group 1) on 15.11 will be held **ONLY ONLINE ON ZOOM!**

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:

- **Word Operation**
- **Feature Extraction**

WORD OPERATIONS

WORDS vs TOKENS

How many words?
How many Tokens?

"Let us learn tokenization."

A **word-based tokenization algorithm** will break the sentence into words. The most common one is splitting based on space.

["Let", "us", "learn", "tokenization."]

A **subword-based tokenization algorithm** will break the sentence into subwords.

["Let", "us", "learn", "token", "ization."]

A **character-based tokenization algorithm** will break the sentence into characters.

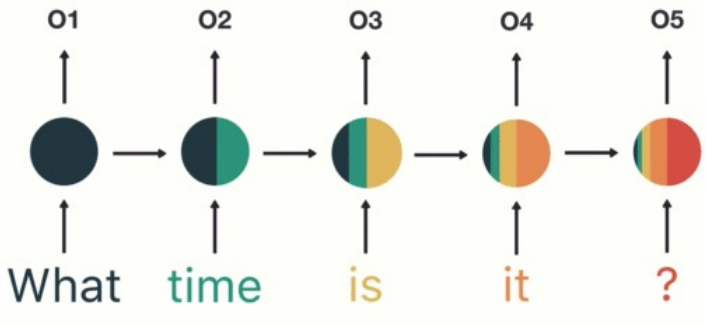
["L", "e", "t", "u", "s", "l", "e", "a", "r", "n", "t", "o", "k", "e", "n", "i", "z", "a", "t", "i", "o", "n", "."]

TOKENIZATION

The true reasons behind tokenization?

As tokens are the building blocks of Natural Language, the most common way of processing the raw text happens at the token level.

For example, Transformer based models – the State of The Art (SOTA) Deep Learning architectures in NLP – process the raw text at the token level. Similarly, the most popular deep learning architectures for NLP like RNN, GRU, and LSTM also process the raw text at the token level.



LEMMATIZATION vs STEMMING

Why do we need this?

- For grammatical reasons, documents are going to use different forms of a word, such as ***organize, organizes, and organizing***.

- Additionally, there are families of derivationally related words with similar meanings, such as ***democracy, democratic, and democratization***.

- In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

LEMMATIZATION

- A **lemma** is a word that represents a whole group of words, and that group of words is called a **lexeme**.

- am, are, is ⇒be
- car, cars, car's, cars'⇒ car

Let's do it together:

- Barack Obama was born in Hawaii.

Word	Lemma
Barack	Barack
Obama	Obama
was	be
born	bear
in	in
Hawaii	Hawaii
.	.

STEMMING

Why Stemming is important?

- To build a robust model, it is essential to normalize text by removing repetition and transforming words to their base form through stemming.

- **Stemming** is a text processing task in which you reduce words to their root, which is the core part of a word.

1. Porter Stemmer – PorterStemmer()

```
Connects ---> connect
Connecting ---> connect
Connections ---> connect
Connected ---> connect
Connection ---> connect
Connectings ---> connect
Connect ---> connect
```

2. Snowball Stemmer – SnowballStemmer()

```
generous ---> generous
generate ---> generat
generously ---> generous
generation ---> generat
```

PRACTICAL SESSION ON JUPYTER NOTEBOOK

FEATURE EXTRACTION

using Sklearn & Pandas

CLASSIFICATION TASK

task	x	y
language ID	text	{english, mandarin, greek, …}
spam classification	email	{spam, not spam}
authorship attribution	text	{jk rowling, james joyce, …}
genre classification	novel	{detective, romance, gothic, …}
sentiment analysis	text	{postive, negative, neutral, mixed}

Given training data in the form of <x, y> pairs, learn the mapping function

$$h'(x) = y$$

which is as close as it gets to the ideal (unknown)

$$h(x)=y$$

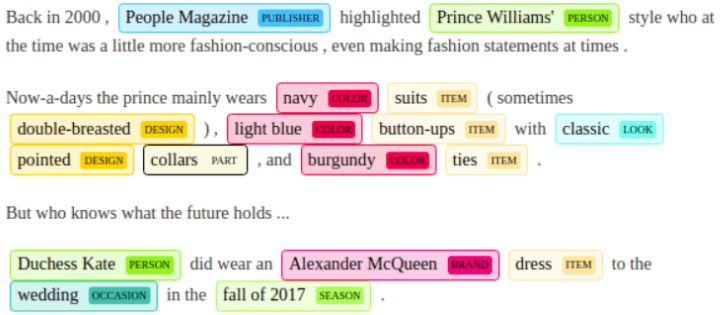
Given your training data samples x with labels y.

WORD REPRESENTATION

How can we represent words?

What can we extract from words?

- IDs
- Frequency
- Part of Speech
- Co-occurrence
- Named Entities



LOOKUP TABLE

Word	Id
and	0
document	1
first	2
is	3
one	4
second	5
the	6
third	7
this	8

Corpus =

- This is the first document.
- This document is the second document.
- And this is the third one.
- Is this the first document?

Sent 1	8	3	6	2	1	
Sent 2	8	1	3	6	5	1
Sent 3	0	8	3	6	7	4
Sent 4	3	8	6	2	1	

COUNT VECTORS

Sent 1	0	1	1	1	0	0	1	0	1
Sent 2	0	2	0	1	0	1	1	0	1
Sent 3	1	0	0	1	1	0	1	1	1
Sent 4	0	1	1	1	0	0	1	0	1

Corpus =

- This is the first document.
- This document is the second document.
- And this is the third one.
- Is this the first document?

Word	Id
and	0
document	1
first	2
is	3
one	4
second	5
the	6
third	7
this	8

ONE-HOT ENCODING

	Restaurant Reviews
R1	Great restaurant and great service !
R2	They can do better to provide better service
R3	Only two thumbs up, worst service ever

Entire Corpus

Set of all the words in the corpus
great
restaurant
and
service
they
can
do
better
to
provide
only
Two
thumbs
up
worst
ever

Set of all the words in the corpus	R1: Great Restaurant and great service !	R2: They can do better to provide better service	R3: Only two thumbs up, worst service ever
great	1	0	0
restaurant	1	0	0
and	1	0	0
service	1	1	1
they	0	1	0
can	0	1	0
do	0	1	0
better	0	1	0
to	0	1	0
provide	0	1	0
only	0	0	1
Two	0	0	1
thumbs	0	0	1
up	0	0	1
worst	0	0	1
ever	0	0	1

PRACTICAL SESSION ON JUPYTER NOTEBOOK

See you next Wednesday!