Introduction, administrivia

Statistical Methods in NLP 1
ISCL-BA-06

Çağrı Çöltekin
/ʧaɾ'ɯ ʧʏltec'in/
ccoltekin@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft

Winter Semester 2026–2027

version: 5f0f272 2025-09-30

What is this course about?

- This is the first course in the two-course series on *Natural Language Processing*
- A course on foundations of machine learning
- Focus on theory and concepts (a bit of practice in the second part)

Prerequisites:

- Formally: ASW-BA-01, ISCL-BA-01, ISCL-BA-04
- Realistically: High-school math and some programming (in Python)

Module: ISCL-BA-06 (formerly called Parsing)

What is NLP?

```
NLP is a set of methods for computational processing of natural
languages.
```

- Is it a subfield of:
  - Artificial Intelligence?
  - Machine Learning?
  - Computer Science?
  - Information Retrieval? – Wikipedia claims so.
- Is it the same as Computational Linguistics?
- How much linguistics is needed / used?

Text (sequence) classification

- Text classification is a very common NLP task
- Given a text we often want to assign one or more labels
- Assignment of the label(s) requires some level of 'language understanding'

Text classification: some examples

Is it spam?

```
From: Dr Pius Ayim <mikeabass15@gmail.com>
Subject: Dear Friend / Lets work together
Dear Friend,
My name is Dr. Pius Anyim, former senate president of the Republic Nigeria
under regime of Jonathan Good-luck.
I am sorry to invade your privacy; but the ongoing ANTI-CORRUPTION GRAFT
agenda of the ruling government is a BIG problem that I had to get your
contact via a generic search on internet as a result of looking for a
reliable person that will help me to retrieve funds I deposited at a
financial institute in Europe.
—
```

* From my 'spambox' which I stopped checking regularly long time ago.

Text classification: some examples

Is the customer happy?

```
I never understood what's the BIG deal behind this album. Yes,
the production is wonderfull but the songwriting is childish and
rubbish. They definitly can not write great lyrics like Bob
Dylan sometimes do. "God Only Know" and "Wouldnt Be nice" are
indeed masterpieces...but the rest of the album is background
music.
```

```
€DB_Bahn mußten sie für den Sauna-Besuch zuzahlen ?
```

- Sentiment analysis is one of the popular applications of text classification

Text classification: some examples

which language is this text in?

```
Član 3. Svako ima pravo na život, slobodu i osobnu sigurnost.
```

- Detecting language of the text is often the first step for many NLP applications.
- Easy for the most part, but tricky for
  - closely related languages
  - text with code-switching

Text classification: More examples

- Who wrote the book?
- Find the author's
  - age
  - gender
  - political party affiliation
  - native language
- Is the author/speaker depressed?
- What is the proficiency level of a language learner?
- What grade should a student essay get?
- What is the diagnosis, given a doctor's report?
- What category should a product be listed based on its description?
- What is the genre of the book?
- Which department should answer the support email?
- Is this news about
  - politics
  - sports
  - travel
  - economy
- Is the web site an institutional or personal web page?

Sequence labeling

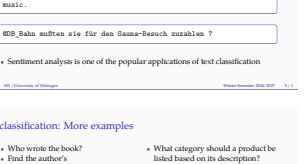
Example: entity recognition

UN	Secretary-General	Antonio	Guterres
ORG	NONE	PER	PER

plans	to	visit	Ukraine
NONE	NONE	NONE	GEO

- Typical entities of interest include: people, organizations, locations
- Can be application specific, e.g., drug/disease names, chemical components, legal entities
- Many NLP problems can be cast as sequence labeling problems: extractive summarization, question answering, ...

Relation extraction



- For many other tasks, we do not only need entities, but the relations between them
- Other similar applications include: dependency parsing, semantic role labeling, ...

Text generation

Example: machine translation



Text generation

Other examples

- Dialogue systems / 'intelligent' agents
- Question answering
- Summarization
- Caption generaion
- Data to text generation

Models capable of generating text (LLMs) is where most of the research is focused currently.

Why is it challenging

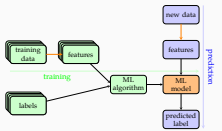
- NLP deals with combinatorial problems: high computational complexity
- Natural language is categorical on the surface, but with many nuanced/continuous differences underneath
- Linguistic input is variable length, most (if not all) ML methods require fixed-length input
- Data sparsity: many linguistic objects (e.g., words) occur rarely
- Natural language is ambiguous, interpretations depends on context

Some examples of ambiguity

(newspaper headlines)

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE
- KIDS MAKE NUTRITIOUS SNACKS
- DRUNK GETS NINE MONTHS IN VIOLIN CASE
- MINERS REFUSE TO WORK AFTER DEATH

Anatomy of a (supervised) data-driven solution



What is in this course?

A bird's eye view

Introductory lectures on

- Background
  - Linear algebra
  - Calculus (mainly derivatives)
  - Probability and information theory
- Generalization, bias, variance
- Some fundamental classification methods
- Unsupervised learning

Course overview

- Lectures (VG 0.02)
  - Tuesday 16:15-17:45
  - Thursday 10:15-11:45
- Tutor: How Chuen Yu (Kyle)
  - Tutorial time TBA
- Public course website: `https://is1-nlp1.github.io/`
- Moodle: `https://moodle.zdv.uni-tuebingen.de/course/view.php?id=1979`
  - Please fill in the 'beginning of semester survey' before the next course session
- GitHub: `https://github.com/is1-nlp1/nlp1-2026`

Literature

- No textbook
- Reading recommendations will be provided for each course (all online, freely accessible)

Coursework and evaluation

- Reading material for most lectures
- Assignments: ungraded, but **required**
- Final (written) exam
- Attendance is not required, but you are unlikely to pass without regular attendance

Assignments

- Some paper & pencil assignments (through Moodle)
- 3 programming assignments (after semester break, through GitHub)
- The programming assignments can be done in pairs (recommended – knowing your classmates, and learning from them, is an important part of the university experience/education)
- This means **working together on the whole exercise**, not sharing parts of an assignment and working on them independently

Final remarks

- Please do not be shy, ask your questions during the lectures
- Please take the assignments seriously, learning ML/NLP requires practice
- Please fill in the 'beginning of semester survey' on Moodle:
  `https://moodle.zdv.uni-tuebingen.de/mod/feedback/view.php?id=122937`
- Next:
  - Linear algebra introduction
  - Suggested ~~reading~~/video lectures:
    `https://www.youtube.com/playlist?list=PLZ-GT3cokr2y29Krbmu3vDLSr44Bv205S`
- Time for your questions

Acknowledgments, credits, references

*