NLP without Annotated Dataset

Course Overview

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

8 January 2021

Today's plan

- Course overview
- NLP overview

Course Overview

About me

- I work as a full time researcher at the National Research Council, Canada in Digital Technologies Research Center.
- 2020: Wrote a book for O'Reilly (http://www.practicalnlp.ai/)
- 2018-19: Senior Data Scientist in software engineering r&d teams in Toronto
- 2016-18: Assistant Professor (tenure track) at Iowa State University, USA
- 2011-15: PhD at University of Tuebingen, Germany
- Before that: software developer, Bachelors/Masters in Engineering

About You

- 22/46 filled the questionnaire so far (at 1600 CET).
- Mostly MA, followed by BA ISCL?
- Background: Mostly Linguistics, and many wrote they know some programming.
- Languages you speak: English, German, Italian, Spanish, Portuguese, Swedish, Arabic, Russian, Mandarin, Korean,Japanese, Thai, Bahasa Indonesia, Acehnese!

About You

- 22/46 filled the questionnaire so far (at 1600 CET).
- Mostly MA, followed by BA ISCL?
- Background: Mostly Linguistics, and many wrote they know some programming.
- Languages you speak: English, German, Italian, Spanish, Portuguese, Swedish, Arabic, Russian, Mandarin, Korean,Japanese, Thai, Bahasa Indonesia, Acehnese!
- Why are you enrolled in this course? What do you want to do later?
  - A common answer: learn practical aspects of NLP and work in the industry
  - Do research on NLP for native languages (non English/German)

Teaching experience

- 2011-13: 2 Hauptseminar courses at Tuebingen (with Prof Meurers)
- 2016-18:
  - Applied Linguistics grad students: Python programming, Introduction to NLP
  - Grad Computer science students: Statistical NLP
  - Undergrad students from all disciplines : "Language and Computers", "Text as Data" (R), Technical Communication
- 2020: Guest course at Munich Graduate School of Economics, Germany (online)

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.
- However, what drives all of them are large collections of annotated corpora.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.
- However, what drives all of them are large collections of annotated corpora.
- What do you do when you don't have access to such datasets, though?

Course Objectives

- Provide an overview of NLP system development pipeline
- Discuss some common approaches for collecting, cleaning and exploring text data
- Introduce some methods to develop labeled data for NLP

Expected Learning Outcomes

Students should be able to:

- Understand the end to end NLP system development pipeline
- Compile and explore labeled/annotated corpora for NLP
- Build some basic text classification and information extraction systems

... upon successful completion of the course..

Pre-requisities

1. Intermediate proficiency in any programming language (Python preferred)
2. Comfortable installing libraries etc on their laptops
3. Knowledge of the usage of virtual environments (venv, anaconda) is useful

What the course can't do

- Don't expect to become an NLP expert with one compact course.
- Contents may not always meet your own expectations, but there is a term paper and a group discussion, which gives you opportunities to explore your specific interests related to this topic.
- The course won't teach you programming.

How we learn and grow

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

One fourth from the teacher, one fourth from own intelligence,
One fourth from classmates, and one fourth only with time.

AchAryAt pAdamAdatte, pAdam shiShyaH swamedhayA |
sa-brahmachAribhyaH pAdam, pAdam kAlakrameNa cha ||

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

Source

Course Logistics

Meeting and Location

- January 8 2021-January 29, 2021, M W F, 17:00 s.t. - 19:30 (Central European Time).
  1. 8th Jan 2021 (Friday)
  2. 11th, 13th, 15th Jan 2021 (Mon, Wed, Fri)
  3. 18th, 20th, 22nd Jan 2021 (Mon, Wed, Fri)
  4. 25th, 27th, 29th Jan 2021 (Mon, Wed, Fri)
- Location: Zoom Meeting-ID: 990 5086 7382
  Kenncode: 296817
- For a one to one meeting, email me to set up a time. I am keeping 1700-1800 free on most days in January for these one to one meetings.

Course Website

- Moodle: https://moodle.zdv.uni-tuebingen.de/course/view.php?id=1301
- Syllabus, Lecture slides and Assignments will be uploaded there.

Course Format + Credits

- Video lectures + Discussion (I may sometimes pick people randomly and ask a question!)
- Assignments (2)
- Team presentations: You are expected to form into groups of 2-4 people, pick a paper from the reading list on the website (or any other relevant paper) and present a brief discussion in a live session (10-15 minutes per group)
- Assignments
- Term paper(optional)

Credits: 3 CP (+ 3 CP if you write a term paper)

Textbooks

1. "Speech and Language Processing" by Jurafsky and Martin (2/3 editions)
2. "Practical Natural Language Processing" by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta and Harshit Surana.
3. NLTK book
4. For Python: "Python for Everybody" Charles Severance

(Details on how to access these books are in the Syllabus document)

Course Topics

1. Introduction (1 session)
2. NLP Pipeline (1 session)
3. Corpus collection, extraction, exploration (1 session)
4. Automatically labeling data (3 sessions)

remaining 4 sessions are for student presentations and review.

Assignments/Grading (for 6 CP)

1. 2 Assignments (30% of the grade)
2. 1 presentation (30% of the grade)
3. 1 term paper (30% of the grade)
4. classroom participation (10% of the grade)

(For 3 CP: Split the term paper grade between two assignments)

Assignments

- Two assignments, already uploaded on Moodle
- They are not difficult - the goal is not to trick you, but to make you think about the challenges of working with NLP problems in real world.
- My preferred programming language is Python, I am okay with Java, R, C, C++, and anything else (note: I can't debug for you. What you submit should run error-free on my machine).

Presentation

- Students can work in teams of 2-4 people and present one of the research papers related to course topics, from a given list of papers.
- Papers are listed in the syllabus document. If you want to present a different paper, talk to me first.
- Pick your teams early (deadline: 13th Jan)

Term Paper

- Work on a short project involving NLP and write a report describing your work (6-8 pages long in single column, latex formatted document)
- Some ideas are listed in the syllabus document. If you want to work on something else, talk to me first.
- If you want to get into NLP research later, explore some of your ideas through this term paper!

Classroom Participation

- Attending live meetings
- Participating in the forum
- Communicating (Asking questions, informing me if something comes up and you can't attend etc)
- Submitting stuff on time

Important Deadlines

1. Decide on a team for group discussion (13th Jan 2021)
2. Decide on a paper for group discussion (15th Jan 2021)
3. Group Discussions (22nd-27th Jan 2021)
4. Assignments 1 and 2 Submission (6th Feb 2021)
5. Decide on term paper topic (29th Jan 2021)
6. Term paper submission (13th Feb 2021)

- Questions so far?