Computer-Based Content Analysis I (Theory)
SURV 703
1 credit/2 ECTS
Winter 2018/2019

Instructor(s)
Christoph Kilian Theil, christoph@informatik.uni-mannheim.de
Heiner Stuckenschmidt, heiner@informatik.uni-mannheim.de

Short Course Description
This course investigates the foundations of Natural Language Processing (NLP) as tool for analyzing natural language texts in the social sciences, thus providing an alternative to traditional ways of data generation through surveys. The course introduces general use cases for NLP, provides a guide to standard operations on text as well as their implementation in the Python-based Natural Language Toolkit (NLTK) and introduces the text mining functionalities of the WEKA Machine Learning workbench.

The theory part of the course worth one credit can be supplemented by an optional project part worth another credit point.

Course and Learning Objectives
By the end of the course, students will be able to ...:
- Understand the possibilities and limitations of automatic text analysis
- Judge the potential benefits of applying automatic text analysis to a given research question
- Preprocess a corpus using the Natural Language Toolkit (NLTK)
- Perform text classification using the WEKA Machine Learning workbench
- Understand the principles of advanced text analysis methods

Prerequisites
Participants need to have attended the following IPSDS courses or have corresponding knowledge:

1. SURV699C Introduction to Python and SQL or necessary knowledge in programming in Python: data types & structures, functions & loops, file I/O
2. SURV736 Web Scraping (recommended)
Class Structure and Course Concept
This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit and then “attending” mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. Just like in an on-site course, homework will be assigned and graded and there will be a final project at the end of the course.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings
Thursday, 7 am ET/13:00 CET
Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Wednesday, 7 am ET/13:00 CET).

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

Grading
Grading will be based on:
- Participation in online meetings (10%)
• Answering questions about the content of the videos – 4 quizzes (15%)
• Practical application of NLP and Machine Learning technologies – 4 assignments (75%)

Dates of when assignments will be due are indicated in the syllabus. Extensions will be granted sparingly and are at the instructor's discretion.

Technical Equipment Needs
The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description
In this course, participants should learn how to use standard methods of Natural Language Processing (NLP) to support social science research through automatic content analysis. For this purpose, the course starts with an introduction of typical use cases for NLP such as information extraction, text classification and topic detection. The participants will acquire a basic understanding of the mature and possible applications of these methods to be able to judge to what kinds of problems they can be applied. Further, participants will acquire practical knowledge of how to implement these methods using the Python library Natural Language Toolkit (NLTK) and the text mining features of the WEKA Machine Learning workbench. We will look at how to generate the specific feature format WEKA needs as input from textual resources and guide the participants through the use of WEKA for performing systematic text classification experiments. Beyond this basic form of text analysis, we will also look at two advanced techniques that go beyond the classification of a text. In particular, we will look at so-called topic models that generate topics that can be identified in a set of documents in terms of a probabilistic assignment of words to the different topics and we will introduce the idea of training word embedding models to capture word-level semantics.
Readings


RegexOne Interactive Tutorial  
http://regexone.com/lesson

http://www.nltk.org/book


http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.1205&rep=rep1&type=pdf#page=96

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and


Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.
Unit 1: Introduction – Potential and Use Cases of Text Analysis Methods
Online meeting (Heiner Stuckenschmidt/Christoph Kilian Theil): Thursday, November 29, 7 am ET/13:00 CET

Online quiz 1 & Assignment 1: both due Friday, November 30, 7 am ET/13:00 CET

Video lecture (Christoph Kilian Theil): available online Thursday, November 22

Readings:
[Grimmer and Steward, 2013]
[O’Connor et al., 2011]

Unit 2: The Natural Language Toolkit
Online meeting (Heiner Stuckenschmidt/Christoph Kilian Theil): Thursday, December 6, 7 am ET/13:00 CET

Online quiz 2 & Assignment 2: both due Friday, December 7, 7 am ET/13:00 CET

Video lecture (Christoph Kilian Theil): available online Thursday, November 29

Readings:
RegexOne Interactive Tutorial http://regexone.com/lesson
[Bird et al., 2014] chapters 1.1–1.4, 3.1–3.10

Unit 3: Text Mining with WEKA
Online meeting (Heiner Stuckenschmidt/Christoph Kilian Theil): Thursday, December 13, 7 am ET/13:00 CET

Online quiz 3 & Assignment 3: both due Friday, December 14, 7 am ET/13:00 CET

Video lecture (Christoph Kilian Theil): available online Thursday, December 6,

Readings:
[Witten et al., 2017] chapters 4.1–4.3, 4.6–4.7, 6.3

Unit 4: Advanced Text Mining
Online meeting (Heiner Stuckenschmidt/Christoph Kilian Theil): Thursday, December 20, 7 am ET/13:00 CET
Online quiz 4 & Assignment 4: both due Friday, December 21, 7 am ET/13:00 CET

Video lecture (Christoph Kilian Theil): available online Thursday, December 13

Readings:
[Blei and Lafferty, 2009]
[Schmidt, 2015]

**Note:** Student access to the course website will be revoked four weeks after the last online meeting.