COMPUTATIONAL LINGUISTICS

Detailed Course syllabus

Teacher

Christophe Coupé

Department of Linguistics, The University of Hong Kong

Email: ccoupe@hku.hk

Course date and time

June 24, 2019 – July 5, 2019

Monday to Friday, 9:30 am - 1:30 pm

No lecture on July 1, Monday, 2019 (HKSAR Establishment Day)

Total number of contact hours: 36 hours

The teacher will be available during the hours of the contact period for further discussion or help with technical aspects.

After July 5, the teacher will be further available through emails.

Short description of the course

Natural Language Processing (NLP) addresses how machines analyze, produce and understand natural language. The field lies at the crossroads of linguistics and computer science and encompasses a wide range of techniques including automatic speech recognition, text mining, machine translation, and speech synthesis. The goal of this course is to provide a thorough introduction to basic methods and issues of NLP, with a specific focus on text mining. No previous experience or prerequisites in programming are required for this course. During hands-on practice, you will experiment with various computational techniques, especially by applying and adapting generic scripts provided by the teacher (the R software environment will be used to this end).

Keywords

Natural language processing (NLP); auto-correction & spell checking; automatic speech recognition; morphological and semantic processing; speech synthesis; text mining; sentiment analysis; machine translation

Intended Learning Outcomes

At the end of the course, you will be able to:

1. Describe the processes, achievements and challenges of several fields of NLP (speech recognition, morphosyntactic analysis, machine translation, text mining, speech synthesis)
2. Apply basic NLP algorithms to the corresponding inputs
3. Apply text mining tools to a text corpus using the R software environment
4. Reflect upon the ethical challenges created by NLP

Overall, the course will try to answer both the "what?" and "how?" questions when it comes to processing natural language with computers. In other words, we will not focus only on why NLP is interesting and what scholars and engineers are doing in this field, but also study some concrete techniques and acquire basic skills through hands-on experience.

Course structure / content

Classes will consist in lectures, learning activities and hands-on practices.

A few software (e.g. Praat, speech synthesizers) will be used during the learning activities. The R programming environment and R Studio will also regularly be used to experiment with different computational techniques. Chunks of code (snippets) will be provided by the teacher, and students will learn by modifying them and applying them to different raw data. In connection with one of the assignments, you will in particular apply text mining techniques to collections of news articles.

After an introduction to the course and to NLP in general, we will start with the issue of correcting spelling errors (just think of the number of errors made each day when sending instant messages on various social platforms). The, during the rest of the first week, we will consider the chain of treatments from recognizing speech to analyzing its content to synthesizing speech. We will thus address questions such as how do we extract vowels and consonants from an audio signal? How do we identify words? How do we extract meaning and solve ambiguities? Etc.

During the second week, we will focus on text mining, a field of NLP which has grown exponentially with the explosion of online text contents (social media, online newspapers, blogs etc.) We will study questions such as how to extract topics from collections of texts? How to analyze sentiments or lexical richness? Can we characterize one's writing style? Etc.

Learning activities

During the learning activities, we will among others:

- Brainstorm about talking robots
- Try to distinguish real human poets from computers
- Create sentences ambiguous to a computer

- Create audio messages with various speech synthesizers
- Study novels (such as Frankenstein) or news articles with text-mining tools

Details of the lectures

Date	Lecture
June 24	**Introduction to the course & Introduction to Natural Language Processing**
June 25	**Auto-correction, spell checkers and beyond**
June 26	**Automatic Speech recognition (ASR)**
June 27	**Processing morphology and semantics**
June 28	**Speech synthesis**
July 2	**Text mining/analytics (Part I)**
July 3	**Text mining/analytics (Part II)**
July 4	**Machine translation: from Warren Weaver to *Google Translate***
July 5	**Hands-on practice and wrapping up**

Assessment

The course is assessed on the basis of course work - there is no examination.

There will be 2 components for the assessment of the course:

- A written report on your investigations of a novel or of news articles with text mining techniques
- An audio file with a synthetic voice describing a real-life application of an NLP technique

The first report will amount to 70% of the final grade, the audio file to 30%.

Written report on studying a novel or news articles with text mining techniques

During the hands-on practice of the contact hours (second week), you will discover how to apply various text mining techniques to a novel or to a collection of news articles. These techniques can help someone to get a better understanding of textual documents from a quantitative rather than qualitative point of view. Upon choosing a set of documents (either among some predefined options, or one you create

yourself), you will generate a number of quantitative analyses from them, experimenting with various parameters to get different outputs and perspectives. Your job will be to prepare an analysis of these different outputs, and thus try to answer the question: what do I learn from these documents when using these techniques? (e.g. what are the texts about? Are they rather sad, angry, joyful? What are the different topics, and how do they connect with the previous emotions? How is the style, and how rich is it? Are there significant differences between the texts?)

Target number of pages for the report: 10 to 12 pages (figures included, standard margins, Time New Roman 11 pt, single spacing)

Audio file with a synthetic voice describing a real-life application of an NLP technique

You will have to prepare a short report on a real-life application of one of the NLP techniques seen during the lectures (e.g. text mining in social networks for early detection of bad buzz, reliance on automatic speaker recognition in judicial trials etc.). More precisely, you will have to describe the issue at hand, the techniques used, the current successes and challenges, as well as the ethical issues. You will be given a set of predefined topic and some resources, or you'll have the option to focus on your own topic of interest.

The report will take the form of an audio file with a synthetic voice presenting the report. Your grade will primarily be about informative your report is, but also how attractive, convincing and interesting it is.

Target duration of the audio file: 3 to 5 minutes

Relationship of the assessments to the intended learning outcomes (ILO)

Written report on studying a novel or news articles with text mining techniques:

- ILO 2: Apply basic NLP algorithms to the corresponding inputs
- ILO 3: Apply text mining tools to a text corpus using the R software environment

Audio file with a synthetic voice describing a real-life application of an NLP technique:

- ILO 1: Describe the achievements and challenges of several fields of NLP (speech recognition, morphosyntactic analysis, machine translation, text mining, speech synthesis)
- ILO 2: Apply basic NLP algorithms to the corresponding inputs
- ILO 4: Reflect upon the ethical challenges created by NLP

Grading

For the written report on studying a novel with text mining techniques, the document will be evaluated as follows:

A. Content (15 points)

    a. Clarity: Is the content easy to understand? (3 points)
    b. Structure: Is the argumentation well structured? (2 points)
    c. Quality: Are the arguments/ideas convincing/correct? (4 points)
    d. Quantity: Is there sufficient information? (3 points)
    e. Personal reflection: Are there original ideas? (3 points)
B. Surface (6.5 points)
    a. Figures: Are there figures and are they easy to understand? (3 points)
    b. Formatting: Is the document well formatted and presented overall? (2 points)
    c. Language: Are the orthography and grammar good enough? (1.5 points)

The grade over 21.5 points will be divided by 5 to obtain a grade over 4.3 points, in order to follow the current grading system at HKU.

For the description of a real-life application of an NLP technique, the audio file will be evaluated as follows:

A. Content (15 points)
    a. Clarity: Is the content easy to understand? (3 points)
    b. Structure: Is the argumentation well structured? (2 points)
    c. Quality: Are the arguments/ideas convincing/correct? (4 points)
    d. Quantity: Is there sufficient information? (3 points)
    e. Personal reflection: Are there original ideas? (3 points)
B. Surface (6.5 points)
    a. Voice: How good is the voice / How much attention has been paid to the synthesis? (2 points)
    b. Attractiveness: Is the audio message catching one's attention? Is it lively/fun/entertaining? (3 points)
    c. Language: Is the English good enough? (1.5 points)

The grade over 21.5 points will be divided by 5 to obtain a grade over 4.3 points, in order to follow the current grading system at HKU.

The grades of both components of the assessment with be average with the aforementioned weights (70% and 30%). Your final grade will thus be between 0 and 4.3.

Deadline for handing in the assignments

Both assignments will be due on **July 28 at midnight** (the night from July 28 to July 29). An electronic version will be uploaded on an online platform or sent to the teacher.

Study load

The course will require approximately 120 hours of work, divided as follows:

- Contact hours: 36 hours
    - Lectures: around 24 hours
    - Learning activities and hands-on practice: around 12 hours
- Assessment: 20 hours
- Self-study: 65 hours

Recommended reading

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing (2nd Edition). Upper Saddle River, NJ, USA: Prentice-Hall. (for the 3rd edition, check https://web.stanford.edu/~jurafsky/slp3/)

Robinson, D., & Silge, J. (2017). Text Mining with R. O'Reilly Media.

Annex: Current grading system at HKU

Grade	Standard	Grade Point	General Expectations of Student Performance
A+	Excellent	4.3	**Excellent result.** A thorough grasp of the subject as demonstrated by original, creative or exceptionally astute analysis and synthesis of ideas or critical interpretation of texts/issues/other course content or reflection on learning experience. Ample evidence of familiarity with relevant reading and research as well as very effective organizational, rhetorical and presentation skills as appropriate to the assessment task. Students display excellent knowledge and performance in areas such as grammar, vocabulary, and oral and aural competencies.
A		4.0	
A-		3.7	
B+	Good	3.3	**Good to very good result**. A good to very good grasp of the subject as demonstrated by generally persuasive analysis and synthesis of ideas or critical interpretation of texts/issues/other course content or reflection on learning experience. Some evidence of a generally sound understanding of relevant reading and research as well as effective organizational, rhetorical and presentation skills as appropriate to the assessment task. Students display good to very good knowledge and performance in areas such as grammar, vocabulary, and oral and aural competencies.
B		3.0	
B-		2.7	
C+	Satisfactory	2.3	**Satisfactory to reasonably good result**. A reasonable grasp of the subject as demonstrated by some analysis of ideas or interpretation of texts/issues/other course content or reflection on learning experience. Familiarity with relevant reading and research is adequate but tends to be rather descriptive with little evidence of critical reflection but organizational, rhetorical and presentation skills, as appropriate to the assessment task, still contribute to overall coherence satisfactorily. Students display reasonable knowledge and performance in areas such as grammar, vocabulary, and oral and aural competencies.
C		2.0	
C-		1.7	
D+	Pass	1.3	**Barely satisfactory result**. A minimal grasp of the subject with little analysis of ideas or critical interpretation of texts/issues/other course content or reflection on learning experience. Hardly any evidence of familiarity with relevant reading or research as required for the assessment task. Ideas presented are generally not well organized or well argued but still largely comprehensible. Students display minimal knowledge and performance in areas such as grammar, vocabulary, and oral and aural competencies.
D		1.0	
Fail	Fail	0	**Unsatisfactory result**. A poor grasp of the subject with negligible or largely inaccurate analysis of ideas or interpretation of texts/issues/other course content or reflection on learning experience. A general lack of familiarity with relevant reading or research, as required for the assessment task. Work presented is poorly organized, largely irrelevant and incoherent. Students display poor knowledge and performance in areas such as grammar, vocabulary, and oral and aural competencies. Plagiarism or non-submission of coursework will also result in a Fail.