Lessons from a Massive Open Online Course (MOOC) on Natural Language Processing for Digital Humanities

Simon Clematide, Isabel Meraner, Noah Bubenhofer, Martin Volk

Institute of Computational Linguistics
University of Zurich, Switzerland

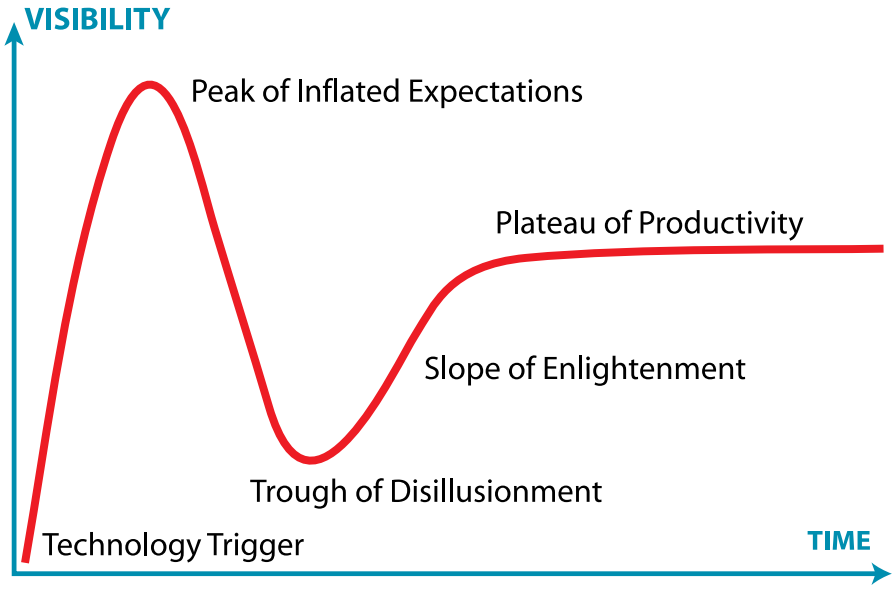
September 12, 2017

Teach4DH Workshop @ GSCL 2017 Berlin

Massive Open Online Courses (MOOCs)

Hype Cycle: Have MOOCs reached the plateau of productivity?

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run." (Roy Amara)



- MOOC $\approx$ Mainly video-based distance learning for higher education
- Worldwide, around 60 million people have signed up for MOOCs [Ubell, 2017]
- Commercial (like Coursera) and nonprofit (like edX) platforms compete for (paying) students for their open courses

Digital Scholarship and Automatic Text Analysis

More and more scientific disciplines use automatic text analysis

- humanities: corpus linguistics, quantitative cultural studies ("distant reading"), corpus-based discourse analysis, ...
- computational social science: media monitoring
- bio-medical text mining, ...

But ...

applying NLP methods to texts requires special knowledge and skills

Our Introductory MOOC on NLP for Digital Humanities

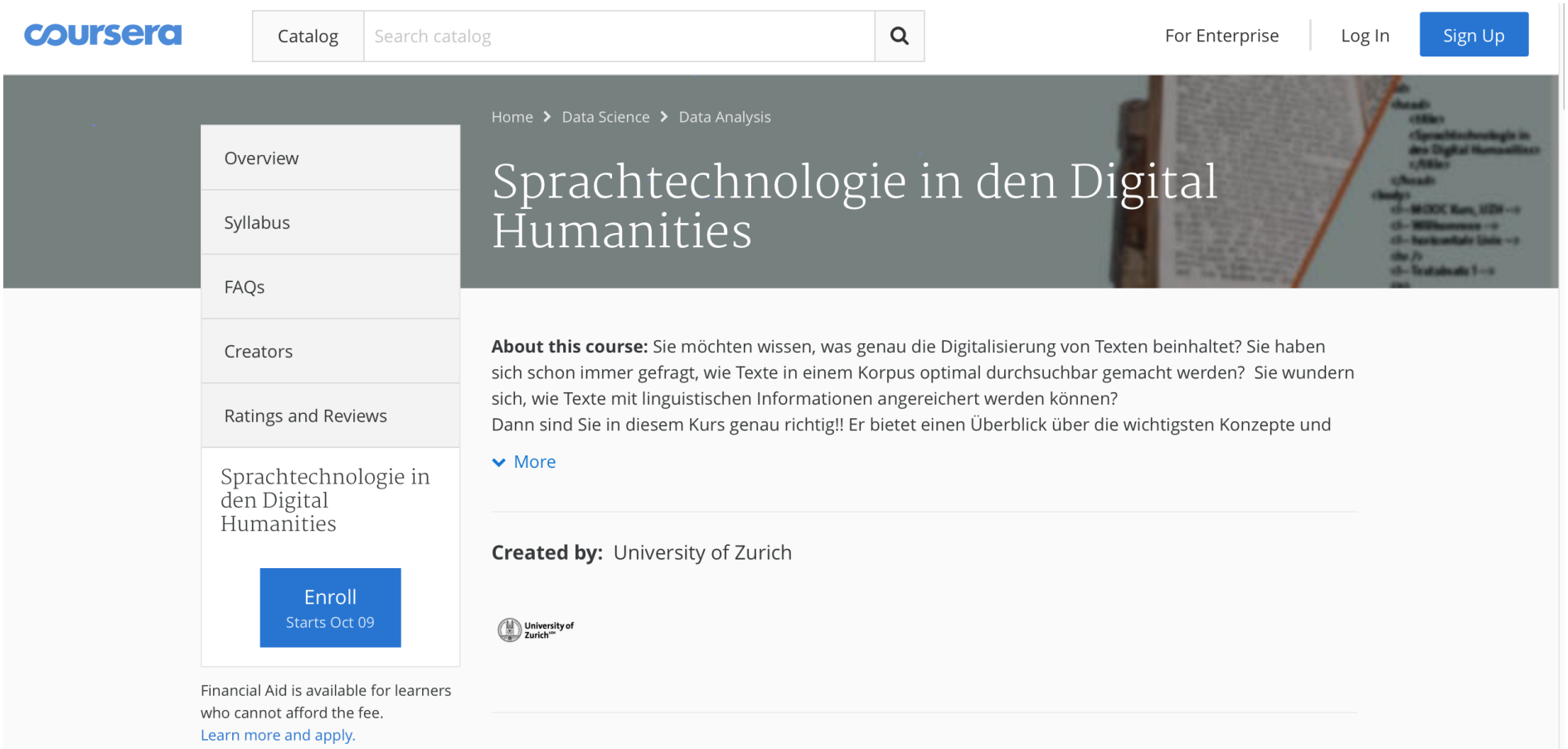
. . . does not teach any NLP programming skills.

Our main goal is

- a broad and illustrative overview on important concepts, problems and techniques
- for automatically enriching and exploiting text corpora
- via visual exploration, and allowing for sophisticated corpus queries.

Thereby introducing

- the process of digitization, corpus creation, text representation, statistical analysis, visualization,
- automatic and manual annotation on different linguistic levels (including their quantitative evaluation)
- as well as the challenges and benefits of multilingual document collections.



An open course on Coursera provided by the University of Zurich and held in German

Some Hard Facts

- 6 weekly modules: $\approx$ 2-3 study hours per week for students
- 3 initially inexperienced video lecturers: Dr. Simon Clematide, Dr. Noah Bubenhofer, Prof. Dr. Martin Volk
- 2 student tutors: Sara Wick (initial course implementation, video production) for the 2015 session; Isabel Meraner (subtitling, course migration on new Coursera platform) for the 2017 sessions
- 1 (small) course production budget: 25,000 CHF (plus a 5% part-time student tutor (forum support and integration of small adjustments from user feedback) while the course is running)
- A lot of good and free technical support from "Digitale Forschung und Lehre" and the multimedia production services of the University of Zurich
- 46 certificates of accomplishments in 2015 (out of 883 learners that actively visited the course at least once)
  $\rightarrow$ yes,..., typically, only 5 to 12% of all registered course users successfully complete a course [Ubell, 2017].

Why on Earth in German?

- Good question. . . most MOOCs are held in English, the global language of science and business
- Less participants (although some learners are motivated by their "hidden agenda" of learning a foreign language)
- Focus on multilingual diachronic text corpora (our running example is the Text+Berg corpus of yearbooks of the Swiss Alpine Club (1864-2015))
- Occupying a niche for working on German texts
- For an introductory level, a course in mother tongue might still be beneficial (and the videos are easily reusable for our Bachelor program students)
- Coursera has/had some interest in promoting non-English courses
- Subtitles can be translated (but less so the illustrative text material)
- Forum activity probably suffers (but we explicitly allow for English or German posts)

Content and Course Design

- 3 lecturers agreed an the overall structure, content and presentation style
- Each lecturer was responsible for fine-tuning his own modules (slides, background material, tools, demos)
- Each lecturer was presenting his favorite topics
- Each lecturer had experience in teaching these topics
- Each lecturer needed a lot more time than expected for fitting his learning material into video episodes of a reasonable length for online learning (and they are still too long according to current standards)

Module 1: "Paths into the Digital World" (Volk)

- Digitization: OCR (and OCR post-correction/crowd-correction), OLR, acquisition of text corpus material, including digital-born documents and the challenges one encounters with them
- Explained and illustrated by the digitization project Text+Berg
- Short interviews about the relevancy of digitization and practical large-scale digitization techniques with two experts from the (digitization center of the) Zurich central library

Module 2: "Structured and Sustainable Representation of Corpus Data" (Clematide)

Character and structured text representation

- Character encoding (ASCII and Unicode), textual storage formats (UTF-8)
- XML Markup language and the TEI P5 standard for structured text representation

Automatic sentence and word segmentation

- Tokenization
- Dealing with punctuation and abbreviations:
  $\rightarrow$ Exemplary discussion of rule-based, supervised, and unsupervised approaches

Module 3: "Properties of Corpora and Basic Methods for Analysis" (Bubenhofer)

Statistical properties of text corpora

- Term frequencies, n-grams, collocations
- Corpus query languages and tools (hands-on)

Visualization and exploitation

- "Visual linguistics" [Bubenhofer, 2016]: Tools for displaying interesting text properties in a creative, interactive and illustrative way
- Exploratory "distant-reading-like" investigations of corpora

Module 4: "Automatic Corpus Annotation Using NLP Tools" (Clematide)

- Lexical and syntactic corpus annotation methods: part-of-speech tagging, stemming, lemmatization, chunking, parsing
- Shallow semantic processing: Named Entity Recognition (mention detection and coarse-grained entity classification) and Entity Linking

Module 5: "Manual Annotation and Evaluation of Corpus Data" (Clematide)

- Efficient combination of manual and automatic annotation (along the paradigm of "Manual Annotation for Machine Learning" [Pustejovsky and Stubbs, 2013]
- Their MATTER annotation process model
- Relevant evaluation metrics (precision, recall, f-measure) for quantifying the quality of NLP applications
- Inter-rater reliability for assessing the quality/inter-subjectivity of manual annotations

Crowdsourcing Manual Annotation

- Introduction of typical crowdsourcing paradigms: gamification, paid microwork, citizen science (volunteer work)
- Expert truth vs. crowd truth

Module 6: "Challenges in Multilingual Text Analysis" (Volk)

- Automatic language identification in large-scale multilingual text collections
- Tools for automatic alignment of documents, sentences, and words of parallel corpora

Initiatives, Resources, and Tools Mentioned

Many things are mentioned

(a) digitization initiatives (Projekt Gutenberg, Europeana, TextGrid); (b) OCR crowd-correction and crowd-sourcing in general (TypeWright, Crowdflower, Artigo); (c) online corpora and corpus query tools (COSMAS II/DeReKo, DWDS, CQPweb); (d) parallel corpora (EuroParl, Canadian Hansard); (e) sentence and word alignment tools for parallel corpora (InterText, HunAlign, GIZA++); (f) language identification (lingua-ident, LangId); (g) text representation standards (Unicode, UTF-8, XML, TEI-P5); (h) annotation standards (STTS, Universal tags and dependencies); (i) standard lexical and syntactic NLP tools (Porter Stemmer, Durm Lemmatizer, TreeTagger, Connexor-Tagger; chunkers and parsers); (j) named entity recognition (Open Calais, Stanford NER); (k) tools for manual annotation of linguistic structures (and/or querying the annotations) (WebAnno, ANNIS, EXMARaLDA, RSTTool); (l) visualization (Graphviz, Leaflet, Gephi).

Assessments and Active Learning

- Traditional multiple-choice quizzes at the end of each module
- In-video quizzes and reflection questions for re-captivation of the learner's attention

Peer Assessments: hands-on and critical thinking

- Each student solves an open task according to well-defined criteria
- Each student assesses the quality of the solutions of other students w.r.t. these criteria
- PA1 in Module 3: Find an interesting diachronic corpus query, look at its visualization and interpret the result
- PA2 in Module 5: Perform NER with a standard tool (Stanford NER tagger/ Open Calais) and evaluate its precision and recall

Active learning is more demanding for the students $\rightarrow$ rather high dropout rate on these (obligatory) tasks in our course

Community

- Distant learning has more to offer then just streamed video recordings.
- Discussion forums can replace some of the missing in-class communication of classroom teaching.
- However, there was not that much discussion between the participants in our rather technical course
- Exceptions: difficult unexplained concepts (e.g. using the term dependency parsing before properly explaining it in a later module)
- Unclear cases: the NER evaluation assessment raised questions whether the expression “Mittelmeerraum” (Mediterranean) should be recognized as a toponym or not.
- Observation: imperfections, omissions, uncertainties awaken the community.
- Perfection puts it to sleep.

Motivation

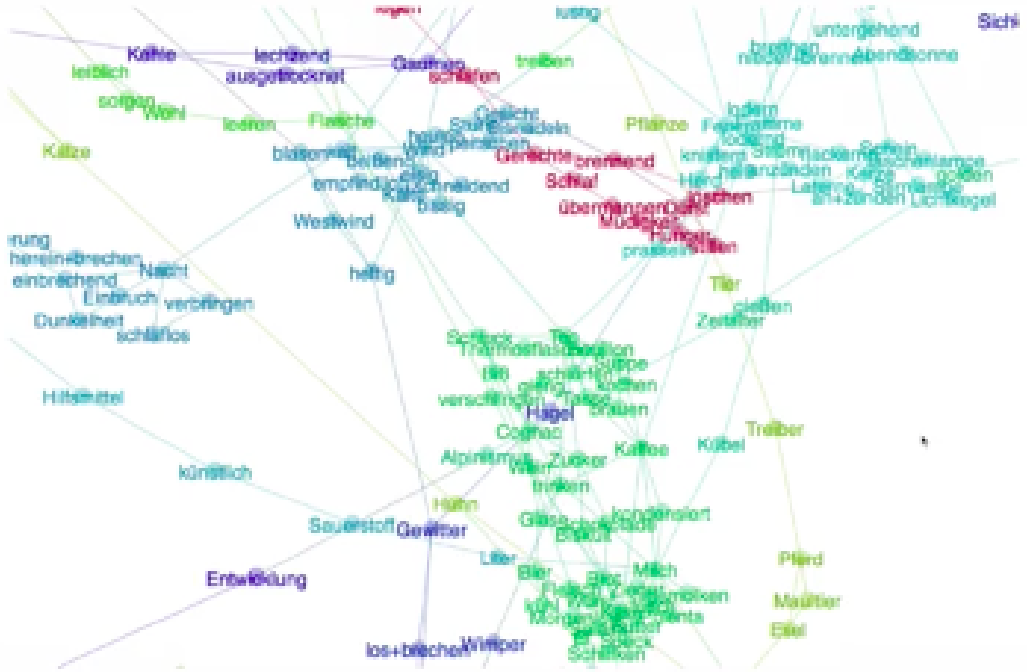


Abbildung: Noah Bubenhofer: Die typische Bergtour: www.bubenhofer.com/sprechtakel/

03:23 / 32:33

3.3 - Visualisierung und graphische Darstellung

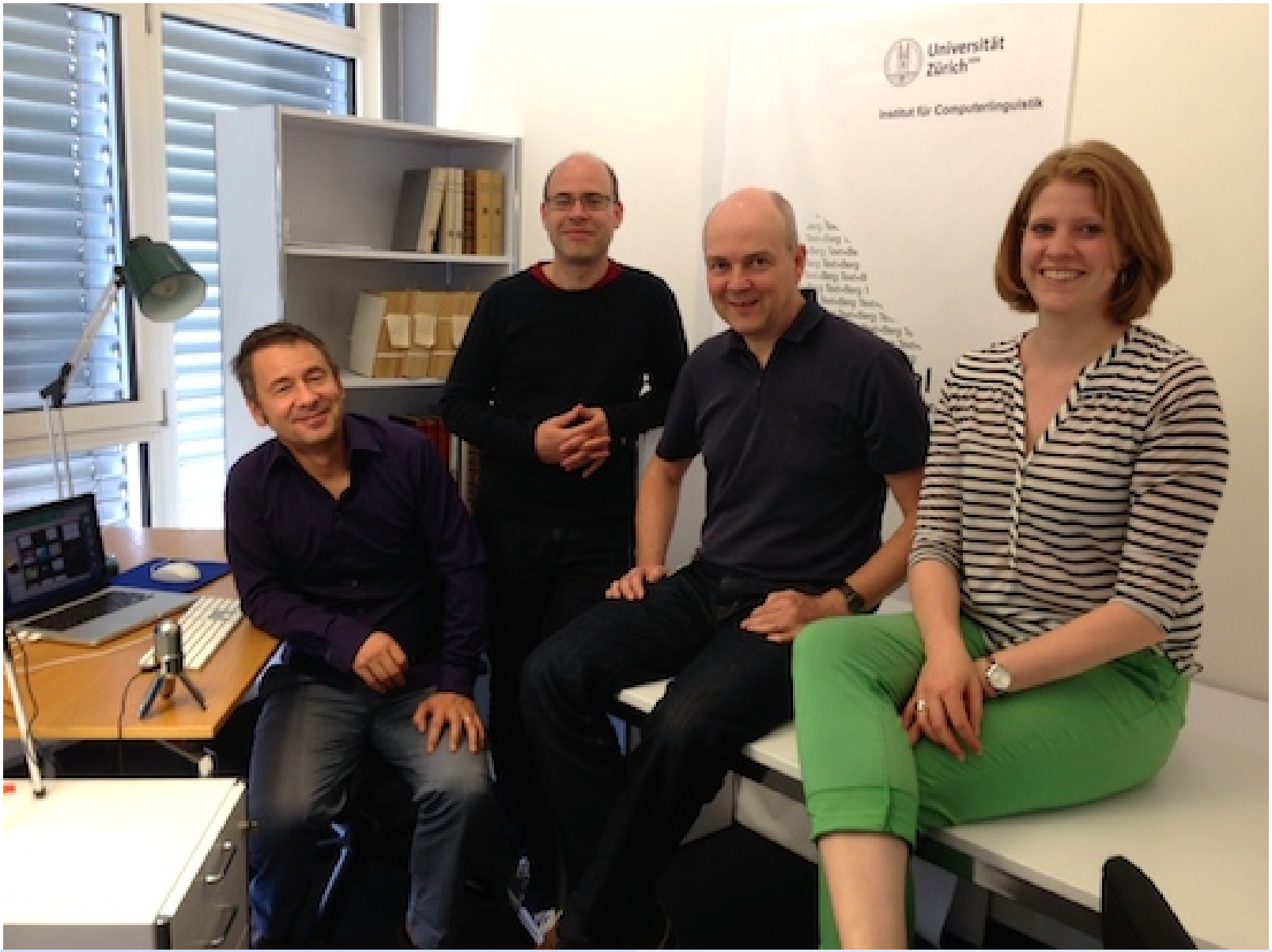
Production Experience

- Self-made video recordings in an office turned into a makeshift studio
  → gave us some flexibility and relaxedness
- Professional help (lightning, camera position, "talking to the camera, and not to the slides") in the beginning for the setup
  → a good microphone is important, however, our new one turned out to be defective
- Classical, "unambitious" talking head with slides: during video cut, some visual effects (zooming, highlighting, annotations) were added for clarity and for avoiding monotony

– Publication on Coursera's platform requires a lot of point, select and click
  → no support for course exchange formats (e.g. SCORM),

\+ Coursera offers good support (course design) and infrastructure for course authors

Happy Faces at the End of the Production Phase



NLP: A Rapidly Evolving Discipline

Paradigm Changes in the Last 25 Years

1. Handwritten rules and application-specific algorithms
   → linguistic structures are key
2. Statistical systems using supervised machine learning with annotated training material
   → feature engineering is key
3. Deep and/or recurrent neural networks with end-to-end architectures without interpretable intermediary representations (goal: "from characters directly to application-specific output")
   → general architectures and numeric optimization are key

Our course reflects the stages 1 and 2 and their different requirements (e.g. annotated training material),
. . . and so far ignores the "deep learning tsunami" [Manning, 2015] that hit the NLP area.

Classical White Pipelines vs Black Boxes

Our Course: Classical NLP Pipeline Architecture

- Language identification, tokenization, POS tagging, lemmatization, NER, syntactic analysis
- Better suited for students with a typical DH background in arts and humanities: the problems and challenges of automatic text analysis have an interpretable form in this paradigm.

Neural Black Boxes

- High performance on the task, but difficult to interpret
- Tricky question: Should we advocate the performance-oriented use of "magic" tools?

Still, an intermediate NLP course has to cover distributional (word embeddings, topic modeling) and neural approaches. This requires more mathematical and programming skills.

Summary

- Presentation of the conception and realization of an on-going open video-based introductory course on classical NLP techniques held in German on Coursera
- Some reflections on the "right kind" of NLP for DH
- Maybe some stimulus for discussion...

The End

Thank you for your attention.

Comments? Questions?

Please visit
`https://www.coursera.org/learn/digital-humanities`
Next cohort starts in October

Acknowledgments

- "Digitale Lehre und Forschung (DLF)" from the Faculty of Arts of the University of Zurich (UZH), especially *Anita Holdener* (DLF) for her technical support.
- "Multimedia & E-Learning-Services (MELS)" of the UZH, especially *Lukas Meyer*
- *Sara Wick*, our initiative student tutor and production assistant in 2015

Discussion Topics of (my) Interest

- Which topics does our course miss?
- Which programming skill are necessary for DH?
- Which frameworks, tooling, programming languages build a solid and reasonable basis in higher education?
- What is the difference between a Digital Humanist and an NLP specialist /text miner?

Bibliography

Bubenhofer, N. (2016). Drei Thesen zu Visualisierungspraktiken in den Digital Humanities. *Rechtsgeschichte Legal History - Journal of the Max Planck Institute for European Legal History*, (24):351–355.

Manning, D. C. (2015). Last words: Computational linguistics and deep learning. *Volume 41, Issue 4 - December 2015*, pages 701–707.

Pustejovsky, J. and Stubbs, A. (2013). *Natural language annotation for machine learning*. O'Reilly Media, Sebastopol, CA.

Ubell, R. (2017). Moocs come back to earth. *IEEE Spectrum*, 54(3):22–22.