Digital Humanities

TAU 2016-2017

Nachum Dershowitz, Kfir Bar
Digital Humanities

• The intersection of digital technology and humanities disciplines

• The humanities is the discipline that focuses on the arts, literature, music, dance, theater, architecture, philosophy, and other expressions of human culture
General process

Raw data (analog)

Converting analog to digital (e.g., Scanning + OCR, Recording + Voice-to-text)

Digital data

Analysis (e.g., machine learning, NLP, Computer vision)

insights

Visualization

Graphic interface
Related CS topics

• Machine learning
• Vision and sound analysis
• NLP
• Graph analysis
Digital Humanities (DH) conference

• Some of 2016’s topics:
  • 3d printing
  • agent modeling and simulation
  • anthropology
  • archaeology
  • archives, repositories, sustainability and preservation
  • art history
  • asian studies
  • audio, video, multimedia
  • authorship attribution / authority
  • bibliographic methods / textual studies

http://dh2016.adho.org/abstracts/browse/topics/
Digital Humanities (DH) conference - continue

• classical studies
• concording and indexing
• content analysis
• philosophy
• image processing
• film and cinema studies
• text analysis
• theology
• visualisation

http://dh2016.adho.org/abstracts/browse/topics/
Project examples
Digitization example: **Mapping Republic of Letters**

http://republicofletters.stanford.edu/
Digitization example: London Lives

- A fully digitized and searchable archive of a wide range of primary sources about eighteenth-century London

- Documents were manually transcribed twice by two different persons

- Named entities were marked up using dictionary and NLP based methods
On the Origin of Species (Darwin)

- Published first on 1859, and revised 5 times.
- https://fathom.info/traces/
Google Books $n$-gram viewer

- Ngram viewer
  [https://books.google.com/ngrams](https://books.google.com/ngrams)

- Project published @Science
  Quantitative Analysis of Culture Using Millions of Digitized Books
The Friedberg Genizah project

https://fgp.genizah.org
THE BERKELEY FOLK MUSIC FESTIVAL PROJECT

Principal Investigator
Michael J. Kramer
History and American Studies
Northwestern University
email: mj@northwestern.edu
phone: 847 942 5882
website: www.michaeljkramer.net

Overview

http://www.michaeljkramer.net/cr/the-berkeley-folk-music-festival-project/
General concepts
Structure vs. Unstructured data

- Structured data uses extra elements (such as labels), data structures or other means to add an extra level of interpretation to the data.

- Unstructured data refers to texts, images, sound files, etc. that has not had a known format.
Mark-up language

• A way to structure the data
• Adding information, which helps analyzing the data (title, chapter, gender, proper noun, etc.)
  • Widely used by libraries to represent digital resources
In April 1957, the Museum of New Mexico in Santa Fe honored 84-year-old Olive Rush with a retrospective of her work from the past forty years. The Hoosier-born painter, considered “one of the city’s most lovable personalities” and “greatest artists” had “brought fame, dignity, and beauty” to the vibrant artist center that she helped to develop in the
Data classification

• A higher order of organization

• Organizing texts, physical objects, files, images, recordings etc. into classes, or labels, which form a higher level structure

• Examples:
  • Thompson. Motif-index of folk-literature (http://www.ruthenia.ru/folklore/thompson)
Data classification

• Manual vs. Computational approaches for classification

• Computational approaches
  • Rule based
  • Data driven (e.g., using Bayes)

\[
P(C | X) = \frac{P(X | C)P(C)}{P(X)}
\]

\[
Posterior = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}
\]
Ontologies and taxonomies

• In general: a naming system, lists of terms/names, organized in a hierarchical structure

• Used to represent a domain of knowledge in a standard way

• No real difference between taxonomy and ontology. If at all, one may say that a taxonomy is a tree and ontology is a graph (for more read [http://www.ideaeng.com/taxonomies-ontologies-0602](http://www.ideaeng.com/taxonomies-ontologies-0602))

• Examples: Wikipedia, Wordnet
Example: WordNet

**synonyms:** sofa = couch

**antonyms:** good/bad, life/death, come/go

**hyponyms / hypernyms:**
class inclusion: cat < mammal < animal

**meronyms:** the part-whole relation: line < stanza < poem
Example: Traffic ontology

https://www.researchgate.net/figure/220974898_fig2_Fig-3-Example-ontology-in-Traffic-Services-domain-defining-concepts-like-Sensor-and
Automatic data tagging/labeling

• Extracting ontology-based tags/labels from unstructured data using rules/data driven approaches
  • Computer vision
  • Text analytics (natural language processing)
Visualization

• Showing features/patterns in a condensed form

• A dataset can be visualized in different formats
Graph analysis

• A system of elements or entities that are connected by explicit relations

• good examples of networks are social networks, traffic networks, communication networks

• Snap (http://snap.stanford.edu/)

• RDF/XML - standard for representing a graph using triplets (subj, predicate, obj)
Natural language processing

• Relevant applications:
  • Document classification (e.g., Tibetan original vs. translated source)
  • Summarization (e.g., Yahoo Summly)
  • Machine translation (e.g., Google translate)
  • Chat bots (e.g., Alexa, Siri)
  • Search (e.g., Morphological/semantic search)
  • Question answering (e.g., IBM Watson)
NLP layers

• **Phonetics/phonology:** how words are actually sound?
• **Morphology:** what words (or sub-words) are we dealing with?
• **Syntax:** what phrases are we dealing with? Which words modify one another?
• **Semantics:** what’s the literal meaning?
• **Pragmatics:** what should you conclude from the fact that I said something? How should you react?
Eliza (Weizenbaum, 1966)

- Remarkably simple “Psychologist”
- Uses pattern patching to carry on limited form of conversation
- Seemed to pass the Turing Test! (Machines Who Think, McCorduck, 1979)
- Demos:

“A computer would deserve to be called intelligent if it could deceive a human into believing that it was human”
Ambiguity

I made her duck
Ambiguity

I made her duck ??

- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (ceramic?) duck she owns
- I caused her to quickly lower her upper body
- I waved my magic wand and turned her into undifferentiated waterfowl
NLP-based labeling tools

• Part of speech tagger

I/pro-noun would/aux like/verb to/prep buy/verb a/det ticket/noun to/prep New/proper-noun York/proper-noun

• Named entity recognizer

Yaron/B-person London/I-person traveled/O to/O London/B-location
Machine learning

• Supervised – when you have labeled data and you want to automatically learn how to label unseen instances

• Unsupervised – when you have unlabeled corpus, and you want to find an interesting structure in the corpus

• Reinforcement learning – predicting actions in an environment so as to maximize some notion of cumulative reward
A generic supervised learning approach

**Training**
- Labeled data
  - Feature extraction
  - Vector representation
  - Learning
  - Classifier (model)

**Predicting**
- Single instance
  - Feature extraction
  - Vector representation
  - Predicting
  - Label
  - Classifier
Learning algorithms

• Generative models– given an instance, predicting the distribution of the labels (e.g., Naïve Bayes, GMM, LDA, HMM)

• Discriminative models – finding the dependency between a label and the input data (e.g., SVM, Logistic regression, NN, CRF)
Relevant links

- http://dh101.humanities.ucla.edu/
- http://4humanities.org/
- https://amicus.uvt.nl/
Seminar requirements

• Presenting a paper in class (1 hour presentation + discussion)
• Attending all classes
• Taking part in class discussions (reading the papers before class)
• 500-1000 word summary of what learned in the seminar (to be submitted on the last meeting)