Notes on Digital Humanities

02.137DH Introduction to Digital Humanities, Term 4 2019

Wei Min Cher

5 October 2019

Contents

1 W1: Introduction 4
- 1.1 What are the Digital Humanities? 4
- 1.2 Benefits of digital tools in the Humanities 4
  - 1.2.1 Better Curation 4
  - 1.2.2 More Interpretations 4
- 1.3 Disadvantages of the Digital Humanities 4
- 1.4 The Hermeneutical Spiral 4

2 W2: Measurement / quantification and the Humanities 6
- 2.1 Digital Humanities as Anti-Cartesianism 6
- 2.2 Hermeneutica 6
  - 2.2.1 Problems with Hermeneutica 6
- 2.3 Voyant 6
  - 2.3.1 Features of Voyant 6

3 W3: Concordance and Analysis: Introduction to Voyant 7
- 3.1 The Remix 7
- 3.2 Concordances 7
- 3.3 Defintion of key terms 7
- 3.4 Contexts and dimensionality reduction 7

4 W5: Thematic analysis in the Humanities: Topic Modeling 8
- 4.1 From Week 4 8
- 4.2 Metadata 8
- 4.3 Topic Models 8
- 4.4 Model Checking 9
- 4.5 Pitfalls in Topic Modeling 9
- 4.6 Limitations of Topic Modeling 9

5 W8: The notion of ”style” – Towards Computational Criticism **10**
5.1 Russian formalism . . . 10
5.2 Jockers Ch. 9 . . . 10

6 W9: WordNet **11**
6.1 Overview of WordNet . . . 11
6.2 Difference between WordNet and word vectors . . . 11
6.2.1 WordNet . . . 11
6.2.2 Word vectors . . . 11
6.3 WordNet relationships . . . 11
6.4 Different kinds of relationships in WordNet . . . 12
6.5 Synsets . . . 12
6.6 Things that can be done with WordNet . . . 12
6.7 Verifying the word-sense/synset . . . 12
6.8 Extended Open Multilingual WordNet . . . 12
6.9 History and Genealogy of WordNet . . . 13
6.10 Before WordNet . . . 13
6.11 Problems with Semantic Networks . . . 13
6.12 Spreading activation . . . 13
6.13 Cross-Part of Speech and Adjective Peculiarities . . . 13

7 W10: Hands-on with Classifiers **15**
7.1 Support Vector Machine (SVM) . . . 15
7.2 Sentiment and Emotion . . . 15
7.3 Counting . . . 15
7.4 “Bag-of-words” approach . . . 16
7.5 Subjectivity in sentimental analysis . . . 16
7.6 TextBlob . . . 16
7.7 Pitfalls in sentiment analysis . . . 16

List of Weekly Readings

Week 1

- Owens, Trevor. "Defining Data for Humanists: Text, Artifact, Information or Evidence?" *Journal of Digital Humanities* 1, no. 1 (2011)
- Rockwell & Sinclair, *Hermeneutica*, Chapter 1

Week 2

- Rockwell & Sinclair, *Hermeneutica*, Chapter 2

Week 3

- Rockwell & Sinclair, *Hermeneutica*, Chapters 3-4

Week 4

- Rockwell & Sinclair, *Hermeneutica*, Chapters 5-7

Week 5

- Jockers, *Macroanalysis*, Chapters 5-6

Week 6

- All Stanford Literary Lab pamphlets

Week 8

- Jockers, *Macroanalysis*, Chapters 7-9

Week 9

- Montfort, Section on Wordnet in *Exploratory Programming for the Arts and Humanities*

Week 10

- Montfort, Section on classifiers in *Exploratory Programming for the Arts and Humanities*

Week 11

- Rockwell & Sinclair, *Hermeneutica*, Chapter 9

1 W1: Introduction

1.1 What are the Digital Humanities?

- Humanities with the use of digital tools
- A new approach

1.2 Benefits of digital tools in the Humanities

Digital tools can be used to:

- Create an experience for an audience **(Better curation)**
- Derive new meaning from an existing artefact **(More interpretations)**

1.2.1 Better Curation

- Similar to the Model-View-Controller (MVC) paradigm in user experience design
- Focus on visualisation for this course

1.2.2 More Interpretations

- Allow new patterns and irregularities in big data to be discovered
  - Discovery of relationships within data (by tracking correlations, co-occurences, etc.)
  - Discovery of trends over time (or over any continuous category)
  - Discovery of anomalies (across continuous or discrete categories)

1.3 Disadvantages of the Digital Humanities

- Over-reliance on digital tools
  - Black box problem
- Possibility of violating copyright laws

1.4 The Hermeneutical Spiral

- Rockwell and Sinclair aim to add computational thinking into the hermeneutical spiral
  - Developed the Voyant series of tools for text analysis
- With Digital Humanities, we now have another way to look at texts.
  - Before: Reader $\longleftrightarrow$ Text
  - After: $$\begin{tikzcd} & \text{Reader} \arrow[dl, Leftrightarrow] \arrow[dr, Leftrightarrow] & \\ \text{Text} \arrow[rr, Leftrightarrow] & & \text{Tool} \end{tikzcd}$$

- Individual and collective sense-making of value in the experience in a social context
    - Anu Helkkula ‘Characterizing Value as an Experience’
- Hermeneutica (hermenutical tools) are in Digital Humanities’ tradition of *problematizing* methods through developing tools

2 W2: Measurement / quantification and the Humanities

2.1 Digital Humanities as Anti-Cartesianism

- Descartes: ”I think, therefore I am”
  - Emphasizes solitary thought over groupthink
  - Inner monologue
  - Thought is fundamental
- Digital Humanities: Uses computational tools together with humanistic skills to interpret texts
  - Collaborative in nature
  - Dialogical
  - Text and tools are fundamental

2.2 Hermeneutica

- Small embeddable ”toys” that can be woven into essays
- Computational tools used to complement interpretation of texts

2.2.1 Problems with Hermeneutica

- Researchers using tools without understanding how they work
  - Tools can become too ”ready-at-hand” and therefore ”non-disclosing” (not open to scrutiny and critique)
  - *vide* Heidigger’s distinction between ”ready-at-hand” and ”ready-to-hand” tools
- Over-reliance on tools instead of humanistic interpretation could sidetrack the real conversation, which is about understanding texts in context
- Modernist commitment to (possibly false) progress through technique

2.3 Voyant

Accessible via https://voyant-tools.org/.

- Web-based reading and analysis environment

2.3.1 Features of Voyant

- Collocates graph
- Distribution graph: Uses stop-words to filter out common words
- Concordance and more...

3 W3: Concordance and Analysis: Introduction to Voyant

3.1 The Remix

- Everything is a Remix
  - Documentary by Kirby Ferguson
  - “Remixing is a folk art but the techniques are the same ones used at any level of creation: copy, transform, and combine. You could even say that everything is a remix.”
- Rearrangeable texts
  - Commonplace book in the West e.g. *Pride and Prejudice and Zombies*
  - Narrative paintings from West Bengal, India

3.2 Concordances

- Provides a new view of a corpus to support a consultative reading
- Was originally created for the Bible
- Was expensive to create, can be easily created computationally today
  - e.g. New York Times’ interactive concordance of 75 Years of the State of the Union Addresses

3.3 Defintion of key terms

- **Bag of words:** Per page or per document representation of words
- **Term Frequency - Inverse Document Frequency (Tf-Idf):**
  Statistic indicating how important a term is relative to a particular document
- **Semantics:** vector space

3.4 Contexts and dimensionality reduction

- With a large number of contexts (dimensions), we need to perform dimensionality reduction to visualise information, methods include:
  - Correspondence analysis
  - Principal Components Analysis (PCA) [for continuous-valued dimensions]
  - t-SNE (t-Distributed Stochastic Neighbor Embedding)
  - Factor analysis
- **Note:** beyond the scope of the class

4 W5: Thematic analysis in the Humanities: Topic Modeling

4.1 From Week 4

- - Supervised learning
- Carve up space into manageable, identifiable space to draw conclusions from text
- Simplest case
  - One feature in feature-space (normalized average length of each line in a book)
  - Class for each data point is plotted along the y-axis (0 = prose, 1 = poetry)
- More classes = more dimensions
  - Require reduction in dimensions

4.2 Metadata

- Data for data
  - Useful in slicing and dicing the data to discover local, subset-specific trends
- Particularly important in humanistic studies
  - Data is very rarely homogeneous
    - Many micro-trends may be lurking in the data
  - Data is highly subject to various biases
    - Confirmation, selection, sampling bias etc.
    - Data is filtered through many layers of mediation
      - What libraries have found worth preserving
      - What critics have found worth praising
- Metadata is data about data (or second-order data)
  - Truth-claim
  - Was the truth-claim "falsified" (invalidated)?

4.3 Topic Models

- Algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents
- What do topic models do?
  - Organize a collection according to discovered themes

- Assigns every **word** in every **document** to one of a given number of **topics**
    - Topic: distribution of words: a guess made by algorithm about how words tend to co-occur in a document
    - Document: basic unit in terms of which the algorithm is treating the body of text being analyzed; modeled as a mixture of topics in different proportions
    - Algorithm: knows nothing about the content of any document and nothing about the order of words of any given document
- Topic modeling algorithm can create many different topic models
    - Based on different choices of parameters for the model
- For large, unstructured corpora: use the most interpretable topic model

4.4 Model Checking

- How can we compare topic models based on how interpretable they are?
    - Use statistical/mathematical measures to compare
    - Use inspection/visualization methods to compare
        - Usual approach for humanists
- Requirements for successful topic modeling
    - A sufficiently large corpus
        - No. of documents should be in the hundreds
    - Some familiarity with the corpus

4.5 Pitfalls in Topic Modeling

- Can treat junk as an oracle
- Poorly supervised machine-learning algorithm is like bad research assistant

4.6 Limitations of Topic Modeling

- Don't blindly trust the word-cloud visualization for TM
- Be aware that choices made about stopwords can shape results
- Treat topic modeling results as a heuristic, and not as evidence
- Know the limitations of your tool

5 W8: The notion of "style" – Towards Computational Criticism

5.1 Russian formalism

- Precursor to digital humanities
- School of thought that developed in Russia in the early 20th century
- Russian formalists thought of literature (or, more broadly, culture) as "combinatorial"
  - Culture consists of the constant reshuffling of the ceratain pre-existing thematic form in various combinations
    - Certain key elements ocur in different elements occur in different folk tales in various combinations and permutations
  - One development of formalism was into "structuralism"
- Alexander and Alexey Veselovsky
  - Thought beyond single texts or copora to *comparative studies of different literatures and cultures* that together make up the imaginative world of all mankind

5.2 Jockers Ch. 9

- Experiment with British, Irish and Scottish corpora
  - Topic modelling of themes and their relative saliences
  - Most salient theme for British: hounds and shooting sport
  - Most salient theme for Ireland: dialect, Ireland, lords and ladies, tears and sorrow
  - Most salient theme for Scotland: Scottish dialect, Scotland
  - Most salient theme for male corpora: pistons and other guns
  - Most salient theme for female corpora: female fashion
- Misclassified novels are outliers, exceptions to the norms
- We can find pattern among the pattern-breakers by creating a network graph of all the novels
- Who is systemically different for authors
- Understand time and gender influence on theme and authors

6 W9: WordNet

6.1 Overview of WordNet

- A large, human-curated ”lexical database”
- Most frequently cited ”lexicographic resource” in the world
- Originally created at Princeton University in 1986
- A kind of ”semantic network”
- Words as vectors:
    - A vector representation of words in a text corpus
    - Created using unsupervised machine learning by training on the text corpus

6.2 Difference between WordNet and word vectors

6.2.1 WordNet

- Topological representation of the relationship between words
    - Graph structure, with words as vertices and relationship between them as edges
    - Specific types of relationships are *explicit*

Note: *Strictly speaking, the term ”words” here should be replaced by ”word-senses”; we will talk about that when we discuss synsets*

6.2.2 Word vectors

- **Geometrical** representation of the *relationship* between words
- Types of relationships were *implicit*

6.3 WordNet relationships

- Is-a relationship: Inheritance
- Has-a relationship: Possesses property
- Indirect inheritance
- Antonymic relationship

6.4 Different kinds of relationships in WordNet

- Synonymy: words with same meaning
- Antonymy: words with opposite meaning
- Hypernym: word more general than given word
- Hyponymy: word more specific than given word
- Meronymy: word standing in a **part-of** relationship to given word
- Holonymy: opposite of **meronymy**

6.5 Synsets

- Many words are polysemic (multiple meanings, or senses)
- WordNet graphs aren't graphs of words, but are graphs of **word-senses**
- Synset: Associated with each word in Wordnet is a list of synsets
  - A set of synonyms, all of which pertain to a specific word sense
  - Nodes in the graph are not words, but word-senses, each word=sense being represented by a synset
- Each element of a synset corresponds to a distinct word

6.6 Things that can be done with WordNet

- e.g. get all synsets that consist of "noun" senses of the word "stream"
- e.g. compute the path similarity between two word-senses
- e.g. Enumerate all the different words that correspond to the same word-sense
- e.g. Make a text more abstract and general

6.7 Verifying the word-sense/synset

How do we know that we are on the right word-sense (right synset)?

- Check that word-sense (synset) has best (shortest) similarity with nearby synsets

6.8 Extended Open Multilingual WordNet

- Being developed at NTU's Linguistics Dept
- Extension of the original WordNet database developed by Princeton University

6.9 History and Genealogy of WordNet

- Comes from the earlier and import period of Symbolic AI
- Not about Big Data but about smaler datasets, emphasizing generalizability and explainablity
- Arose from work in AI inspired by Cognitive Psychology, Philosophy and Information Science (knowledge representation), called "Semantic Networks" (also known as "conceptual graphs")
    - Extreme example of this is the still-ongoing Cyc project by (Doug Lenat), a humongous semantic network intended to capture all the knowledge in the world that can be obtained from text

6.10 Before WordNet

- Oldest known semantic network drawn in 3rd century AD by Greek philosopher Porphyry in his commentary on Aristotle's categories
    - Poryphyry used it to illustrate Aristotle's method of defining categories by specifying the genus or general type and the differentiae that distinguish different subtypes of the same supertype

6.11 Problems with Semantic Networks

- Real world data may be inconsistent or contradictory
- Exceptions may occur

6.12 Spreading activation

- Solution to problem of how to allocate contextual attention

6.13 Cross-Part of Speech and Adjective Peculiarities

- Majority of WordNet's relations connect words from the same part of speech (POS)
- WordNet consists of four sub-nets (nouns, verbs, adjectives and adverbs) with few cross-POS-pointers
    - Cross-POS relations include the "morphosemantic" links that hold among semantically similar words sharing a stem with the same meaning:
        - Observe (verb), observant (adjective), observation, observatory (nouns)
- Verb synsets are arranged into hierarchies as well
- Verbs towards the bottom of the trees (troponyms) express increasingly specific manners characterizing an event, as in {communicate-talk-whisper}

- Specific manner expressed depends on the semantic field; volume (as in the example above) is just one dimension along which verbs can be elaborated.
- Others are speed move-jog-run or intensity of emotion like-love-idolize. Verbs describing events that necessarily and unidirectionally entail one another are linked: {buy}-{pay}, {succeed}-{try}, {show-see}, etc.

- Adjectives are organized in terms of antonymy. Pairs of ”direct” antonyms like wet-dry and young-old reflect the strong semantic polarization of their members
  - Each of these polar adjectives in turn is linked to a number of ”semantically similar” ones: dry is linked to parched, desiccated and bone-dry, wet to soggy, waterlogged, etc.
  - Semantically similar adjectives are ”indirect antonyms” of the central member of the opposite pole
  - Relational adjectives (”pertainyms”) point to the nouns they are derived from (criminal-crime)
- There are only few adverbs in WordNet
  - Majority of English adverbs are derived from adjectives via morphological affixation

7 W10: Hands-on with Classifiers

7.1 Support Vector Machine (SVM)

- Default classifier provided in Stylo GUI
- SVMs build on the intuition that a “linear” hyperplane separating the candidate classes is easier to induce than a “non-linear” hyperplane
- Solves the problem by introducing a new feature that makes for a linear (hyper)plane as the frontier between classes
    - Take a low dimensional input space and transforming it to a higher dimensional space by applying a special “kernel” function
    - Converts a *not linearly separable* problem to a much more manageable separable one
    - Coordinates of the individual instances are like *”supports”* holding up the classification frontier (hyper)plane
    - Maximise the distance between nearest data point (either class) and hyperplane to decide the right hyperplane
        - Distance is known as the **margin**
        - Margin proportional to robustness of classifier

7.2 Sentiment and Emotion

- What’s the difference?
- Sentiment is like an overall mood (more ambient)
    - Emotion is more event-like (more transactional)
- Why are sentiment/emotion important?
    - Emotion:
        - Provide motivation for specific human actions
            - Important for practical human-robot interactions
        - “Signal” of attitude of writer towards his/her subject
        - Understanding attitude provides an overall context that may be otherwise missing
            - Automatically classify reviews into positive or negative
            - Even, predict the stock market by gauging overall mood from social mediascape

7.3 Counting

- Not probably is going to be useful to get at sentiment

7.4 "Bag-of-words" approach

- Cannot distinguish tone

7.5 Subjectivity in sentimental analysis

- Algorithm must understand subjectivity
- Subjectivity is encoded in relational information between words (syntax)
- "Bag-of-words" model loses information about sequentiality between words
- Solution: We need a richer language model
- TextBlob provides a richer model that takes word order into account

7.6 TextBlob

- Acts as wrapper around Python implementation of the Natural Language ToolKit (NLTK)
- Comes with its inbuilt (pre-trained) classifiers
- Choose your classifier
- Pre-trained classifier is heavily biased towards contemporary "standard English"
- Train the classifier yourself if corpus consists of non-"standard" English
- Does not return a single numeric value but a complex structure with *polarity* and *subjectivity*
  - Positivity: proxy for the confidence with which it is being considered positive or negative
  - Subjectivity: number which is a proxy for whether the sentence is subjective or not

7.7 Pitfalls in sentiment analysis

- Semantic drift
  - Words changing in meaning over time
  - Sentiment-laden adjectives are most vulnerable to this
    - Nouns and verbs are much more stable
- Sarcasm/irony hard to handle
- Beware of potential heterogeneity among people doing the training (if labeling training set)
  - Colloquial sentiment-bearing adjectives or idiomatic uses may be opaque to non-native speakers or people drawn from a different demographics
  - Particularly problematic if you are doing the training distributively e.g. Amazon's Mechanical Turk