Week 8 Video 3

Text Mining

Text Mining

- Related to discourse processing, computational linguistics, natural language processing…

Text Mining

- Is hard
- Is very different from the types of interaction data and course data I’ve discussed throughout the rest of the class

Different Stuff Works

- Stuff that works poorly in interaction data works great in text mining
  - Support Vector Machines

- Stuff that works great in interaction data is less relevant in text mining
  - Bayesian Knowledge Tracing, IRT

Interesting Attributes of Textual Data

- Really high dimensionality
  - Many many words in a corpus of data

- Multiple levels of analysis that look very different from each other
  - From individual phonemes and graphemes to entire books

Analyses often conducted

- At level of whether individual words are seen

- A popular algorithm for this is Latent Semantic Analysis (LSA)
  - Represents utterances or paragraphs such that each row is an utterance or paragraph
  - And each column is a word that can be present (1) or absent (0)
  - Conducts singular value decomposition (a matrix factorization algorithm conceptually similar to factor analysis) to find structure
  - Does not look at syntax of sentences, just what words are present (Landauer, Foltz, & Laham, 1998)
    - Does consider co-occurrence of words across large corpuses

Alternatively, analysis is conducted using

- Pairs of words, in order, called *bigrams*
- Triplets of words, in order, called *trigrams*

- “Colorless green ideas sleep furiously”
- Bigrams: “Colorless green”, “green ideas”, “ideas sleep”, “sleep furiously”
- Trigrams: “Colorless green ideas”, “green ideas sleep”, “ideas sleep furiously”

LightSide

- Toolkit that supports turning utterances into unigrams, bigrams, and trigrams, as well as more powerful feature extraction methods, and then running data set through a range of powerful machine learning algorithms

- http://www.cs.cmu.edu/~cprose/LightSIDE.html

Semantic Tagging

- Another approach is to reduce specific words to semantic categories, such as sports, business, time, prior to analysis

- Allows easier categorization of types of utterances that is less dependent on presence of specific words

Semantic Taggers

- http://www.liwc.net/
- http://ucrel.lancs.ac.uk/wmatrix/

Coherence

- Another type of tool can provide coherence metrics
- A modern, updated version of reading level metrics such as Fleisch-Kincaid
- How hard is a text to read?

Coh-Metrix

- A popular tool that provides several metrics about a text, including coherence
- http://cohmetrix.memphis.edu/cohmetrixpr/index.html
- http://tea.cohmetrix.com/

Coh-Metrix

- Over 100 metrics
- Distilled into five core characteristics of a text

1. Concrete (vs. abstract) words
2. Syntactic complexity
3. Narrativity (vs. expository)
4. Referential coherence
5. Situational coherence

(Graesser, McNamara, & Kulikowich, 2011)

Many uses of text mining in education

- Analysis of sentiment and emotions within learner utterances (D'Mello et al., 2008)
- Studying content of online discussion forums
- Studying pair collaboration online (Dyke et al., 2013)
- Enhancing tutorial dialogues between students and online tutoring systems (Forsyth et al., 2013)
- Studying learner expertise in think-aloud data (Worsley & Blikstein, 2011)

Next lecture

- Hidden Markov Models