Semester review

INFO 3350/6350: Lecture 21

Humanities questions

Technical methods

Humanities questions

How do humanists use texts?

- ~~Interpret single texts closely~~
- Identify differences and similarities between many texts
- Track changes in content and style over time
- Compare texts by different groups
- Use texts as models of communication/narrative
- Correlate textual and social events

Specific examples …

Emotional arcs and narrative archetypes

Gender and racial stereotypes in characterization

Literary traits of canonical or successful novels

Historical evolution of narrative pace

Comparative differentiation and stability of genres

Identifying and mapping topical content

Speaker ideology and influence on political debates

Power and use of geographic space

Adapting pretrained models to work with old books

Divide narrative from non-narrative texts

Find poems similar to a given example

Social media as a model of storytelling

Technical methods

Representations v. applications

Representations convert text into meaningful data

- Bag of words
- Static embeddings
- Contextual embeddings
- Paratextual data
- …

Applications use representations to perform a task

- Clustering
- Classification
- Topic modeling
- Regression
- …

Supervised v. unsupervised learning

Supervised methods seek to identify and use the features that differentiate two (or more) classes of objects on the basis of labeled examples

- Genre *labeling*
- Entity recognition
- Narrative class prediction
- Temporal span labeling
- Persona *prediction*

Unsupervised methods use known features to find structure in unlabeled data

- Genre *detection*
- Topic modeling
- Document similarity
- Fightin' Words
- Persona *modeling*

Methodological progression

- Word counts and dictionary lookups
  - Clustering (*k*-Means, DBSCAN)
  - Classification
  - Topic models
- Feature importance, model inspection
- Static embeddings
- BERT and contextual embeddings
  - Transformer architecture, encoder/decoder systems
- Generative models

Better representations ⇒ better performance (usually)