Review for Final

Natural Language Processing

10. Semantic Similarity

The Basics of Semantic Similarity

- Word similarity
  - BoW statistical similarity vs. structural similarity
- Basic statistical measure: PPMI
  - Dependent vs. independent word probabilities

10. Semantic Similarity

Statistical Word Similarity

- Vector semantics
  - Term-document matrix
    - Raw counts or tf-idf
  - Sparse vs. dense vectors
- LSA
  - Dimension reduction
  - SVD (singular value decomposition)
- Cosine similarity
  - Of words, with respect to a corpus

10. Semantic Similarity

Structural Word Similarity

- Using parse contexts
  - Adjective-Noun, Verb-Object patterns
- Ontological distance
  - Using WordNet hyponym trees

10. Semantic Similarity

Measures of Document Similarity

- Jaccard distance
  - Intersection vs. union
- Cosine similarity, of documents
  - Dot product
  - Eucidian distance of vectors in vector space
  - Advantage of using tf-idf values in the feature vectors (instead of just raw counts)

10. Semantic Similarity

Measures of Document Similarity

- Jaccard distance
  - Intersection vs. union
- Cosine similarity, of documents
  - Dot product
  - Euclidian distance of vectors in vector space
  - Advantage of using tf-idf values in the feature vectors (instead of just raw counts)
- Heilinger distance
  - Documents as probability distributions
- Weaknesses of term-document matrix methods
  - Countering by normalizing to synsets

10. Semantic Similarity

Caveats in Applying Measures of Semantic Similarity

- Understanding importance of lexicography (having collocation headwords like “boot camp”)
- Similarity vs. synonymy, issues of conflation

10. Semantic Similarity

Uses for Word Similarity

- Search expansion
- Tracking linguistic change over time
- Plagiarism detection
- Detection of information gain vs. previous news
- Source criticism (detection of multiauthorship in a text)

DataScience@SMU

Review for Final

Natural Language Processing

11. Document Clustering

Document Clustering

- What is document clustering?
- Clustering vs. classification
- Methods: Centroid vs. hierarchical

11. Document Clustering

Centroid Clustering

- What is a centroid?
- K-means clustering
  - Distance between vs. within clusters
  - Weaknesses
    - Fixed number of clusters
    - Unlucky random seeding

11. Document Clustering

Hierarchical Clustering

- AGNES vs. DIANA
- Ward's minimum variance
- Agglomerative procedure
- Divisive procedure
- "Cutting" a dendogram

11. Document Clustering

Working with Clusters

- Visualizing clusters
  - Challenge of visualizing in a 2D space
  - How PCA helps
  - Labelling with *top n features*
- Applications of clustering
  - Cohort profiling (e.g., clustering top *n* movies)
  - Search result profiling

12. Text Classification

Document Classification

- Subject based vs. descriptor based
- Binary vs. multiclass
- Text classification vs. document classification
  - E.g., clustering news headlines vs. news articles
- Text classification landscape
  (dominated by one type)

12. Text Classification

Content-Based Classification

- With multinomial naïve Bayes
  - *Prior* vs. *posterior* probabilities
  - Probabilities of the *class* vs. of the *predictor*
  - Example application: making a spam filter

12. Text Classification

Content-Based Classification

- With multinomial naïve Bayes
  - *Prior* vs. *posterior* probabilities
  - Probabilities of the *class* vs. of the *predictor*
  - Example application: making a spam filter
- With SVMs
  - Separation in the margins of feature vector space
  - Why it is called a hyperplane
  - Outlier detection
  - Projecting to higher dimensionality: Why?
  - Projecting to higher dimensionality: How?
    - Kernel trick—know one example
  - Multiclass with SVM: cascading binary classification

12. Text Classification

Descriptor-Based Classification

- Two approaches
  - IR approach
  - Empty taxonomy

12. Text Classification

Descriptor-Based Classification

- IR approach
  - Postprocess query results for *strong hits*
  - What is a *strong hit?*
  - Where does the SVM come in?

12. Text Classification

Descriptor-Based Classification

- Taxonomy approach
  - Building queries from empty taxonomies
  - SVM comes again after IR

DataScience@SMU

Review for Final

Natural Language Processing

13. Topic Modeling

Types of Topic Modeling

- Organic topic modeling
- Canonical topic modeling
- Entity-centric topic modeling
- AI community bias in favor of organic

13. Topic Modeling

Organic Topic Modeling

- With LSA
  - Topic-topic matrix
  - Separating documents
- With LDA
  - Procedural difference from LSA
  - Concentration parameters and their effects
- With NMF
  - Implementation differences from LDA
  - Psychological difference from LDA

13. Topic Modeling

Working with Organic Topic Models

- Statistical interference
- Troubleshooting
- Applications

13. Topic Modeling

Canonical Topic Modeling

- What are canonical topic models?
- Examples in real life
- Organic approach—constraining output to canon
- IR approach—intensional vs. extensional comparison of topics

13. Topic Modeling

Entity-Centric Topic Modeling

- Entity weight of topics
- Ways of determining definitive list of entities
  - By reference vs. by description
  - Entity-first vs. topic-first approach

13. Topic Modeling

Curation of Topic Models

- For what? And why?
  - Prune topics, edit labels, etc.

DataScience@SMU

Review for Final

Natural Language Processing

14. Sentiment and Rhetoric

General Sentiment Scoring

- ML approach
- Lexical KB approach

14. Sentiment and Rhetoric

ML Approach

- Pros and cons
- Procedure
  - Choosing a classifier
- Blueprint

14. Sentiment and Rhetoric

Lexical Approach

- Pros and cons
- Procedure
  - Choosing a lexicon
- Blueprint

14. Sentiment and Rhetoric

Advanced Sentiment Analysis

- Pairing sentiment with objects or themes
- Adding dimensionality to pos-neg sentiment
- Sentiment and rhetoric

14. Sentiment and Rhetoric

Attaching objects or themes to sentiment

- Using a chunker
  - Procedure
  - Vulnerabilities (negation, cross-chunk attachments)
- Using a dependency parser
  - Procedure

14. Sentiment and Rhetoric

Adding dimensionality to sentiment

- Typologies of emotion
- Hierarchical listings, e.g., Shaver
- Modifying our blueprint for lexical sentiment analysis

14. Sentiment and Rhetoric

A hybrid approach to adding themes/dimensions

- Semiautomated feature engineering
- Modifying our blueprint for lexical sentiment analysis again
- Example with user reviews of products

14. Sentiment and Rhetoric

Sentiment and Insight

- Presenting roll-ups of sentiment numbers, with examples
- Sentiment and rhetoric
- Sentiment and demographics

DataScience@SMU