Chapter 6 Vector Semantics and Embeddings

9/30/20 7:01 PM

Lexical Semantics

Lemmas and senses
- Each word (citation form) is a lemma
- Each aspects of meaning of a word is a word sense
- Lemmas can be polysemous (a word has multiple senses)

Synonymy
- Words that have same propositional meaning
- Replacing words won't change the condition truth of a sentence

Word similarity and relatedness
- Semantic field: a set of words which cover a particular semantic domain and bear structured relations with each other
- Topic models: apply unsupervised learning on large sets of texts to introduce sets of associated words from text

Semantic frames and roles
- A semantic frame is a set of words that denote perspectives or participants in a particular type of event

Connotations
- Connotations: affective meanings of a word
- Early research shows words vary along three important dimensions of affective meaning:
  - Valence: pleasantness of the stimulus
  - Arousal: intensity of emotion provoked by stimulus
  - Dominance: degree of control exerted by stimulus

Vector Semantics

Embeddings
- Vectors for representing words

Term-document matrix
- Each row represents a word in vocabulary
- Each column represents a document in a collection of documents
- Each cell represents number of occurrence of a word in a particular document

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

- Each document will have a representing vector of |V| dimensions

Information Retrieval (IR)
- Task of finding document that best matches a query

Word-word matrix (term-context matrix)
- Each word will have a representing vector (count of neighbor words)
- Usually vector is of large dimension and sparse

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Figure 6.5 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

- (Cosine) similarity of words: inner product of two word vectors

TF-IDF Algorithm
- Term frequency: frequency of word t in document d
  - $\mathrm{tf}_{t,d} = \mathrm{count}(t,d)$
  - $\mathrm{tf}_{t,d} = \log_{10}(\mathrm{count}(t,d)+1)$
- Document frequency: number of documents that a term t appears in
- Collection frequency: total number of times the term t appears in whole collection
- Inverse document frequency (idf):
  - N: number of documents in collection
  - Df: document frequency of a word

$$\mathrm{idf}_t = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

- TF-IDF weights:
  - $w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$

User case of TF-IDF
- Determine word similarity
- Compute centroid of word vectors of a document to get document vector, then use for computing document similarity

Word2Vec

Skip gram with negative sampling
1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the regression weights as the embeddings.