Extracting and Aggregating Aspect-Level Sentiment from Product Reviews

Desmond Ong, Shane Soh, Matthew Long

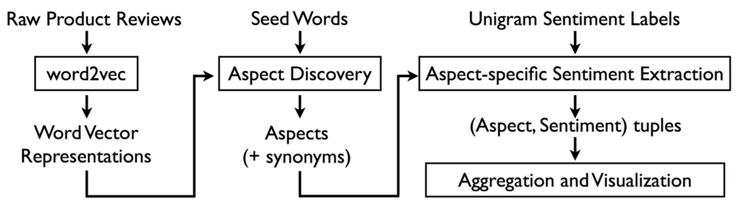
CS224D

Motivation

Consumers not only have access to online stores through which they might purchase anything she desires, but they also have unprecedented access to a deluge of information—most notably, product reviews written by other consumers—with which they can make their decision. Sifting through hundreds of reviews across tens of different websites to acquire specific information about the product and its important attributes (e.g. *battery life* for electronics) is a time-consuming chore. It is difficult to automatically summarize this information, especially because of the heterogeneity of attributes across different product categories. Because of this, there has been much recent research tackling the two separate but connected components of this problem: (1) automatic discovery of aspects, and (2) aspect-specific sentiment analysis.

Problem

- Aspect Discovery
  - *Problem* : Given a short list of attributes ("seed" attributes), we discover an expanded list of attributes by returning top *n* related words vectors based on cosine similarity
  - *Dataset* : 6 million reviews of electronics product from Amazon.com[1]
- Aspect-Specific Sentiment Extraction
  - *Problem* : Given a set of reviews (for a single product) identify (aspect, aspect-related sentiment) tuples
  - *Evaluation* : We will construct word cloud visualizations color coded to express sentiment. We will perform comparisons with "expert review" sites, such as CNET and DPReview, and the aspect-specific ratings on Google Shopping.



Acknowledgments

We would like to thank Richard Socher and the teaching assistants of CS224D for teaching us the skills necessary to make this project a success and the Stanford Deep Social Learning Lab for providing computing resources.

[1]McAuley, J., Targett, C., Shi, J., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. *ACM Special Interest Group on Information Retrieval (SIGIR)*

[2]placeholder

Approach

- Aspect Discovery
  - Tokenized reviews with NLTK's punkt tokenizer
  - Removed all non-alphanumeric characters and replaced digits with *DG*
  - Detected and combined common bigrams
  - Trained word2vec (CBOW, Skipgram) with tuned hyperparameters using Gensim package
  - Each attribute in seed set expanded with top *n* nearest-neighbors in terms of cosine similarity
- Aspect-Specific Sentiment Extraction
  - Needs updating

Models

- Aspect Discover
  - Three models were trained, the most successful of which used window sizes of 10 and 300 dimensional feature vectors. Frequency-based pruning was also performed.
- Aspect-Specific Sentiment Extraction

Results

- The word2vec model was able to learn not only the synonyms for the product aspects we listed in the "seed" set, but also could discover aspects of particular products that were not "seeded". For instance, the query *tripod* returned results *ball_head* and *quick_release*, both of which refer to features of tripods that are non-obvious to a person unfamiliar with photography.

Query	portability	contrast	tripod
Results	(u'portability,', 0.72859996557235718), (u'compactness', 0.64743077754974365), (u'mobility', 0.60842603445053101), (u'versatility', 0.5763777494430542), (u'simplicity', 0.53962129354476929), (u'lightness', 0.53950369358062744), (u'convenience', 0.53897607326507568), (u'ruggedness', 0.5272858738899231), (u'versatility,', 0.5055851936340332), (u'thinness', 0.49253776669502258)	(u'contrast,', 0.65686732530593872), (u'sharpness', 0.62712550163269043), (u'color_saturation', 0.60933655500411987), (u'saturation', 0.57076853513717651), (u'brightness', 0.5553707480430603), (u'gamma', 0.53090476989746094), (u'shadow_detail', 0.52805298566818237), (u'color_accuracy', 0.52408510446548462), (u'dynamic_range', 0.52167940139770508), (u'black_levels', 0.51741272211074829)	(u'monopod', 0.74430769681930542), (u'tripod,', 0.71975594758987427), (u'ball_head', 0.70861411094665527), (u'tripods', 0.68399727344512939), (u'ballhead', 0.60356354713439941), (u'manfrotto', 0.59598124027252197), (u'monopod,', 0.58229464292526245), (u'pole', 0.56997144222259521), (u'quick_release', 0.549965500831604), (u'cold_shoe', 0.5460544228553772)

Maybe new table?

Conclusions