CSE 190 – Lecture 11
Data Mining and Predictive Analytics

Text mining Part 2
• Hints and tips to follow after the lecture
• **Yes,** there’s a weakness in the task 2 evaluation (there was last year too, so I changed the way the test set was constructed, but apparently made it even worse...)
Assignment 1... due in one week (and one day)

• **So,** we’ll be flexible about evaluation
  • Full marks (for that task) if you find the deficiency
  • Full marks for a reasonable solution that doesn’t exploit the deficiency
  • No marks if you don’t attempt anything...
  • \(\frac{1}{2}\) (ish) marks if your solution is about as good as the HW2 solution

• So concentrate most of your effort on Task 1, where most of the grading variance will be
Assignment 1... due in one week (and one day)

- Extra office hours to discuss (on Thursday or Friday)
- **No class on Wednesday** (veteran’s day)
- Kaggle times are in UTC
What kind of quantities can we model, and what kind of prediction tasks can we solve using text?
Does this article have a positive or negative sentiment about the subject being discussed?
SAN FRANCISCO — While Apple has been preparing to release its first wearable computers, the company has also been busy assembling a team to work on an automobile.

The company has collected about 200 people over the last few years — both from inside Apple and potential competitors like Tesla — to develop technologies for an electric car, according to two people with knowledge of the company's plans, who asked not to be named because the plans were private.

The car project is still in its prototype phase, one person said, meaning it is probably many years away from being a viable product and might never reach the mass market if the quality of the vehicle fails to impress Apple's executives.

It could also go nowhere if Apple struggles to find a compelling business opportunity in automobiles, a business that typically has much lower sales margins than

Prediction tasks involving text

Which of these reviews am I most likely to agree with or find helpful?

Most Helpful Customer Reviews

1,000 of 1,928 people found the following review helpful

🌟🌟🌟🌟🌟 Le Creuset on a budget
By N. Lefond on October 24, 2007
Color Name: Caribbean Blue | Size Name: 6 qt. | Verified Purchase
Enamel on cast iron cookware like this, was, until recently, only available from makers like Le Creuset. Lately, several lower cost makers have come on the scene, like Target and Innova. The new budget priced Lodge cookware is in the same price range as the low cost alternatives but completely out performs them.

I have all of the brands I have mentioned. The Lodge is the same weight as the Le Creuset which is much heavier than the other budget models. The ridge where the lid and sides meet is a matt black porcelain on the Lodge and Le Creuset but lis just exposed cast iron for the other budget models (which leads to rusting if you are not careful). The porcelain resists staining (even tomato sauces) in the Lodge and Le Creuset but the other budget models stain very easily. And finally, the Lodge and Le Creuset maintain a very polished interior finish that resists sticking which others do not. So, I see no performance differences at all between the Le Creuset and the Lodge whereas the comparably priced budget models are certainly inferior.

If you plan of using these pots very heavily (every day for example) you might want to upgrade to the higher priced Lodge product. It has 4 coatings of enamel as opposed to 2 in this model. But if you use them once or twice a week I dont think you will need the added wear resistance.

47 Comments | Was this review helpful to you? Yes No

1,105 of 1,164 people found the following review helpful

By J. G. Pavlovich on March 2, 2008
Color Name: Island Spice Red | Size Name: 6 qt. | Verified Purchase
This is a terrific value. The quality and performance match my Le Creuset pieces at a fraction of the price. The only slight design flaw I have found is that the rounded bottom makes browning large pieces of meat awkward. Other than that I have no complaints. Even heating. Easy clean up. I use it several times a week.
UPDATE: I found a second minor problem. The inside rim of the lid has a couple of raised spots which prevent the lid from seating tightly. This causes steam to escape much faster than I would like during a long braise or stew.

Update 2: Three years in I am dropping my rating to three stars. It's still a decent pot at a bargain price, but it will not be an heirloom piece like my Le Creuset. The loose fitting lid turns.

Update: Finally, I am giving it 2 stars. I am returning it. It is a good bargain, but it is not a good replacement for my Le Creuset.
Which of these articles are relevant to my interests?
Prediction tasks involving text

Find me articles similar to this one

Meatloaf That Conquers the Mundane

FEB 13, 2015

City Kitchen
By DAVID TANIS

I was raised on Midwestern meatloaf. My mother's dependable recipe did not vary: Ground beef, grated onion and carrot, and a little oatmeal were the main ingredients, along with a dash of "seasoned salt." A ribbon of bottled chili sauce ran down a gully in the center.

Served hot, accompanied by Tater Tots, it was dinner. Served cold for lunch, it was always a sandwich on white bread, with potato chips on the side. It was usually moist and tasty but never remarkable, and there was no way you could call it anything but meatloaf.

Do I harbor a kind of nostalgia for it? Yes. But would I use that recipe now? I think not.

I have a friend from Brussels who loves to entertain. Of his dinner party repertoire, one dish is most requested and admired. It is pain de veau, served with a vermouth-splashed mushroom sauce. In French, it sounds elegant. Translated into English — veal loaf — it sounds dull.

The Italian word for meatloaf is polpettone. (Polpette are Italian meatballs; polpettone are meatballs, too, but more diminutive.) This substantial family-size meatball, whether ovoid or elongated, plain or fancy, served with tomato sauce or not, is beloved both in Italy and in Italian communities throughout the world. Aside from its melodic, polysyllabic name, polpettone is always well seasoned, prepared with care and served with gusto.

It is usually a combination of different kinds of ground meat, typically beef, pork and veal, in equal Parts. Grated cheese and herbs are
Feature vectors from text

**Bag-of-Words models**

The Peculiar Genius of Bjork

 Solo musician or master collaborator? For her new album, Bjork has merged the two sides of her artistry to create a new experience of music — again.

F_text = [150, 0, 0, 0, 0, 0, 0, ... , 0]

a aardvark zoetrope

musician, who creates her music in an emotional cocoon, tinkering with technologies, concepts and feelings; and Bjork the producer and curator, who seeks out cell thermocouples, aardvarks and zoetropes, among other oddities.
Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.

yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma Medium tan

These two documents have exactly the same representation in this model, i.e., we’re completely ignoring syntax. This is called a “bag-of-words” model.
Let’s do some inference!

**Problem 1:** Sentiment analysis

Let’s build a predictor of the form:

$$f(\text{text}) \rightarrow \text{rating}$$

using a model based on linear regression:

$$\text{rating} \sim \alpha + \sum_{w \in \text{text}} \text{count}(w) \cdot \theta_w$$
CSE 190 – Lecture 12
Data Mining and Predictive Analytics

TF-IDF
When we studied recommender systems, we looked at:

- Approaches based on measuring similarity (cosine, jaccard, etc.)
- Approaches based on dimensionality reduction

Today we’ll look at the same two concepts, but using textual representations
Finding relevant terms

So far we’ve dealt with huge vocabularies just by identifying the most frequently occurring words.

But! The most informative words may be those that occur very rarely, e.g.:

• Proper nouns (e.g. people’s names) may predict the content of an article even though they show up rarely.
• Extremely superlative (or extremely negative) language may appear rarely but be very predictive.
Finding relevant terms

e.g. imagine applying something like cosine similarity to the document representations we’ve seen so far.

e.g. are (the features of the reviews/IMDB descriptions of) these two documents “similar”, i.e., do they have high cosine similarity.
Finding relevant terms

e.g. imagine applying something like cosine similarity to the document representations we’ve seen so far
Finding relevant terms

So how can we estimate the “relevance” of a word in a document?
e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn’t ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we’ve got all the scoop on why Katy is set in her ways. Will these two ever bury the hatchet? Katy Perry & Taylor Swift Still Fighting?

“Taylor’s tried to reach out to make amends with Katy, but Katy is not going to accept it nor is she interested in having a friendship with Taylor,” a source tells HollywoodLife.com exclusively. “She wants nothing to do with Taylor. In Katy’s mind, Taylor shouldn’t even attempt to make a friendship happen. That ship has sailed.” While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor’s managed to mend the fences with Katy’s BFF Diplo, but it looks like Taylor and Katy won’t be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That’s not too much to ask.
So how can we estimate the “relevance” of a word in a document? e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn’t ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we’ve got all the scoop on why Katy is sticking with her stance. Will these two ever bury the hatchet? Katy Perry & Taylor Swift Still Fighting? "Taylor’s tried to reach out to make amends with Katy, but Katy is not going to accept it nor is she interested in having a friendship with Taylor,” a source tells HollywoodLife.com exclusively. “She wants nothing to do with Taylor. In Katy’s mind, Taylor shouldn’t even attempt to make a friendship happen. That ship has sailed.” While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor’s managed to mend the fences with Katy’s BFF Diplo, but it looks like Taylor and Katy won’t be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That’s not too much to ask.
So how can we estimate the "relevance" of a word in a document? e.g. which words in this document might help us to determine its content, or to find similar documents?

Despite Taylor making moves to end her long-standing feud with Katy, HollywoodLife.com has learned exclusively that Katy isn’t ready to let things go! Looks like the bad blood between Kat Perry, 29, and Taylor Swift, 25, is going to continue brewing. A source tells HollywoodLife.com exclusively that Katy prefers that their frenemy battle lines remain drawn, and we’ve got all the scoop on why Katy’s tried to reach out to make amends with Katy, but Katy is not going to accept it nor is she interested in having a friendship happen. That ship has sailed.” While we love that Taylor has tried to end the feud, we can understand where Katy is coming from. If a friendship would ultimately never work, then why bother? These two have taken their feud everywhere from social media to magazines to the Super Bowl. Taylor’s managed to mend the fences with Katy’s BFF Diplo, but it looks like Taylor and Katy won’t be posing for pics together in the near future. Katy Perry & Taylor Swift: Their Drama Hits All-Time High At the very least, Katy and Taylor could tone down their feud. That’s not too much to ask.
So how can we estimate the “relevance” of a word in a document?

**Q:** The document discusses “the” more than it discusses “Taylor Swift”, so how might we come to the conclusion that “Taylor Swift” is the more relevant expression?

**A:** It discusses “the” **no more** than other documents do, but it discusses “Taylor Swift” **much more**
Finding relevant terms

Term frequency & document frequency

**Term frequency** ~ How much does the term appear in the document

**Inverse document frequency** ~ How “rare” is this term across all documents
Finding relevant terms

Term frequency & document frequency

$$tf(t, d) = \text{# times } t \text{ appears in } d$$

$$df(t, D) = \text{# times } t \text{ appears in } D$$

$$idf = tf(t, d) \times \log \left( \frac{1}{df(t, D)} \right)$$
Finding relevant terms

**Term frequency & document frequency**

"Term frequency": \( tf(t, d) \) = number of times the term \( t \) appears in the document \( d \)

e.g. \( tf(\text{"Taylor Swift"}, \text{that news article}) = 3 \)

"Inverse document frequency": \( idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \)

term (e.g. "Taylor Swift") \hspace{1em} set of documents

\( N \)
Finding relevant terms

Term frequency & document frequency

**TF-IDF** is high $\rightarrow$ this word appears much more frequently in this document compared to other documents

**TF-IDF** is low $\rightarrow$ this word appears infrequently in this document, or it appears in many documents

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$
Finding relevant terms

Term frequency & document frequency

$\textbf{tf}$ is sometimes defined differently, e.g.:

$$tf'(t, d) = \delta(t \in d)$$

$$tf''(t, d) = \frac{\text{frequency of word}}{\text{frequency of most common word in document}}$$

Both of these representations are invariant to the document length, compared to the regular definition which assigns higher weights to longer documents.
Finding relevant terms

How to use TF-IDF

- Frequently occurring words have little impact on the similarity
- The similarity is now determined by the words that are most "characteristic" of the document
Finding relevant terms

But what about when we’re **weighting** the parameters anyway?

e.g. is:

$$\text{rating} \approx \alpha + \sum_{w \in \text{text}} \text{count}(w) \cdot \theta_w$$

really any different from:

$$\text{rating} \approx \alpha + \sum_{w \in \text{text}} \text{tfidf}(w, d, D) \cdot \theta_w$$

after we fit parameters?
But what about when we’re weighting the parameters anyway?

Yes!

• The **relative** weights of features is different between documents, so the two representations are not the same (up to scale)

• When we regularize, the scale of the features matters – if some “unimportant” features are very large, then the model can overfit on them “for free”
Finding relevant terms

But what about when we're **weighting** the parameters anyway?
Finding relevant terms

But what about when we’re **weighting** the parameters anyway?
Etc.

Not today...

See Michael Collins & Regina Barzilay’s NLP mooc if you’re interested:
Further reading:

- Original TF-IDF paper (from 1972)
  “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”
  [http://goo.gl/1CLwUV](http://goo.gl/1CLwUV)
Dimensionality-reduction approaches to document representation
Dimensionality reduction

How can we find low-dimensional structure in documents?

What we would like:

87 of 102 people found the following review helpful

You keep what you kill, December 27, 2004
By Schtinky "Schtinky" (Washington State) - See all my reviews

This review is from: The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from Pitch Black to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to Pitch Black Fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats. So each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of “The Chronicles of Riddick”)
Singular-value decomposition

Recall
(from dimensionality reduction lectures)

\[ R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix} \]

(e.g.) matrix of ratings

(square roots of) eigenvalues of \( R R^T \)

eigenvectors of \( R R^T \)

eigenvectors of \( R^T R \)

\[ R = U \Sigma V^T \]
Singular-value decomposition

Taking the eigenvectors corresponding to the top-K eigenvalues is then the "best" rank-K approximation.

\[ R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & \cdots & 1 \\ 3 & 1 & \cdots & 3 \\ 2 & 2 & \cdots & 4 \\ 1 & 5 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix} \]

\( R \approx U^{(k)} \Sigma^{(k)} V^{(k)T} \)

(square roots of top k) eigenvalues of \( RR^T \)

(top k) eigenvectors of \( RR^T \)

(top k) eigenvectors of \( R^T R \)
### Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

The matrix $X$ is a $T \times D$ matrix whose columns are bag-of-words representations of our documents.

$$X = \begin{pmatrix}
1 & 0 & \cdots & 4 \\
0 & 2 & & 0 \\
31 & 23 & & 97 \\
0 & 98 & & 1 \\
473 & 88 & & 347 \\
\vdots & \vdots & \ddots & \vdots \\
11 & 34 & \cdots & 13
\end{pmatrix}$$

- **$T$** = dictionary size
- **$D$** = number of documents
Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

\[ X^T X \] is a \( D \times D \) matrix.

\( U^{(k)} \Sigma^{(k)} \) is a low-rank approximation of each document eigenvectors of \( X^T X \)

\[ XX^T \] is a \( T \times T \) matrix.

\( V^{(k)} \Sigma^{(k)} \) is a low-rank approximation of each term eigenvectors of \( XX^T \)
Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?
Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?
Using our low rank representation of each document we can...

- Compare two documents by their low dimensional representations (e.g. by cosine similarity)
- To retrieve a document (by first projecting the query into the low-dimensional document space)
- Cluster similar documents according to their low-dimensional representations
- Use the low-dimensional representation as features for some other prediction task
Using our low rank representation of each word we can...

- Identify potential synonyms – if two words have similar low-dimensional representations then they should have similar “roles” in documents and are potentially synonyms of each other.
- This idea can even be applied across languages, where similar terms in different languages ought to have similar representations in parallel corpora of translated documents.
This approach is called **latent semantic analysis**

- In practice, computing eigenvectors for matrices of the sizes in question is not practical – neither for $XX^T$ nor $X^TX$ (they won’t even fit in memory!)
- Instead one needs to resort to some approximation of the SVD, e.g. a method based on stochastic gradient descent that never requires us to compute $XX^T$ or $X^TX$ directly (much as we did when approximating rating matrices with low-rank terms)
This approach is called **latent semantic analysis**

\[ \text{obj} \quad \| B_{ow} \cdot x_t \cdot y_d \| \]

\[ \text{obj} \quad x_t \quad \ldots \ldots \]

Singular-value decomposition
Today...

Using **text** to solve predictive tasks

- Representing documents using bags-of-words and TF-IDF weighted vectors
- Stemming & stopwords
- Sentiment analysis and classification

**Dimensionality reduction approaches:**

- Latent Semantic Analysis
Further reading:

• Latent semantic analysis
  http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

• LDA
  “Latent Dirichlet Allocation” (Blei, Ng, & Jordan, 2003)
  http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

• Plate notation
  http://en.wikipedia.org/wiki/Plate_notation
  “Operations for Learning with Graphical Models” (Buntine, 1994)
CSE 190 – Lecture 11
Data Mining and Predictive Analytics

Case study – opinion dimensions in beer
A (quick) case study

How can we estimate which words in a review refer to which sensory aspects?

‘Partridge in a Pear Tree’, brewed by ‘The Bruery’

Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.

Feel: 4.5 Look: 4 Smell: 4.5 Taste: 4 Overall: 4
Aspects of opinions

There are lots of settings in which people’s opinions cover many dimensions:

Wikipedia pages:

Cigars:

Beers:

Audiobooks:

Hotels:
Aspects of opinions

Further reading on this problem:

• Brody & Elhadad
  “An unsupervised aspect-sentiment model for online reviews”
• Gupta, Di Fabbrizio, & Haffner
  “Capturing the stars: predicting ratings for service and product reviews”
• Ganu, Elhadad, & Marian
  “Beyond the stars: Improving rating predictions using review text content”
• Lu, Ott, Cardie, & Tsou
  “Multi-aspect sentiment analysis with topic models”
• Rao & Ravichandran
  “Semi-supervised polarity lexicon induction”
• Titov & McDonald
  “A joint model of text and aspect ratings for sentiment summarization”
Aspects of opinions

If we can uncover these dimensions, we might be able to:

- Build sentiment models for each of the different aspects
- Summarize opinions according to each of the sensory aspects
- Predict the multiple dimensions of ratings from the text alone

- But also: **understand** the types of positive and negative language that people use
Aspects of opinions

Task: given (multidimensional) ratings and plain-text reviews, predict which sentences in the review refer to which aspect

Input:

‘Partridge in a Pear Tree’, brewed by ‘The Bruery’

Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.

Feel: 4.5 Look: 4 Smell: 4.5 Taste: 4 Overall: 4

Output:

‘Partridge in a Pear Tree’, brewed by ‘The Bruery’

Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.

Feel: 4.5 Look: 4 Smell: 4.5 Taste: 4 Overall: 4
Aspects of opinions

Solving this problem depends on solving the following two sub-problems:

1. Labeling the sentences is easy if we have a good model of the words used to describe each aspect
2. Building a model of the different aspects is easy if we have labels for each sentence

• **Challenge:** each of these subproblems depends on having a good solution to the other one
• So (as usual) start the model somewhere and alternately solve the subproblems until convergence
Aspects of opinions

Model:

\[ P(\text{aspect}(s) = k | \text{sentence } s, \text{rating } v) = \frac{1}{Z} \exp \sum_{w \in s} \left\{ \theta_{k,w} + \phi_{k,v_k,w} \right\} \]

- **Normalization over all aspects**
- **Sum over words in the sentence**
- **Weight for a word \( w \) appearing in a particular aspect \( k \)**
- **Weight for a word \( w \) appearing in a particular aspect \( k \), when the rating is \( v_k \)**
Aspects of opinions

Intuition:

\[ P(\text{aspect}(s) = k | \text{sentence } s, \text{rating } v) = \]

\[ \frac{1}{Z} \exp \sum_{w \in s} \left\{ \theta_{k,w} + \phi_{k,v_k,w} \right\} \]

**Nouns** should have high weights, since they describe an aspect but are independent of the sentiment.

**Adjectives** should have high weights, since they describe specific sentiments.
Aspects of opinions

Procedure:

1. Given the current model (theta and phi), choose the most likely aspect labels for each sentence

\[ \max_{\text{aspect labels for each sentence}} P_{\theta, \phi}(\text{aspect}(s) = k | \text{sentence } s, \text{rating } v) \]

2. Given the current aspect labels, estimate the parameters theta and phi (convex problem)

\[ \max_{\theta, \phi} P_{\theta, \phi}(\text{aspect}(s) = k | \text{sentence } s, \text{rating } v) \]

3. Iterate until convergence (i.e., until aspect labels don’t change)
Aspects of opinions

Evaluation:
In order to tell if this is working, we need to get some humans to label some sentences

- I labeled 100 sentences for validation, and sent 10,000 sentences to Amazon’s “mechanical turk”
  - These were next-to-useless
- So we hired some “experts” to label beer sentences

![Network diagram showing 30% agreement between me, turkers, and oDesk, with 90% agreement between me and oDesk, and 30% agreement between turkers and oDesk labeled “beer experts”)]
Aspects of opinions

Evaluation:

• 70-80% accurate at labeling beer sentences (somewhat less accurate for other review datasets)
• A few other tasks too, e.g. summarization (selecting sentences that describe different opinions on a particular aspect), and missing rating completion
## Aspects of opinions

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Aspect words $\theta_k$</th>
<th>Sentiment words (2-star) $\phi_{k,2}$</th>
<th>Sentiment words (5-star) $\phi_{k,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feel</td>
<td>light, carbonation, bottle, little</td>
<td>thin, watery</td>
<td>perfect, amazing, silky, velvety</td>
</tr>
<tr>
<td>Look</td>
<td>head, finish, nice, like</td>
<td>yellow, beautiful</td>
<td>gorgeous, pitch, huge, perfect</td>
</tr>
<tr>
<td>Smell</td>
<td>caramel, hops, malt, notes, fruit</td>
<td>corn, macro</td>
<td>amazing, awesome, incredible, exceptional, absolutely, wow</td>
</tr>
<tr>
<td>Taste</td>
<td>malt, flavor, sweet, dark, flavor, bitter</td>
<td>watery, water, bland</td>
<td>perfect, delicious, perfect, absolutely, awesome</td>
</tr>
<tr>
<td>Overall impression</td>
<td>beer, one, well, glass, drain, worst, disappointment, water, ship, worst, boring, water, bland, water, worst, avoid</td>
<td>bland, disappointing, water, ship, worst, boring, water, bland, water, worst, avoid</td>
<td>amazing, simply, wonderful, every, incredible, every, incredibly, absolutely, awesome</td>
</tr>
</tbody>
</table>
Aspects of opinions

Moral of the story:

• We can obtain fairly accurate results just using a bag-of-words approach
• People use very different language if the have positive vs. negative opinions
• In particular, people don’t just take positive language and negate it, so modeling syntax (presumably?) wouldn’t help that much
Further reading:

• Linguistics of food

“The language of Food: A Linguist Reads the Menu”
http://www.amazon.com/The-Language-Food-Linguist-Reads/dp/0393240835
A few assignment 1 tips

Task 1

1. Helpful/ness

2. Keep test simple

3. user / id: root
cat
A few assignment 1 tips

Task 2

Purchase

1) Figure out deficiency

2) \[ \sigma(\alpha + \beta n + \beta i) + \delta \alpha \cdot x \]

3) Handle edge case
• The 2\textsuperscript{nd} assignment is also out
• I know nobody wants to start it just yet, but start forming a team, and I’ll talk about project ideas next week
• The 4\textsuperscript{th} homework is also out this week
• Hopefully you don’t have to do it! But by next week you should know where your grades stand for everything but Assignment 2, so will have an idea of whether you need the extra grades
• I’ll put graded HWs etc. outside my office (which seems to be easier than trying to sift through them at the beginning of class)