Text Analysis: Day 2

Rachel Porter
Odum Institute for Research in Social Science
rachsur@live.unc.edu

October 3, 2019

Outline

1. Collection
   - Web-scraping
   - Application Programming Interface
2. Processing
3. Analysis
   - Dictonary-based Approaches
   - Topic-based Approaches

A Quick Note

- https://www.datacamp.com/courses/free-introduction-to-r

What is a Dictionary-basd Approach?

- Dictionary-based methods find the total sentiment of a piece of text by adding up the individual sentiment scores for each word in the text.
- People often use a mixture of sentiment when they talk about topics, which one dominates?

For Example. . .

Topo offers a beautiful view of Franklin Street that's perfect for people watching in the afternoon. During the day, it's a casual spot to grab dinner with family or friends and the food is pretty decent tasting but can be pricey for a predominately college aged crowd. The drinks are always pretty good and they have a good selection of beers on tap with a selection of their own vodka that they distill.

During the night, this place turns into your typical college free for all with a DJ, dancefloor, loud music and college kids packed in like sardines. The outdoor patio offers a nice relief from it all when you need a water break from dancing. It's a pretty decent spot, a fun nighttime bar and a staple in Chapel Hill you have to visit at least once just to say that you did.

Corey C. voted for this review

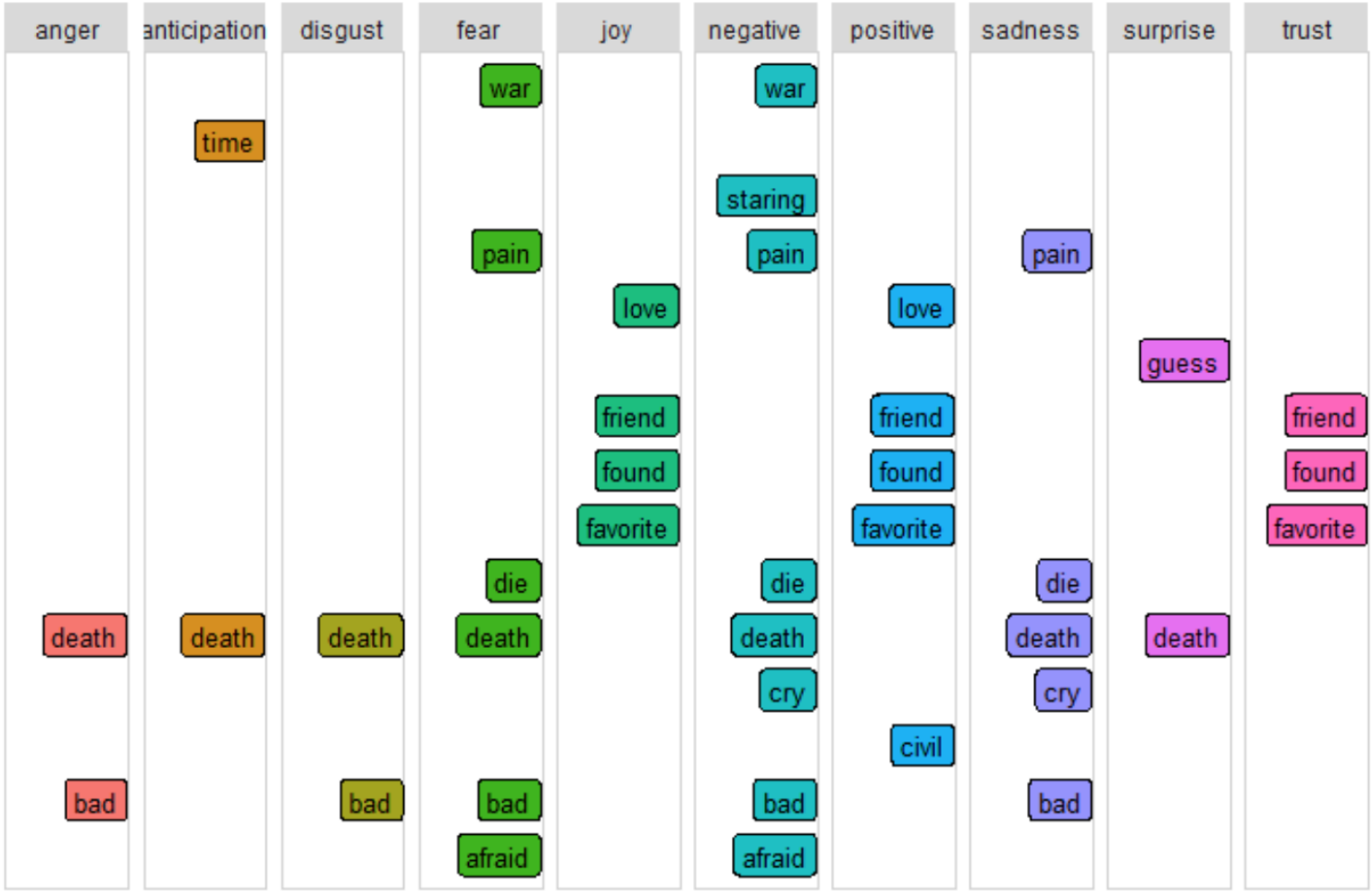
Useful 1 | Funny | Cool

I came here today during their afternoon menu hours, which are 3:00-5:00. The limited menu is definitely smaller than the lunch and dinner menus, but overall had a fair number of choices. Unfortunately, none of them particularly stood out to me, but I know I can be picky. I ended up ordering the Bavarian pretzels with beer cheese, which were okay. The pretzels tasted fine, but were stale and possibly reheated. The cheese had already started to congeal by the time it was brought to the table. It tasted fine, but nothing special. They were quite busy, so the service was alright given the number of people who were there. A great bonus is the location and the patio seating. It's on the top floor of a building right on the corner of downtown Chapel Hill, and the view was beautiful today. They also have covered outdoor seating for those rainy days. I'll have to come back during dinner some time because I've heard that the short ribs are quite good, and I've been wanting to try them. Overall, it's a great spot to hang out. They're known for their bar and beer selection, since they have their own brewery. It does, however, get very crowded with college students on weekends and during events.

Nicole L. voted for this review

Useful 1 | Funny 1 | Cool 1

Another Example. . .



Sentiment Analysis of "Sometimes It Snows In April"

Sentiment Analysis

- Sentiment analysis is a type of text mining which aims to determine the opinion and subjectivity of its content.
- Also called opinion mining, detecting whether a span of text expresses some kind of judgement
  - positive vs. negative
  - happy vs. sad
  - liberal vs. conversation
  - in favor vs. against
- Sentiment is most often thought of as a dichotomous distinction type of analysis (positive vs. negative), but it can also be a more fine-grained (for instance, pinpointing a specific emotion)

Steps for Sentiment Analysis

- Create or find a list of words associated a particular sentiment (ex. liberal and conservative)
- Count the number of liberal and conservative words in a text
- Analyze the proportion of liberal to conservative words.
  - More liberal words indicates a liberal sentiment, more conservative words indicates a conservative sentiment

Tools for Measuring Sentiment

- **Lexicon:** the word list used to define a particular sentiment; think of it as a dictionary of terms
- Personal dictionary

```
taxes <- c("budget", "spending", "defecit", "pork")
healthcare <- c("insurance", "premium", "universal")
```

- Ready-made dictionaries
  - University of Pittsburg (http://mpqa.cs.pitt.edu/)
  - Liu and Hu lexicon (https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon)
  - Gold standard: Liguisitic Inquiry and Word Count (LIWC)

Challenges

- Word strength (liked vs. loved)
- Negation (liked vs. didn't like)
- Mediating language (could have liked, would have liked)
- Topic-specific text features
- Sarcasm

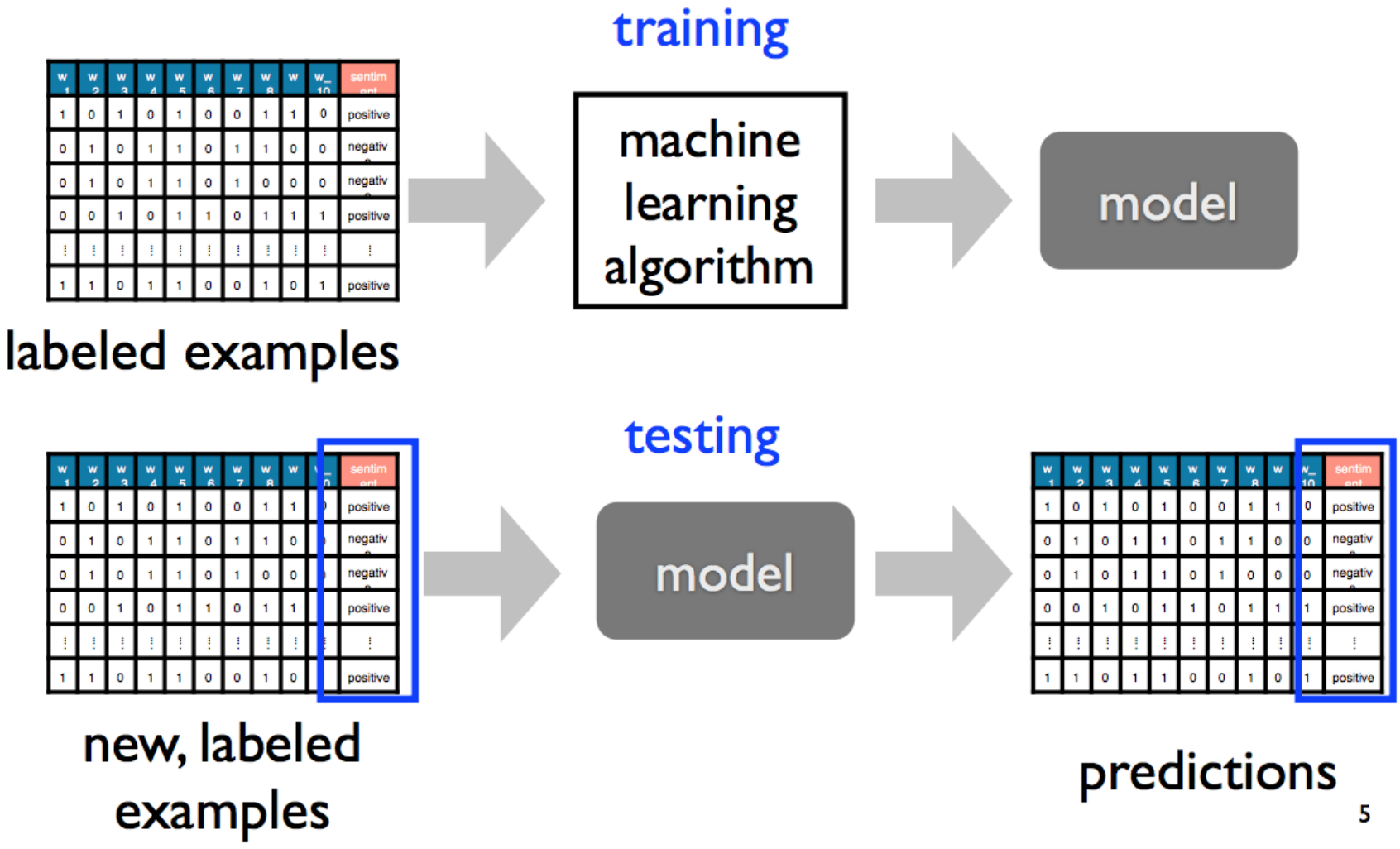
Predictive Analysis

- We can also use statistical models to make predictions on previously unseen data
- For instance, if we have text and we want to know its sentiment, we can use a statisical model to evaluate the text and make determinations about the sentiment without reading the text
- This is **predictive analysis** and has many real world applications
  - The recommended movies on your Netflix account
  - The mail in your junk folder

Predictive Analysis

- To produce a predictive model that does what we want it to, we first need to evaluate the accuracy of that model — how often it correctly predicts the sentiment of a given text
  - In order to do this, we first need data that is labeled (ie. we know the sentiment)
  - Next, we can produce a model using some of that labeled data — this is called the **training** set
  - To evaluate our model, we can run it on the rest of our labeled data — this is called the **test set** — and check its accuracy

Predictive Analysis in Pictures



What does evaluating accuracy look like?

		true	
		pos	neg
predicted	pos	a	b
	neg	c	d

$$\mathcal{A} = \frac{(a+d)}{(a+b+c+d)}$$

Some Important Notes

- Why do we need a separate training and test set?
- If we train a model on a **training** set, can we include these data in our **analysis**?

Some (More) Important Notes

- Accuracy is different than...
  - **Precision:** the percentage of positive predictions that are truly positive
  - **Recall:** the percentage of true positives that are correctly predicted positive

Alternatives to Sentiment Analysis

- **Topic Modeling:** groups of words together—instead of counting them individually–in order to capture how the meaning of words is dependent upon the broader context in which they are used in natural language.
  - Each document can be made up of a number of topics
  - We don't need to define the topics! The model does that for us
  - Great alternative to reading and coding thousands of documents
  - R Package of Interest: stm
- **Plagirism Analysis:** evaluate text against another corpus to determine the proportion of similiarity between the two texts
  - Resources of Interest:
    - https://cran.r-project.org/web/packages/SimilaR/index.html
    - https://www.r-bloggers.com/similar/
    - https://github.com/ropensci/textreuse

Lab Exercise

Thanks to Brice Acree, Jaime Arguello, and Chris Bail for making their materials available. Portions of this presentation draw on their resources.