STAT 474 – Techniques for Large Data Sets

Fall 2018

November 5, 2018

Text mining - Review

- Collection of methods and techniques to extract, summarize, and analyze useful information from text data

Text mining - Review

- Collection of methods and techniques to extract, summarize, and analyze useful information from text data
- Text mining project:
  - Text preprocessing: cleaning up, tokenization, normalization, etc.
  - Feature generation, e.g., word frequencies
  - Data mining. e.g., classification, regression

Text mining - Review

- Collection of methods and techniques to extract, summarize, and analyze useful information from text data
- Text mining project:
  - Text preprocessing: cleaning up, tokenization, normalization, etc.
  - Feature generation, e.g., word frequencies
  - Data mining. e.g., classification, regression
- **What else have we done?**

What is Sentiment?

- Sentiment = feelings
  - Also means attitudes, emotions, opinions
  - Sentiments are subjective impressions, not facts.

What is Sentiment?

- Sentiment = feelings
  - Also means attitudes, emotions, opinions
  - Sentiments are subjective impressions, not facts.
- When working with sentiments, a binary opposition in opinions is assumed
  - For/against, like/dislike, good/bad, etc.
  - A few projects work with other types of emotions, but not common.

What is Sentiment Analysis?

- Using NLP (natural language process), statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit
- Sometimes referred to as **opinion mining**, although the emphasis in this case is on extraction

What is Sentiment Analysis?

- Using NLP (natural language process), statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit
- Sometimes referred to as **opinion mining**, although the emphasis in this case is on extraction
- Questions sentiment analysis might ask:
  - Is this product review positive or negative?
  - Is this customer email satisfied or dissatisfied?
  - How have bloggers' attitudes about the president changed since the election?

Other related taskes for sentiment analysis

- Information extraction (discarding subjective information)
- Question answering (recognizing opinion-oriented questions)
- Summarization (accounting for multiple viewpoints)
- "Flame"/Insulting detection
- Identifying child-suitability of videos based on comments
- Bias identification in news sources

An application in Business Analytics

- Question: "Why aren't consumers buying our laptop?"
- We know the concrete data: price, specs, competition, etc.

Text

An application in Business Analytics

- Question: "Why aren't consumers buying our laptop?"
- We know the concrete data: price, specs, competition, etc.
- We want to know subjective data: "the design is tacky", "customer service was horrible"
- Misperceptions are also important, e.g., "updated drivers aren't available" (actually, they are!!)

An application in Business Analytics

- Question: "Why aren't consumers buying our laptop?"
- We know the concrete data: price, specs, competition, etc.
- We want to know subjective data: "the design is tacky", "customer service was horrible"
- Misperceptions are also important, e.g., "updated drivers aren't available" (actually, they are!!)
- It is very difficult to survey customers who **didn't** by the company's laptop
- Instead, you could use sentiment analysis to
    - search the web for opinions and reviews of this and competing laptops: blogs, amazon, tweets, etc.
    - create condensed versions or a digest of consensus points.

Challenges in sentiment analysis

- People express opinions in complex ways.
- In opinion texts, lexical content alone can be misleading
- Intra-textual and sub-sentential reversals, negation, topic change common
- Rhetorical devices/modes such as sarcasm, irony, implication, etc.

Example - a fictitious letter to a hardward store

> “Dear <hardware store>
>
> Yesterday I had occasion to visit <your competitor>. They had an excellent selection, friendly and helpful salespeople, and the lowest prices in town.
>
> You guys suck.
>
> Sincerely,”

Example - a fictitious letter to a hardward store

> “Dear <hardware store>
>
> Yesterday I had occasion to visit <your competitor>. They had an excellent selection, friendly and helpful salespeople, and the lowest prices in town.
>
> You guys suck.
>
> Sincerely,”

- Humans are subjective creatures and opinions are important. Being able to interact with people on that level has many advantages.

Where to start

- Decide on the building blocks of sentiment expression.
    - Many possibilities: users, texts, sentences, words, tweets/updates

Where to start

- Decide on the building blocks of sentiment expression.
  - Many possibilities: users, texts, sentences, words, tweets/updates
- It is common to use words as building blocks
  - Short phrases may be just as important (or moreso) as words, e.g., "lowest prices", "high quality"

Where to start

- Decide on the building blocks of sentiment expression.
  - Many possibilities: users, texts, sentences, words, tweets/updates
- It is common to use words as building blocks
  - Short phrases may be just as important (or moreso) as words, e.g., "lowest prices", "high quality"
- There seems to be **some** relation between positive words and positive sentiments
  - Can we come up with a set of keywords to identify sentiments?

Keyword methods

- By hand: labor intensive. Amazon's Mechanical Turk for NLP annotation
  - Roughly $1 for 1,000 labels
  - 5 non-expert annotators achieve equivalent accuracy to 1 expert annotator

Keyword methods

- By hand: labor intensive. Amazon's Mechanical Turk for NLP annotation
  - Roughly $1 for 1,000 labels
  - 5 non-expert annotators achieve equivalent accuracy to 1 expert annotator
- Data-driven methods can be used to generate keyword lists
  - model better than human-generated lists (might due to the volume)
  - Comments, tweets with emoticons/smileys
  - Reviews (Amazon, Yelp) with product/item ratings.

Keyword methods

- By hand: labor intensive. Amazon's Mechanical Turk for NLP annotation
  - Roughly $1 for 1,000 labels
  - 5 non-expert annotators achieve equivalent accuracy to 1 expert annotator
- Data-driven methods can be used to generate keyword lists
  - model better than human-generated lists (might due to the volume)
  - Comments, tweets with emoticons/smileys
  - Reviews (Amazon, Yelp) with product/item ratings.
- Sentiment-oriented data set is highly domain sensitive and very difficult to create/collect.

Availability of sentiment-oriented data sets

- Many linguists open their databases for public use.
  - `tidytext` package has 4 different lexicons for sentiment analysis
  - **What are their differences?**

Availability of sentiment-oriented data sets

- Many linguists open their databases for public use.
  - `tidytext` package has 4 different lexicons for sentiment analysis
  - **What are their differences?**

The three general-purpose lexicons:

- `AFINN` from Finn Arup Nielsen: with scores from -5 to 5.
- `bing` from Bing Liu and collaborators: with positive/negative annotation.
- `nrc` from Saif Mohammad and Peter Turney: many different types of emotions

Demo on Sentiment Analysis

- Examine how the sentiment changes across the novels.
- The narrative arc is positive if the total "score" is positive, and is negative if the total "score" is negative
    - Difference in counts of positive and negative words (using `bing`)
    - Total sum of sentiment score when using `affin`
- Define a book section as 80 lines of text
    - Too small sections might not have enough words (not all words have sentiment scores)
    - Too long sections might wash out narrative structure.

Issues with keyword methods

- Words may not be enough
  - *ridiculous comedy* versus *ridiculus drama*
  - *cheap construction* versus *cheap deal*

Issues with keyword methods

- Words may not be enough
    - *ridiculous comedy* versus *ridiculus drama*
    - *cheap construction* versus *cheap deal*
- We might want to assign sentiment scores to certain kinds of phrases, e.g "high quality", etc.
- Binary sentiment don't capture nuance

Issues with keyword methods

- Words may not be enough
  - *ridiculous comedy* versus *ridiculus drama*
  - *cheap construction* versus *cheap deal*
- We might want to assign sentiment scores to certain kinds of phrases, e.g "high quality", etc.
- Binary sentiment don't capture nuance
- Negations (in general **valence shifters**) changes the sentiment score of a word
  - Example: "I am not very happy"

Looking at units beyond just words

- Some sentiment analysis algorithms look beyond only unigrams (i.e. single words) to try to understand the sentiment of a sentence as a whole.
- Token can be sentences or even chapters.
    - It is harder to clean up the text.

Looking at units beyond just words

- Some sentiment analysis algorithms look beyond only unigrams (i.e. single words) to try to understand the sentiment of a sentence as a whole.
- Token can be sentences or even chapters.
    - It is harder to clean up the text.
- R packages: `coreNLP`, `cleanNLP`, `sentimentr`, etc.
- Example of `sentimentr`

```
sentiment('I am not very happy')
```

```
##    element_id sentence_id word_count   sentiment
## 1:          1           1          5 -0.06708204
```

- It takes more computational power (and time) to analyse with bigger token structure.