Dictionaries

Measuring Known Categories in Text

What is a Dictionary?

What is a Dictionary?

A list of words associated with one or more categories.

Words can also be weighted.

Financial Sentiment Dictionary

SCALE	NO. OF WORDS	SAMPLE WORDS
Negative	2,337	termination, discontinued, penalties, misconduct, serious, noncompliance, deterioration, felony
Positive	353	achieve, attain, efficient, improve, profitable
Uncertainty	285	approximate, contingency, depend, fluctuate, indefinite, uncertain, variability
Litigiousness	731	claimant, deposition, interlocutory, testimony, tort
Weak Modal Words	27	could, depending, might, possibly
Strong Modal Words	19	always, highest, must, will

Use: Provide a score for each text

An “uncertainty score” might be:

- a boolean for whether any word in the category occurred
- the total number of times any word from the category occurs
- the percent of total words that are in the uncertainty category
- a relative score compared to other texts in the corpus
- a ranked score compared to other texts in the corpus

Types of Dictionaries

- Expert-created
  - Proprietary
    - LIWC (psychology)
    - Diction (political science
  - Free
    - Harvard General Inquirer
    - MPQA
- Manually-created
- Crowdsourced
  - NRC Word-Emotion Association Lexicon
  - Computational Story Lab (University of Vermont)

Uses

- Identify prevalence of a particular theme (e.g. populism, inequality)
- Measuring affect (sentiment analysis)
- Political tone
- Categorize text
- Examples:
  - Happiest states in the U.S.
  - Happiness and days of the week
  - How to get re-tweeted

Pitfalls (there are many!)

- The same word can be used many different ways and contexts

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time
  - Awesome

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time
  - Awesome
- Sarcasm, irony, humor

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time
  - Awesome
- Sarcasm, irony, humor
- Standard dictionaries do not translate to other domains

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time
  - Awesome
- Sarcasm, irony, humor
- Standard dictionaries do not translate to other domains
- Custom dictionaries are difficult to validate

Pitfalls (there are many!)

- The same word can be used many different ways and contexts
  - Crude
  - Unanticipated
  - Combine with part-of-speech, etc., still not perfect
- Change over time
  - Awesome
- Sarcasm, irony, humor
- Standard dictionaries do not translate to other domains
- Custom dictionaries are difficult to validate
- Statistical significance