Reading Week 02 - Dictionary methods

Add your name here

May 20, 2021

Abstract

This week we covered corpus-based and dictionary-based methods for analyzing text. This week's reading will cover methods of creating commonly used dictionaries section 2 and interesting applications of some of these dictionary-based methods section 3.

1 Instructions

For each reading, include the following two paragraphs:

- Summary: 3-4 sentence summary
- One paragraph with at least the following:
  - 1 sentence about something in particular that you like
  - 1 sentence about something you didn't like or something you found confusing and you'd like me to explain
  - 1 question for future work or one sentence on how you'd apply the reading for a final project

2 Dictionary Methods

Please read **one** of the following papers that describe a commonly used dictionary method:

1. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods [9] https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf
2. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text [4] http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

VADER Review

3 Application of Dictionary Methods

Please read **two** of the following papers. These papers come from a wide range of fields, e.g. psychology, economics, political science, and public health.

1. What can software tell us about political candidates?: A critical analysis of a computerized method for political discourse [5] http://coms2710.barnard.edu/readings/Kangas2014Politics.pdf
2. Measuring the happiness of large-scale written expression: Songs, Blogs, and Presidents. [3] https://arxiv.org/pdf/1703.09774.pdf
3. Quantifying the Effects of COVID-19 on Mental Health Support Forums [2] https://www.aclweb.org/anthology/2020.nlpcovid19-2.8.pdf
4. Language use of depressed and depression-vulnerable college students [7] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4752&rep=rep1&type=pdf

5. Word use in the poetry of suicidal and nonsuicidal poets [8] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.544.1791&rep=rep1&type=pdf

6. Measuring Economic Policy Uncertainty Market [1] https://academic.oup.com/qje/article/131/4/1593/2468873

7. Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks [6] https://journals.sagepub.com/doi/pdf/10.1177/2332858420940312

Word Use in Poetry Review

Stirman and Pennebaker attempt to understand psychological differences between suicidal and non-suicidal poets by applying a CTA lens to viewing their bodies of work. The authors use sociological theory (heavily relying on the Durkheim) conceptualize what elements would embody an artist's shift towards more self-isolating/self-harming tendencies. Using Durkheim's "social integration/disengagement model" the authors analyzed corpora of multiple authors , each author having three corpora that spanned the early, middle, and late parts of their careers (2). Stirman et. al. used LIWC as their framework/method of analysis, and they concluded from their work that while there was a statistical difference in pronoun usage (larger use of first person singular), other extrapolations (movement away from first person plural in late works) were not statistically valid.

I enjoyed the disciplinary application of the text analysis and the integration of sociological theory into the use of CTA - it seemed to provide a critical framework for the larger interpretation of the analysis's results. I didn't enjoy the possible ethical implications of the analyzing art for psychological maladies via text since it could lead to art as delegitimizing (i.e. Nazis and other fascist groups used to make the argument that modern art was 'degenerate' and made by 'sick' people). A question that arises for my own project is to what extent do I want to integrate established theory into my hypothesis/analysis.

Measuring Happiness Review

Dodds and Danforth forefront questions surrounding happiness and the ability to derive this latent information from the text in a computational manner. Using the ANEW data set, the authors analyze corpora of Michael Jackson lyrics, blog posts, and State of the Union addresses, and they attempt to analyze and visualize the features (mainly word valence shift) using a variety of different parameters (day of the week, age, latitude, etc.). The authors "compare individual word prevalence changes... in what [the authors] term a 'Valence Shift Word Graph'" (5). This new method of calculating and visualizing the change in a word's usage over a period of time allows for researchers to understand new or previously unseen relations between word usage and concurrent temporal realities. Since the authors analyzed a diverse array of textual genres, they come up with a interesting mix of interpretations asserting that "we also take the viewpoint that human assessment remains superior" with regards to the outcome of their results.

I enjoyed seeing how the authors attempts to exploit and explore the wide array of information that is available in the world. Rather than being myopic about a specific phenomena in the world, they explored a theoretical framework that could be applied largely, and their experimentation seems to result in more lively and interesting outcomes - albeit, at the expense of any significant, statistically verified results. In this, I found it interesting what phenomena they were attempting to explore, but I did feel like there paper could get a little overwhelming from all the different ways they were trying to explore this idea and how they jump from corpus to corpus. In relating this to my project, I want to think about how myopic I want to be with my text sourcing and general analysis.

VADER Sentiment Review This article was fascinating in how statistical and linguistic methodologies have advanced so far as to create a largely general-use model for sentiment analysis – this would have been largely unthinkable even 20 years earlier. Hutto and Gilbert explain the origins of the LIWC lexicon, which had to be manual compiled, and its limitations in being able to deal with a variety of new and emerging genre of writing (but also many old genres, for that matter). Manually compiled sentiment lexicons are important, but the time and effort it takes to compile and validate them makes them largely impractical (especially since they cannot deal with important linguistic features such as emoticons, slang, etc.). The authors explain their semi-supervised method of compiling the lexicon via Amazon's Mechanical Turk, verification systems, incentives, crowd-sourced evaluations, and experimental validation, and Hutto et. al. really showed their brilliance when it came to devising a methodology which protect and furthered its own integrity.

I enjoyed reading about this major shift in lexicon production and validation methods, as well as how VADER (a valence-based lexicon) has become a model for how sentiment analysis lexicons can be compiled more efficiently and accurately. I felt a little uncomfortable with the proposed concept that a computer could be more accurate in interpreting human speech, and what the implications for such a paradigm could be. In relation to my project, I want to think more about what statistics and variables I want to attempt to extract, and what models, lexicons, etc. would be the best use for my specific corpus and features that I wish to extract.

Note for this Week's Readings Many of these papers include other methods that we have not covered yet, e.g. topic modeling and word embeddings. It is okay if you do not understand some of those parts of the papers. The focus of this week's reading is the use case of dictionary-based methods. Please still read those other sections of these papers. We will cover more of these methods through out the course and you might want to employ them for your final projects.

References

[1] Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

[2] Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. Quantifying the effects of COVID-19 on mental health support forums. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics.

[3] Peter Sheridan Dodds and Christopher M Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, 11(4):441–456, 2010.

[4] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[5] Sara Kangas. What can software tell us about political candidates?: A critical analysis of a computerized method for political discourse. *Journal of Language and Politics*, 13, 05 2014.

[6] Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312, 2020.

[7] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[8] Shannon Wiltsey Stirman and James W Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522, 2001.

[9] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.