Imagisaurus: An Interactive Visualizer of Valence and Emotion in the Roget's Thesaurus

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

The form of a thesaurus often restricts its use to word look ups and finding related words. We present *Imagisaurus*, an online interactive visualizer for the *Roget's Thesaurus*, which not only provides a way for word lookups but also helps users quickly grasp the nature and size of the thesaurus taxonomy. Imagisaurus connects thesaurus entries with a large valence and emotion association lexicon. Easy-to-use sliders give the user fine control over depicting only those categories with the desired strength of association with positive or negative sentiment, as well as eight basic emotions. A second interactive visualization is used to explore the emotion lexicon. Both the *Roget's Thesaurus* and the emotion lexicon have tens of thousands of entries. Our visualizers help users better understand these lexical resources in terms of their make up as a whole.

1 Introduction

The *Roget's Thesaurus* (Roget, 1852) was created by Peter Roget in 1852 and originally included about 15,000 English words. Since then a number of newer versions of the thesaurus have been published, and each has included more terms than the previous version. *Roget's* taxonomic structure, which is inspired by philosophical work of Leibniz on symbolic thought (Leibniz and Parkinson, 1995; Leibniz, 1923), groups words into six *classes*: words expressing abstract relations, words relating to space, words relating to matter, words relating to the intellectual faculties, words relating to the voluntary powers, and words relating to the sentient and moral powers. These six classes are further partitioned into thirty nine *sections*, which are in-turn divided into one thousand *categories*. Each category lists about 20 to 200 related words and expressions. These categories can be thought of as coarse concepts.

Widely used by writers, lexicographers, students, and the lay person, the thesaurus is most commonly accessed to identify a word or phrase that best captures what one wants to communicate. Researchers in many fields find use for the thesaurus, for example those exploring literary, social science, psychological, and cognitive theories involving word usage. Not surprisingly, there is a vast and growing body of work in Computational Linguistics that makes use of the *Roget's Thesaurus*, including Masterman (1957), Morris and Hirst (1991), Yarowsky (1992), Mohammad and Hirst (2006), Mohammad (2008), and Grefenstette (2012). However, despite its substantial range and scope of use, manual access to information in the thesaurus is often restricted to looking up a word and finding its neighbors. Existing online portals for the *Roget's Thesaurus* present a very traditional, non-interactive, text-only interface.[1]

We present an online interactive visualizer for the *Roget's Thesaurus*, which we call *Imagisaurus*.[2] Imagisaurus allows users to access information about words, classes, sections, and categories through four separate sub-visualizations that are linked to each other. Clicking on a unit selects it and filters information in all other sub-visualizations, showing information that is relevant only to the selection. The hierarchical structure of the thesaurus is shown in proportion to the size of its components—where size is defined to be the number of words included in the thesaurus unit (category, section, etc.). This allows users to determine which thesaurus units are more populous. Additionally, Imagisaurus links the *Roget's*

[1] http://www.roget.org
http://machaut.uchicago.edu/rogets

[2] Imagisaurus: http://www.purl.com/net/imagisaurus
Imagisaurus currently uses the copyright-free Project Gutenberg version of the thesaurus (Roget, 1911).

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 85–91,
Lisboa, Portugal, 17 September, 2015.

Thesaurus with a large emotion lexicon and lets users interactively discover categories strongly associated with various affect categories: positive and negative valence (sentiment), as well as emotions of joy, sadness, fear, trust, disgust, anticipation, anger, and surprise. Easy-to-use sliders give the user fine control over depicting only those categories with the desired strength of association with an affect category.

Word–Affect association lexicons, such as the NRC Emotion Lexicon, are themselves large semantic resources used not only by computational linguists, but also by researchers in psychology, marketing, advertising, and public health. Thus we developed a second online interactive visualization for the NRC Emotion Lexicon.

Both the *Roget's Thesaurus* and the emotion lexicon have tens of thousands of entries. Thus obtaining a feel for them by manually reading every entry is prohibitive. Our visualizers, created using the visualization tool Tableau, help users better understand these lexical resources in terms of their make up as a whole.[3] Both visualizers are made available online and are free to use.[4]

2 Affect Associations

Many words such as *good* and *delighted* express affectual states such as positive sentiment, negative sentiment, joy, anger, and so on. Apart from literal, denotative, meaning, words also have *associations* with sentimental, emotional, cultural, and social overtones. For example, *skinny* and *slender* primarily convey information about girth, but additionally *skinny* is associated with a slight negative sentiment, whereas *slender* is associated with positive sentiment. Similarly, *party* is associated with joy whereas *test results* is associated with anticipation.[5] The *Roget's Thesaurus* groups related terms within the same category, and this means that a category can include terms associated with many affect categories.

The thesaurus itself makes no claims on the affect associations of its constituent words (denotative or connotative). However, recently large resources have been created that capture the affect associations of thousands of words: The General Inquirer (GI) has sentiment labels for about 3,600 terms (Stone et al., 1966). Hu and Liu (2004) manually labeled about 6,800 words and used them for detecting sentiment of customer reviews. Affective Norms for English Words (ANEW) has pleasure (happy–unhappy), arousal (excited–calm), and dominance (controlled–in control) ratings for 1034 words.[6] The WordNet Affect Lexicon (WAL) (Strapparava and Valitutti, 2004,) has a few hundred words annotated with associations to the six Ekman emotions. The NRC Emotion Lexicon has association labels for over 14,000 words with positive and negative sentiment, as well as the set of eight Plutchik emotions (Mohammad and Turney, 2010; Mohammad and Turney, 2013).[7] These labels were compiled through crowdsourcing. Lexicons for word–affect associations are used in automatic classification systems as well as systems that track affectual words in text (for example in literary analysis and for assessing well-being in social media posts).

We use the NRC Emotion Lexicon in Imagisaurus because of its large coverage and associations with both sentiment and emotions. However, other affect lexicons can also be plugged into the same visualization design.

3 Imagisaurus

Figure 1 shows a screenshot of Imagisaurus. (The tooltip info box, which shows information about the taxonomic unit over which the mouse pointer is hovering, can be ignored for now.) Observe that there are four sub-visualizations: Index, Classes, Sections, and Categories. On the top right corner is a legend showing the colors in which the six thesaurus classes are shown. (The colors were chosen somewhat at random, the only requirement being that they be easily distinguishable.) Below the legend are ten sliders corresponding to affect densities of ten affect categories (two sentiments and eight emotions).

The Index shows the index of the thesaurus, that is, it lists all the words in the thesaurus in alphabetical order along with the categories they are included in. The hierarchical structure of the thesaurus, in terms of its classes, sections, and categories, is shown through the three treemap visualizations—one for each level of the hierar-

[3] http://www.tableau.com

[4] Imagisaurus: http://www.purl.com/net/imagisaurus
Emotion Lexicon Viz.: http://www.purl.com/net/EmoLexViz

[5] Some of these connotations may be cultural, for example, dating may be seen unfavorably in some cultures, however, many connotations add to the denotative meanings of words and are commonly known.

[6] http://csea.phhp.ufl.edu/media/anewmessage.html

[7] www.purl.com/net/NRCemotionlexicon

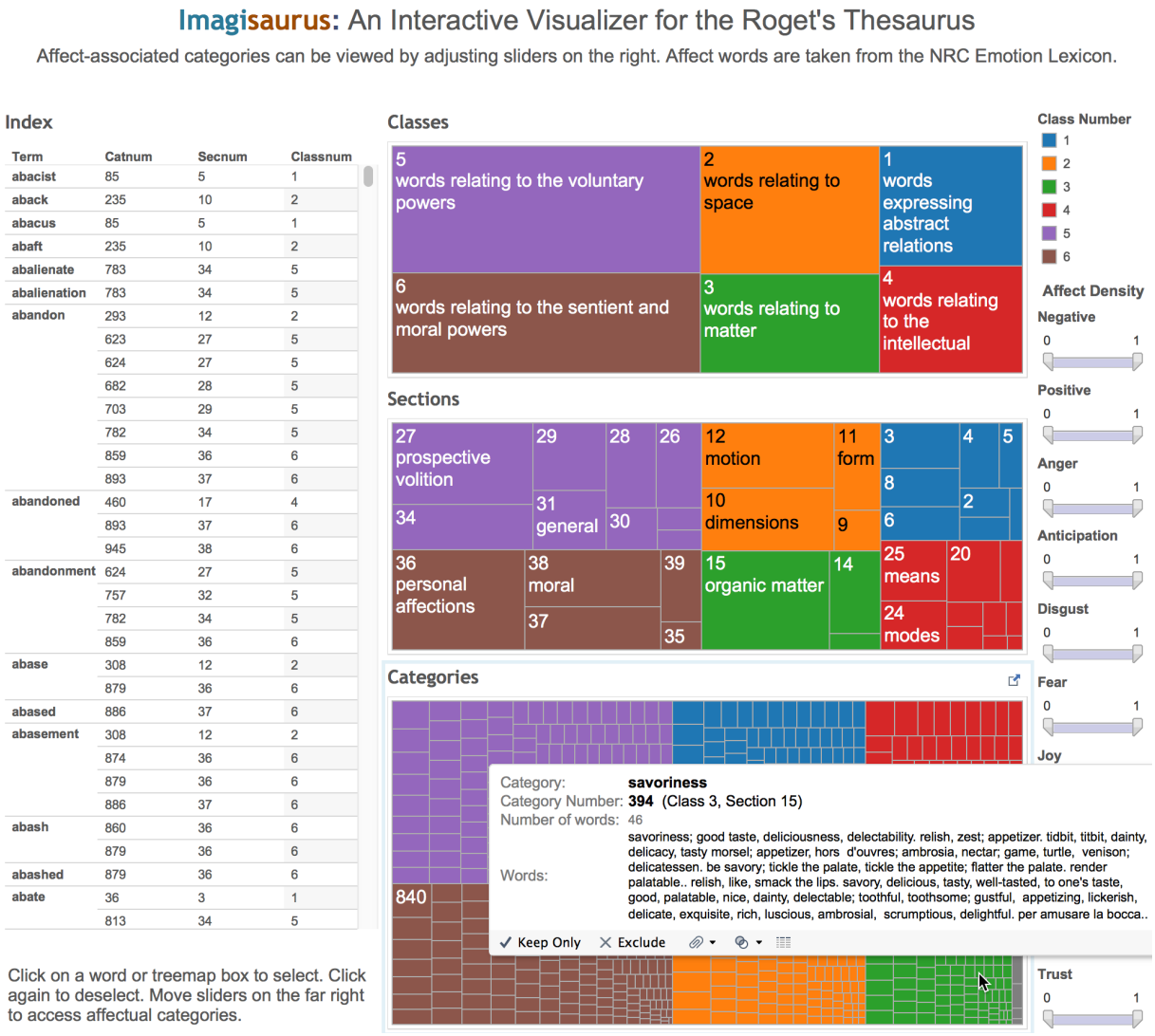

Figure 1: Screenshot of Imagisaurus when one moves the mouse pointer over one of the boxes in the Categories treemap. The tooltip info box pops up showing information pertaining to what is being hovered on—in this case category number 394 (savoriness).

chy. A treemap is a kind of visualization that partitions a large box representing one level into many smaller boxes pertaining to the descendant nodes.

If the box size permits, the name and number of the taxonomic unit is printed in it. For example, the name–number information for all classes and some sections is printed in the default view. This information is not shown for most of the categories in the default view, but as described ahead, when certain selections are made to reduce the number of categories, then this information appears even for the categories. Hovering over any box will always give the corresponding name-number information through a tooltip info box.

We describe each of the four sub-visualizations in the subsections below.

3.1 Index

The Index lists the words in alphabetical order. Users can scroll down the list to quickly locate the word they are interested in. They can then see which thesaurus categories the word is listed in (second column, Catnum), and also the corresponding section number (Secnum), and Class number (Classnum). Clicking on the word filters out information in all four sub-visualizations, leaving information pertaining only to the chosen word. For example, Figure 2 shows a screenshot of the treemaps for when the user clicks on the Index entry *abandon*. Observe that the three treemaps now show a blowup of information relevant only the chosen word: specifically, the classes, sections, and categories *abandon* is listed in.

Figure 2: Filtered view in Imagisaurus when one clicks on the word *abandon* in the index.

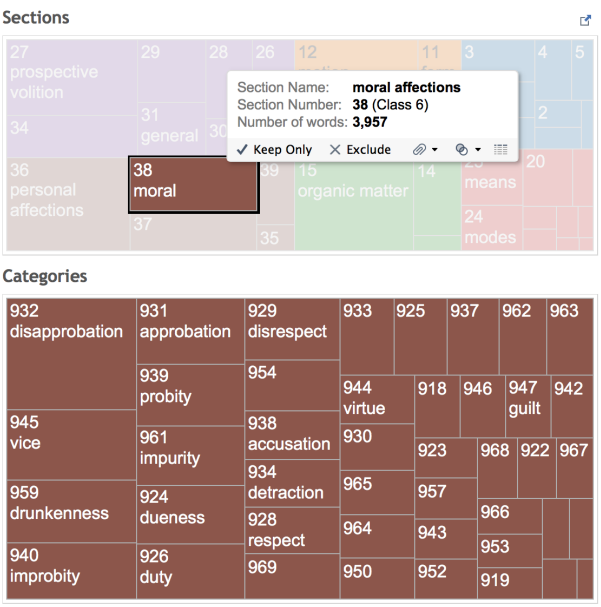

Figure 3: The Sections and categories Treemaps when one clicks on the Section 38 (moral).

3.2 Classes

The Classes treemap shows the six thesaurus classes as boxes. The size of each box is proportional to the number of words in the class. The treemap places the biggest boxes on the top left and the smallest boxes on the bottom right. This allows users to instantly gain a rough estimate of how large each class is. One can see for example that Section 5 has the most words and Section 4 the least. When selections are made in one of the other sub-visualizations and the Classes treemap filters to show relevant information (as in Figure 2 for example), one can then examine the sizes of the now-relevant classes. (For example, in Figure 2, one can now see the relative sizes of the three classes that list *abandon*.)

3.3 Sections

The Sections treemap shows all (or a selection) of sections in the *Roget's Thesaurus*. (Clicking on a particular class filters the Sections treemap to show only the relevant sections.) The sections are first grouped by class, and then within each of these groups they are ordered as per number of words in the sections. This allows users to quickly determine which sections are more dominant within a class. Clicking on a section filters information as one would expect. Figure 3 shows how the Sections treemap and the Categories treemap appear when one clicks on section 38 (moral). Observe that the Categories treemap now shows only those categories that are within section 38. The Index also filters to show rows for only those words that are listed in section 38.

3.4 Categories

The Categories treemap shows all (or a selection) of categories in the thesaurus. (Clicking on a class or section filters the categories treemap.) The categories are first grouped by class, and then within each of these groups they are ordered as per number of words in the categories. This allows users to determine which categories are more populous. Hovering on top of a category reveals a tooltip info box that shows not only the category name and number, but also the number of words in the category and a list of all these words. Recall that Figure 1 shows an example of this tooltip info box. Clicking on a category filters information in the Index to show only the rows for the words in the chosen category. The Class and Section treemaps show the class and section of the category.

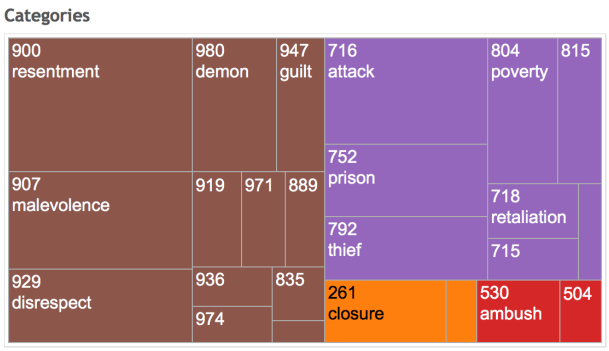

Figure 4: Categories treemap with the **anger** density slider set to range 0.7–1.

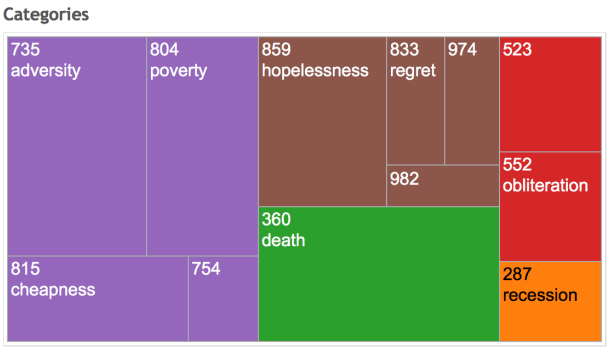

Figure 5: Categories treemap with the **sadness** density slider set to range 0.7–1.

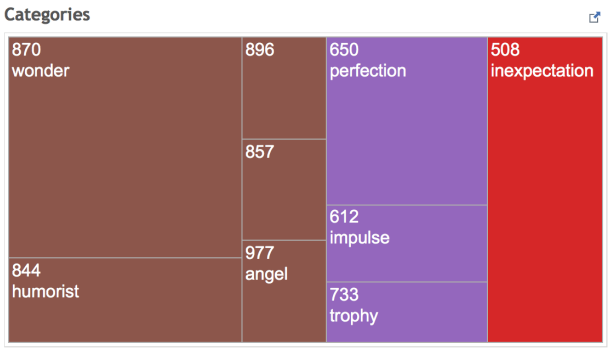

Figure 6: Categories treemap with **surprise** density and **positive** density sliders both set to > 0.4.

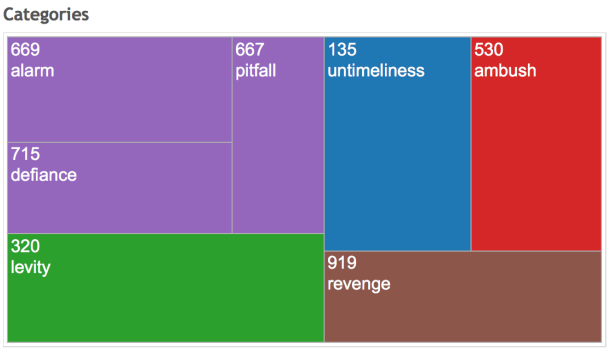

Figure 7: Categories treemap with **surprise** density and **negative** density sliders both set to > 0.4.

3.4.1 Identifying Affectual Categories

We now discuss how the *Roget's Thesaurus* is linked with the NRC Emotion Lexicon to display categories that have strong associations with various sentiments and emotions.

For each category *cat*, we calculate affect density for affect *aff* using the formula shown below:

$$\textit{Affect Density (cat,aff)} = \frac{\textit{NumAssociated}}{\textit{NumTotal}} \quad (1)$$

where *NumAssociated* if the number of words in *cat* associated with *aff* and *NumTotal* is the number of words in *cat* that are listed in the NRC Emotion Lexicon. Thus, for example, if a category has 50 words, 40 of which are listed in the NRC Emotion Lexicon, and 30 of these are associated with positive sentiment, then the category has a positive affect density of 30/40 = 0.75.

We calculated affect densities for both sentiments and all eight emotions covered in the NRC Emotion Lexicon. For each of these affects, Imagisaurus shows density sliders on the far right. Both the lower end (to the left) and the upper end (to the right) of the slider can be moved with the mouse pointer. Adjusting a slider filters the Categories treemap to show only those categories with affect densities within the range of the slider. For example, Figure 4 shows the Categories treemap as it appears when the lower end of the anger density slider is moved to 0.7 and the upper end is left at 1. One can compare it to Figure 5 which shows the categories with sadness density between 0.7 and 1. Observe that the former shows categories such as resentment, attack, and ambush, whereas the latter shows categories such as adversity, hopelessness, and death. One can even manipulate multiple sliders to create multiple filters that apply at the same time. For example, Figure 6 shows categories with surprise and positive densities each greater than 0.4. We see categories such as wonder, humorist, and perfection. On the other hand, Figure 7 shows categories with surprise and negative densities each greater than 0.4. We see categories such as alarm, untimeliness, and ambush.

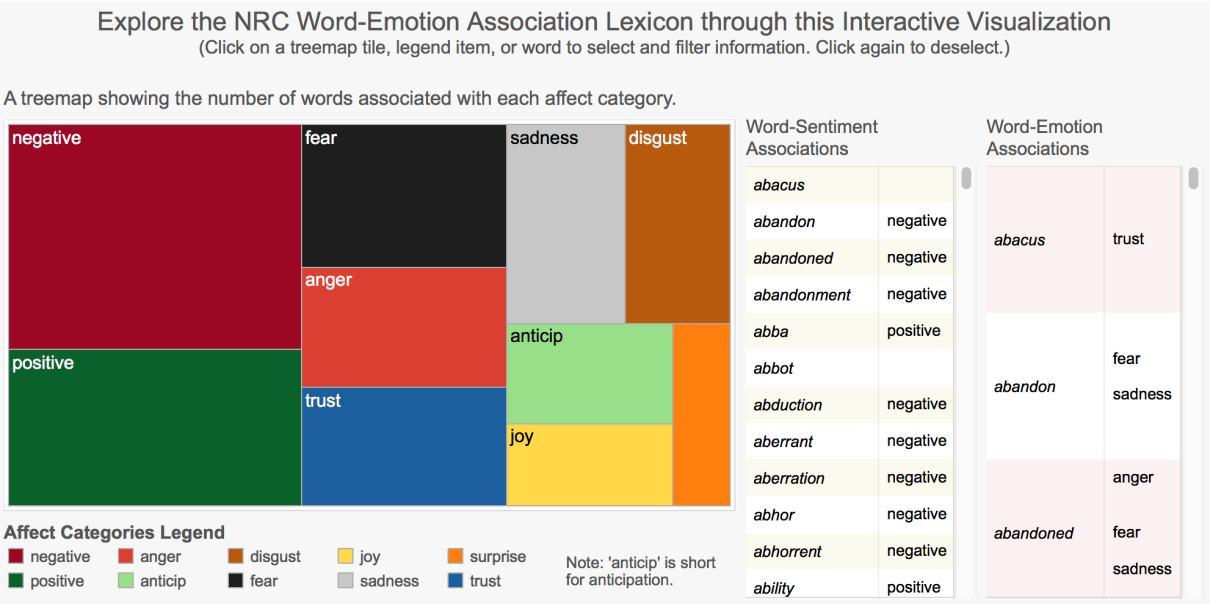

Figure 8: An interactive visualizer for the NRC Emotion Lexicon.

4 Visualizing words–affect associations

We developed a second online interactive visualization to explore word–emotion and word–sentiment associations directly in the NRC Emotion Lexicon. Figure 8 shows a screenshot of this visualization. The treemap on the left shows the various affect categories. The sizes of the boxes in the treemap are proportional to the number of words associated with the corresponding affect. Observe that, word associations with negative sentiment are more frequent than associations with positive. The associations with fear, anger and trust are much more frequent compared to associations with joy and surprise. On the right are two index views for word–sentiment and word–emotion associations respectively. Clicking on a word in one of the index views, filters information in all of the other sub-visualizations to show information relevant to that word. Clicking on a box in the treemap, filters information in all other sub-visualizations to show information relevant only to the chosen affect category.

5 Summary and Future Work

We developed an online interactive visualizer for the *Roget's Thesaurus* called Imagisaurus. Imagisaurus allows users to access information about thesaurus words, classes, sections, and categories through four separate sub-visualizations that are linked to each other. The structure of the thesaurus is shown in proportion to the size of its components—where size is defined to be the number of words included in the thesaurus unit (category, section, etc.). Clicking on a unit selects it and filters information in all other sub-visualizations. We also link the thesaurus with an emotion lexicon such that manipulating simple sliders allows users to view categories associated with affect categories. With its intuitive and easy-to-use interface that allows interactive exploration of the *Roget's Thesaurus*, we believe Imagisaurus will benefit researchers, practitioners, and the lay persons alike. We also developed a second visualization to explore the NRC Emotion Lexicon. Both visualizers are made freely available online.

This work explores the *Roget's Thesaurus* and the NRC Emotion Lexicon, but the same framework can be used to explore other lexical resources too: for example, other thesauri in English and other languages; semantic networks such as WordNet and VerbNet; versions of the NRC Emotion Lexicon in other languages; and sentiment lexicons such as the NRC Hashtag Sentiment lexicon and Sentiment 140 Lexicon (Mohammad et al., 2013; Kiritchenko et al., 2014).[8] Our future work will extend previous work on visualizing literature (Mohammad and Yang, 2011; Mohammad, 2012) by incorporating interactivity among sub-visualizations and by capturing affectual information associated with characters and plot structure.

[8] http://www.purl.com/net/lexicons

References

Gregory Grefenstette. 2012. *Explorations in automatic thesaurus discovery*, volume 278. Springer Science & Business Media.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (in press)*.

Gottfried Wilhelm Leibniz and George Henry Radcliffe Parkinson. 1995. *Philosophical writings*. Everyman.

GW Leibniz. 1923. Calculus ratiocinator. *Samtliche Schriften und Briefe. Reichel, Darmstadt*.

Margaret Masterman. 1957. The thesaurus in syntax and semantics. *Mechanical Translation*, 4(1-2):35–43.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 35–43. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

Saif Mohammad and Tony Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.

Saif Mohammad. 2008. *Measuring semantic distance using distributional profiles of concepts*. Ph.D. thesis, University of Toronto.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Peter Mark Roget. 1852. *Roget's Thesaurus of English Words and Phrases*. Longman Group Ltd., Harlow, Essex, England.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases*. TY Crowell Company.

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France.