Decoding the Color Lexicon

Shuodi Yin
23006440
01/12/2023

Introduction

Color, a powerful visual language, is omnipresent in our lives. It can evoke emotional responses and carry distinct messages. Singh (2006) highlighted the significance of color, noting that 62-90% of an individual's initial assessment of people or products is based on color within the first 90 seconds. This insight ignited my curiosity to delve into the relationship between color and its symbolic meanings. Thus, this study is dedicated to uncovering the societal semantic relationships of color across various fields and enriching our understanding of color as a medium of communication.

Using web scraping to extract data from Wikipedia articles, this study delves into an array of ten colors, amassing datasets from diverse fields such as art and culture, natural sciences, psychological symbolism, and sports. Thematic modeling was employed to uncover the subtle themes associated with these hues. Visualization tools like network graphs and word clouds have been instrumental in elucidating potential correlations between colors and their associated topics. Moreover, a Graphic User Interface (GUI) has been developed to provide an interactive exploration. This study is anticipated to offer fresh perspectives for creative designers and marketing professionals, bolstering their use of color in evoking desired emotional and cultural responses and informing their decision-making processes in practical applications.

Background

Color theory research indicates that each color is linked to specific emotions and cultural meanings. Elliot (2014) has shown how colors influence emotional responses and behaviors. For instance, blue is known to significantly stimulate the sympathetic nervous system (Kido, 2000) and is widely favored across cultures (Wiegersma, 1988). In contrast, yellow and orange are associated with happiness, while red, black, and brown often represent sadness (Cimbalo, 1978). This study aims to uncover the latent topics and emotional trends of these colors, examining their contextual connections to provide a comprehensive understanding of their symbolic significance.

Methodologically, the research leverages the Python libraries Requests and BeautifulSoup for efficient web scraping from Wikipedia, enabling substantial data collection. Two thematic modeling techniques, Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA), are employed to analyze the underlying themes in texts. The project's aesthetics are enhanced by using NetworkX (Stack Overflow, 2019) and word clouds (BJ Peng Da, 2020) for visualization, while TextBlob is employed for sentiment analysis of key terms, integrated with TF-IDF values. Additionally, a Tkinter-based GUI (CSDN, 2022) enhances user interaction with the analysis, offering a more intuitive experience. This technical approach aims to elucidate color's profound meanings across varied cultural and social contexts through advanced data analysis and visualization.

Method

My study combined the Requests and BeautifulSoup libraries to scrape data from Wikipedia on predefined color-related topics, using a User-Agent to mimic real-user requests to prevent multiple access blockage and parsed HTML content to extract text. I extracted the text content from all paragraphs and replaced spaces in topic titles with underscores to construct accurate URLs. Preprocessing was crucial for dataset clarity. It involved removing HTML tags, using the 'isdigit()' function to eliminate numerical interference in the analysis, and adding meaningless words to the stop word list through testing. Additionally, lemmatization was applied to standardize the vocabulary, distilling the essence of each text. Through this cleansing, each text's essence was distilled. Eventually, five topics were set.

The heart of the study lies in the analytical comparison between LDA and LSA. The results revealed that LDA excels at uncovering hidden topics in documents through iterative optimization to discover the most explanatory topic. In contrast, LSA, with its term-document matrix and Singular Value Decomposition (SVD), often results in significant overlap or similarity in keywords between topics. LDA was selected for its robust thematic separation, aligning with the research's aim. Subsequently, I used a network graph to visually present the results of LDA. Since each color contained 50 keywords, I represented keyword weight through the node area and used different shades of colors to distinguish topics, enhancing the aesthetic appeal. This graph illustrates the relationships between colors and topics.

For sentiment analysis, I combined TextBlob to analyze the sentiment scores of individual documents and calculated the term frequency-inverse document frequency (tf-idf) value of words in documents. This approach offered a depiction of the emotional information associated with each keyword, enriching the interpretability of results. This nuanced understanding was further visualized in a word cloud, where keyword size reflected thematic weight, and color intensity signaled sentiment. Since most sentiment scores were below 0.1 and varied little from each other, I normalized the scores to a range between 0 and 1 for a more prominent display in the word cloud. Additionally, I created a color bar to map the range of sentiment score changes, enhancing user observation. This code combined traditional word frequency visualization with sentiment analysis, offering a visual tool for presenting data.

The culmination of this study was the development of a tkinter-based GUI, which interactively displayed the analytical journey from color selection to thematic understanding. The GUI comprised three interfaces, with the main window initiating the main loop and subsidiary windows allowing for dynamic content loading. Dropdown menus facilitated user navigation, enhancing the exploratory experience under each color. In summary, this project not only synthesized a vast array of data into a coherent narrative but also provided an accessible platform for users to navigate through the multifaceted relationship between color and language.

Results

In the data collection phase, I gathered data from Wikipedia about ten colors. Each color was explored across five domains: color, art & culture, science and nature, psychology, symbolism, and sports. In total, I collected 253 texts, each averaging 4,568 words, totaling around 1,155,631 words.

Dataset		
Color (Folder)	Number of txt Files	Number of Words
Red	29	164,283
Orange	21	87,974
Yellow	28	143,421
Green	27	119,036
Cyan	23	74,610
Blue	28	121,177
Purple	26	110,123
Black	28	158,326
White	22	112,594
Pink	21	64,087
Total	253	1,155,631

Figure 1: Overview of Dataset

An in-depth LDA and LSA topic modeling analysis was performed on the dataset, taking the red folder as an example (Figure 2). LSA has a lot of overlapping words between different topics. For instance, topics 3 and 4 have the same keyword 'scarlet' with significant weights. On the contrary, LDA shows clearer differences between topics, and the highest-weighted keywords in each topic are representative and have associations with other words within the same topic.

Red Folder	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
LDA	person: 65.63	goal: 52.25	area: 68.55	vermilion: 61.20	body: 58.27
	expression: 67.78	psychologist: 52.46	system: 77.41	national: 64.19	star: 62.20
	rome: 70.99	research: 55.62	classic: 80.12	surface: 67.61	thermodynamic: 77.20
	emotional: 72.18	chariot: 71.67	power: 87.50	association: 69.13	culture: 83.73
	social: 82.93	performance: 93.83	high: 93.99	player: 73.39	dwarf: 84.20
	theory: 90.72	ferrari: 114.20	period: 108.20	planet: 74.92	temperature: 92.04
	roman: 129.40	psychology: 152.14	chinese: 146.52	scarlet: 91.61	system: 93.12
	war: 175.00	race: 165.48	maya: 302.20	earth: 94.22	transfer: 101.20
	anger: 314.19	athlete: 177.20	laser: 305.20	football: 114.20	energy: 118.43
	emotion: 346.10	sport: 309.06	china: 381.92	mar: 230.10	heat: 206.06
LSA	maya: 0.09	behavior: 0.08	dwarf: 0.13	pigment: 0.09	ink: 0.09
	culture: 0.10	rgb: 0.08	spot: 0.14	vermilion: 0.09	painter: 0.10
	sport: 0.10	brand: 0.08	sunrise: 0.15	rgb: 0.09	rgb: 0.11
	renaissance: 0.10	emotional: 0.10	cmyk: 0.15	football: 0.11	painting: 0.11
	laser: 0.11	commentator: 0.10	wavelength: 0.15	commentator: 0.12	anger: 0.12
	emotion: 0.12	athlete: 0.15	mar : 0.15	race: 0.12	vermilion: 0.14
	scarlet: 0.13	sport: 0.19	opsin: 0.16	sport: 0.13	artist: 0.16
	war: 0.14	psychology: 0.20	ink: 0.17	ferrari: 0.13	emotion: 0.16
	china: 0.15	anger: 0.28	laser: 0.24	laser: 0.16	renaissance: 0.20
	roman: 0.15	emotion: 0.39	rgb: 0.24	scarlet: 0.32	scarlet: 0.20

Figure 2: A comparison of LDA and LSA topic modeling results (red Folder)

Based on the LDA, I proceeded with network graph analysis. Figure 3 displays the weight relationships of the top 50 keywords associated with red. In contrast to the traditional approach of using edge thickness to represent weight, this study opted for using node area to represent weight. This decision was made because there were too many words and even with normalization, it was not convenient for observation, as shown in Figure 4.

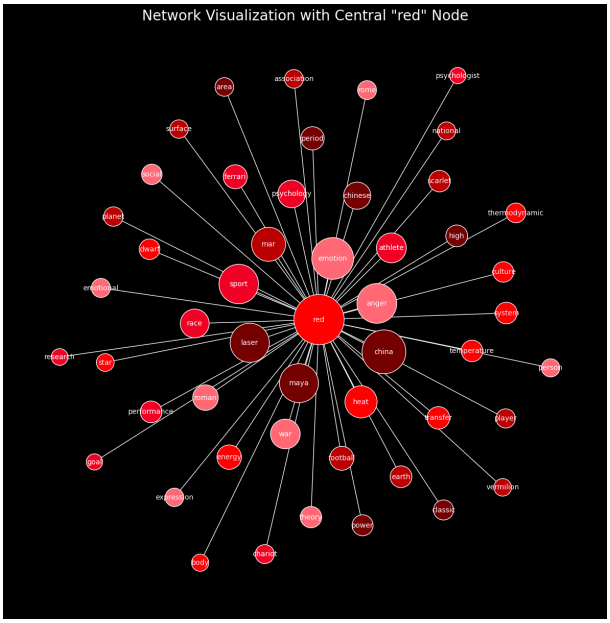


Figure 3: Network with 'Red' Node (node size)

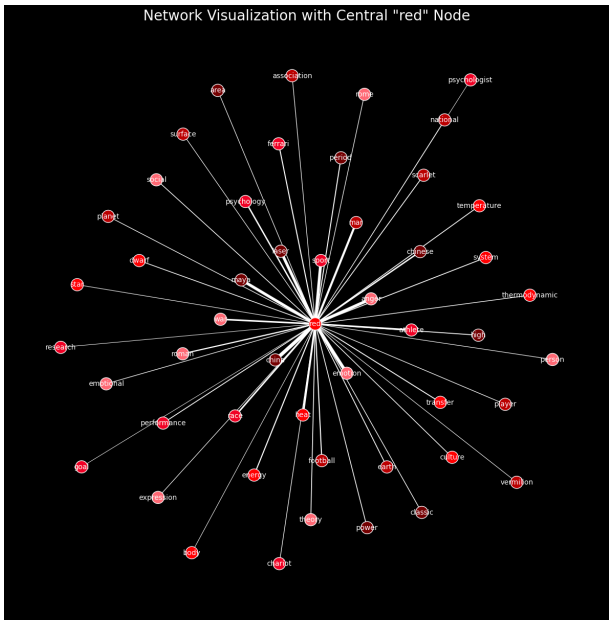


Figure 4: Network with 'Red' Node(edge thickness)

Figure 3 indicates that key terms such as 'emotion', 'anger', 'China', 'war,' and 'culture' prominently appear due to their high weights, showing their strong relevance to the color red. These top-weighted words showcase 'red's' extensive cultural and emotional associations: 'anger' is often depicted with a red emoji and 'China' is represented by its red national flag. Yet, the network graph's representation might not fully capture the depth of connections among these keywords. To probe further, sentiment analysis was employed. Figure 5 captures these sentiment scores, providing insight into the complex relationships that 'red' maintains with these dominant words.

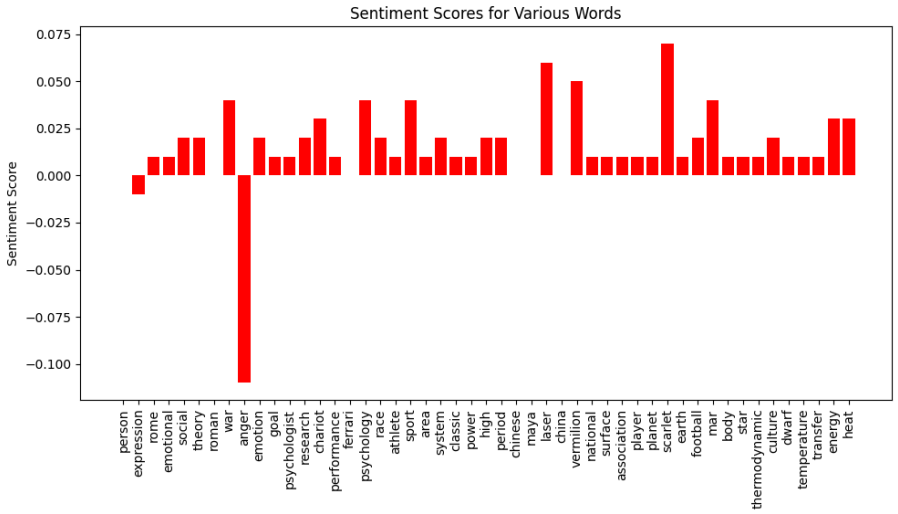


Figure 5: sentiment scores for target words

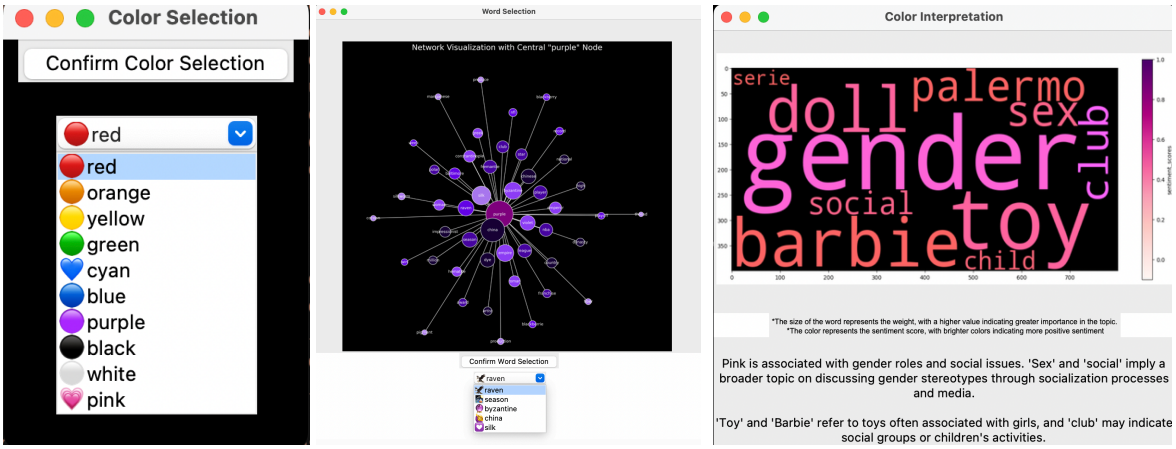
The scores range from -0.11 to 0.07, indicating that most words are associated with neutral or slightly positive sentiment. 'Scarlet' and 'laser' have relatively high positive scores (0.07 and 0.06, respectively), indicating that these words tend to appear in positive contexts. Many words such as 'person', 'Roman', 'Ferrari', 'Chinese', etc., have scores close to zero, which may suggest that they are typically associated with neutral sentiment in the documents, more as objective descriptions. Additionally, 'Anger' has the lowest score of -0.11, suggesting that it is typically associated with negative sentiment. However, 'war' has a score of 0.04, which might be because, in some documents, it is used to describe positive contexts like victory or heroism, despite 'war' generally being considered negative.

A subsequent word cloud integrated both keyword weights and sentiment scores, presenting an insightful visual where the size and color brightness of words corresponded to their thematic significance and emotional valence, respectively.

Figure 6: 'Red' Word cloud

The word cloud centered around "emotion" suggests red's association with strong feelings, particularly 'anger', aligning with its representation in contexts of social theory, where it symbolizes nationalism, conflict, and revolution. The presence of words like "social", "person" and "expression" highlights red's broader historical and cultural significance. It's seen as a marker of happiness and good fortune in some social customs, whereas ancient Rome linked it to power and authority. This visualization thus captures the deep societal and emotional themes connected to the color red.

The analysis generated tailored texts for each word cloud, integrated into a user-friendly GUI with three interfaces. In the first interface, users pick a color from a dropdown menu, and then 'Confirm Color Selection' leads to the second interface displaying a network graph and a new menu. Selecting topics from the graph opens the third interface, showing a word cloud and text analysis for the chosen keyword. Each step is intricately connected, and with the use of color and emojis for visual enhancement, the user experience is enriched.



First Interface: color selection Second Interface: word selection Third Interface: word cloud

Figure 6: GUI Interface

Discussion

This study's analysis of Wikipedia's color articles uncovered varied symbolic meanings. While keywords were chosen from hyperlinked terms on color pages, personal judgment in dataset selection may have introduced biases. Future research should use systematic methods and a wider range of colors for a thorough examination of color's relationship with culture and emotion.

A comparative analysis of topic modeling using LDA revealed methodological contrasts. LDA excelled in separating topics, while LSA showed more keyword overlap, especially in documents with less data, highlighting its limitations in differentiating thematic elements. Future research might yield more precise results by comparing additional topic modeling techniques. On the other hand, network graph analysis, by representing keyword importance with node size instead of traditional line thickness, enhanced data visualization. However, relying solely on color differences seems insufficient. Implementing interactive elements like dynamic charts could improve user engagement. Moreover, a custom function combining sentiment scores and tf-idf values offered a multi-dimensional perspective, but lacked standardization in sentiment analysis, requiring further validation.

Lastly, the GUI's guided design effectively demonstrated the complex interplay between colors and themes. While the current interface offers a structured exploration, I recognize the importance of user feedback in further refining the tool. Future enhancements will focus on integrating mechanisms to collect user insights, aligning more closely with user needs and preferences.

Conclusion

From initial data collection to the creation of interactive GUIs, this project has highlighted the complex symbolic significance and communicative power of colors. Our analysis using networks and word clouds shows colors' visual impact and cultural-emotional associations but also highlights the need for nuanced interpretation. While providing insights into creative and marketing fields, the project reveals the limitations of relying solely on open-source data like Wikipedia for understanding color symbolism. Future enhancements could focus on diversifying data sources and incorporating user feedback to refine the tool's accuracy and cultural relevance, ensuring a more comprehensive understanding of color's role in communication.

Ethical considerations

This project employs Wikipedia to ensure a broad, objective view while minimizing biases. However, it also recognizes the limitations of open-source data, such as potential inaccuracies. In discussing color symbolism, we are careful to avoid stereotyping individuals.

LLM disclaimer

I utilized large language models (LLMs) as an auxiliary tool, which assisted me in fixing errors in the code, as well as in translating the essay writing.

Word count:1940

Bibliography

[1] BJ Peng Da (2020). Generating Word Clouds Based on Word Frequency (Implemented in Python with Wordcloud Library). Sunxiupeng Blog, 21 February. Available at: https://sxpujs.github.io/2020/02/21/generate-wordcloud-from-frequencies/(Accessed 28 November 2023).

[2] Cimbalo, R.S., Beck, K.L. and Sendziak, D.S. (1978) .Emotionally toned pictures and color selection for children and college students. *Journal of Genetic Psychology*, Vol. 33 No. 2, pp. 303-4.

[3] CSDN (2022). Python GUI Interface Design with Tkinter. Available at: https://blog.csdn.net/smallfox233/article/details/112093464(Accessed 28 November 2023).

[4] Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65, 95-120.

[5] How to use beautifulsoup's find_all() method: ScrapeOps (no date). ScrapeOps RSS. Available at: https://scrapeops.io/python-web-scraping-playbook/python-beautifulsoup-findall/ (Accessed 20 November 2023).

[6] Kido, M. (2000). Bio-psychological effects of color. *Journal of International Society of Life Information Science*, Vol. 18 No. 1, pp. 254-62

[7] Networkx graph plot node weights. (2019).Stack Overflow.Available at: https://stackoverflow.com/questions/56294715/networkx-graph-plot-node-weights(Accessed 25 November 2023).

[8] Python Assets. (2022). Drop-down list (Combobox) in TK (tkinter), Python Assets. Available at: https://pythonassets.com/posts/drop-down-list-combobox-in-tk-tkinter/(Accessed 28 November 2023).

[9] Python get wordnet pos. (No date). Available at: https://www.programcreek.com/python/?CodeExample=get+wordnet+pos (Accessed 20 November 2023).

[10] Python tkinter - toplevel widget. (2021). GeeksforGeeks. Available at: https://www.geeksforgeeks.org/python-tkinter-toplevel-widget/(Accessed 28 November 2023).

[11] Removing numbers from string, Stack Overflow. (2012). Available at: https://stackoverflow.com/questions/12851791/removing-numbers-from-string(Accessed 21 November 2023).

[12] ScrapeHero. (2023). How to Scrape Websites Without Getting Blocked. Available at: https://www.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping(Accessed 20 November 2023). Published 14 November 2023.

[13] Shah, P. (2020). My absolute go-to for sentiment analysis - text blob., Medium. Available at: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524 (Accessed 25 November 2023).

[14] Singh, S. (2006). Impact of color on marketing. *Management decision*, 44(6), 783-789.

[15] Standalone color bars. (no date). Standalone color bars. Available at: https://matplotlib.org/stable/users/explain/colors/colorbar_only.html#sphx-glr-users-explain-colors-colorbar-only-py(Accessed 28 November 2023).

[16] Training, P.T.P. (2023). Tutorial: How to normalize data in Python, Pierian Training. Available at: https://pieriantraining.com/tutorial-how-to-normalize-data-in-python/(Accessed 28 November 2023).

[17] Wenliam. (2023). Python NetworkX Co-occurrence Graphs through LDA Topic Keyword Co-occurrence. CSDN, 11 September. Available at: https://blog.csdn.net/weston95/article/details/132620651(Accessed 25 November 2023).

[18] Wiegersma, S. and Van der Elst, G. (1988). Blue phenomenon: spontaneity or preference? *Perceptual & Motor Skills*, Vol. 66 No. 1, pp. 308-10.

[19] Anonymous. (2022). Python Tkinter Multi-Window (Form) Programming Example. CSDN, 18 April. Available at: https://blog.csdn.net/cnds123/article/details/123425360 (Accessed 28 November 2023).

[20] How to resize an image using Tkinter? (no date). Online Tutorials, Courses, and eBooks Library. Available at: https://www.tutorialspoint.com/how-to-resize-an-image-using-tkinter(Accessed 28 November 2023).