**MINING & MODELING FOOD TRENDS**

Leveraging social media and public APIs to detect food trends & strategies through natural language processing

**MICHAEL PAGAN & MATTEO BUCALOSSI**

**12/10/2021**

**Table of Contents**

* Introduction
  + Introduction/Background
  + Problem Statement
  + Problem Elaboration
  + Motivation
  + Project Scope
* Literature Review
  + Relevant Research
* Methodology
  + Data Collection
  + Dataset Description
  + Data Preprocessing
  + Data Modeling & Visualizations
* Results & Analysis
* Conclusion
  + Conclusion
  + Project Limitations
  + Future Research
* Appendix

**Glossary of Terms**

* **ACL:** Association for Computational Linguistics
* **API:** Application Programming Interface
* **BERT:** Bidirectional Encoder Representations from Transformers
* **LDA:** Latent Dirichlet Allocation
* **NLP:** Natural Language Processing
* **QSR:** Quick-Service Restaurant
* **R&D:** Research & Development
* **S-BERT:** Sentence-BERT
* **TF-IDF:** Term frequency-inverse document frequency
* **UI:** User-Interface
* **W2V:** Word2Vector

**Introduction**

Introduction / Background:

This project was conceived as part of Matteo’s internship at Inspire Brands where the Data Science Director expressed interest in investigating the applications of data science and machine learning to expand his team’s work across their company, Inspire Brands.

Inspire Brands is the second largest food-service company in the U.S.[[1]](#footnote-0) and the owner and franchisor of several restaurant brands nationally and overseas, including Arby's, Buffalo Wild Wings, Sonic Drive-In, Jimmy John's, Rusty Taco, Dunkin' and Baskin-Robbins. As the company grows larger in the QSR industry, so does the offering of products and cuisines, creating a need for Inspire to invest in research and development to stay competitive in all the markets involved.

Understanding the foods consumers are interested in is vital for companies with a vast footprint across the country like Inspire, especially in the current era of technology and social media where trends come and go. Thus it is beneficial to the goals of Inspire and their Data Science team as they seek to improve and support the R&D and menu planning parts of the business. Using textual data that is available via public APIs and social media, we applied unsupervised machine learning techniques to investigate and model data and to present the results to stakeholders, as well as develop a baseline model for future enhancements at Inspire.

For this capstone project, we proposed a prototype product that can gather information around what people are saying about food from Twitter and provide these insights to Inspire executives and researchers about what is trending for a given geographical area of interest.

Problem Statement:

The restaurant industry struggles with gaining a competitive advantage as food trends readily change in a dynamic market influenced by social media. Being able to innovate food offerings and understand trends means success in this fluctuating market.

Using current data from social media to understand food popularity by dish and region, industry leaders can keep up with these trends when considering new menu items or updates, and illustrate to consumers where these trends are available in their area.

Problem Elaboration:

Foodservice holdings and companies find themselves within a competitive market, often struggling with and using significant resources on innovation. Providing new products and menu items relevant with current trends embraced by the general public means success in a small profit-margin industry.

In addition to demand forecasting and incentive modeling, restaurant companies have to be creative to stay relevant and compete in a dynamic advertisement environment. Bringing innovations to their product offerings and providing novel, trendy menus that cater to the taste and hype of the masses are quickly becoming vital within this industry.

Huge amounts of data are poured daily into social media, blogs, reviews, and food-related websites. Tapping into public APIs via Python to build extensive and manageable scraping frameworks will allow us to implement a system that will stay on top of current trends in tastes and flavors.

The analysis on these data will provide insight and understanding of the customer base to foodservice companies, such as Inspire Brands, that will strengthen their R&D processes and will gain competitive advantage when bringing new ideas and products to the market.

Motivation:

Innovation of food products is often a lengthy process and thereby streamlining and improving the process could profit businesses significantly. Rather than surveying subject matter experts or simply reacting to competitors’ initiatives, data could provide the head start that is often needed in business to predict consumer trends.

Thus, Inspire Brands anticipated the possibility of using data science tools and technology to support this type of work. While the company’s data science team had always worked to support marketing and operations, the ability to work in this space directly is appealing as it could prove the value of data science in more than just its typical applications.

The barrier to obtaining public social media data is now lower than ever and Inspire Brands is keenly interested in leveraging this fact to gain insights on consumer preferences across space and time and, importantly, confirm their business intuitions to direct decision-making processes as they strive to stay current with popular tastes through the use of NLP and machine learning applications.

Scope:

Our project focused developed over four different steps:

* Data Collection
  + Gathering data from Twitter through the Tweepy package
  + Capturing zip code-level restaurant information using Documenu’s API
* NLP:
  + Using basic NLP techniques, such as token and n-grams frequency,
  + LDA clustering to explore the dataset.
* Machine learning:
  + Vectorizing free text using W2V and S-BERT so users can retrieve relevant text from tweets by cosine similarity from their searches.
  + Using the TF-IDF algorithm to identify trending chatter among the retrieved data.
* Dashboard:
  + The results are finally presented in a Dash-based UI for users to find trendings foods in their area.

**Literature Review**

Researching into the use of NLP within the restaurant and food industry led to some very interesting previous work conducted on this space. In Particular, we were able to find similar approaches adopted on relevant datasets on common machine learning mediums and blogs, such as Cuisine Classification & Topic Modeling[[2]](#footnote-1) and Foodie Discovery.[[3]](#footnote-2) Moreover, as we were surveying business applications, we discovered vendors that have been precursors of what we were trying to build in-house for Inspire Brands, and the most successful among those were Tastewise[[4]](#footnote-3) and Spoonshot.[[5]](#footnote-4)

Relevant Research:

Academic research has also touched on food themes and provided innovative models trained on such data, or entirely new corpora for training. For instance, on the corpus creation and data gathering side of the project, RECIPE1M+[[6]](#footnote-5) is a massive-scale project to gather recipe text and image data and learn cross-modal embeddings for this specific task through neural networks. Similarly, Large-scale Recipe and Meal Data Collection as Infrastructure for Food Research[[7]](#footnote-6) focuses on recipes’ corpus construction and less on the trending of food components within popular preferences or modeling of those. On the analysis side of the research, Wiegand, M., Roth, B., & Klakow, D’s paper provides a starting point for classifying in a weakly-supervised fashion a corpus of food-related terms, where the goal is to find the distinction between composite foods from simple ones[[8]](#footnote-7). Finally, Food Recipe Alternation and Generation with Natural Language Processing Techniques[[9]](#footnote-8) touched upon language generation techniques to provide novel inputs for recipes, providing an inspiring precedent for what is possible with NLP and the latest deep learning approaches.

Overall, our review of recent and available literature on the topic taught us that there is enough interest on similar projects in both academic and professional sectors, and that the newest models have a strong potential of disrupting the industry, as well as providing insightful tools to businesses. Notably, though, is a lack of peer-reviewed literature on standardized methodologies to query social media sites like Twitter for a given business problem.

**Methodology**

Data Collection:

*Twitter data*

To analyze food trends, it was important to not only pull relevant data from Twitter i.e., tweets that are food-related, but also popular. Though topics do “trend” on Twitter[[10]](#footnote-9), relying on trending food topics is not a reliable method for our problem as it would leave our query for tweets to random chance of food topics trending at that time. Nonetheless, we can understand how topics are performing on Twitter through the use of RiteTag[[11]](#footnote-10), which provides insights on how well a given hashtag is performing on Twitter. Using the Selenium library[[12]](#footnote-11), we supplied a list of food words[[13]](#footnote-12) and other hashtags (**Supplementary Figure 1**) to RiteTag and scraped the results to create a dataset that captured unique tweets, retweets, and views per hour for each item supplied to the site. Hashtags that had more than 10,000 views in the past hour were retained for the list of hashtags to search Twitter for and all others were discarded. A cutoff of 100,000 views was initially implemented but proved to be too restrictive of a search.

We then built a wrapper around the Tweepy library[[14]](#footnote-13) that allowed us to both scrape and stream tweets that contained the list of hashtags retained to look back at seven days’ worth of tweets and capture tweets as they were published, respectively. The statistics on the entire corpus of hashtags to search for was refreshed each hour to allow the query to search for the hashtags that met the 10,000-view cutoff. Altogether, we were able to scrape/stream approximately one month’s worth of tweets using this methodology.

*Restaurant menu data*

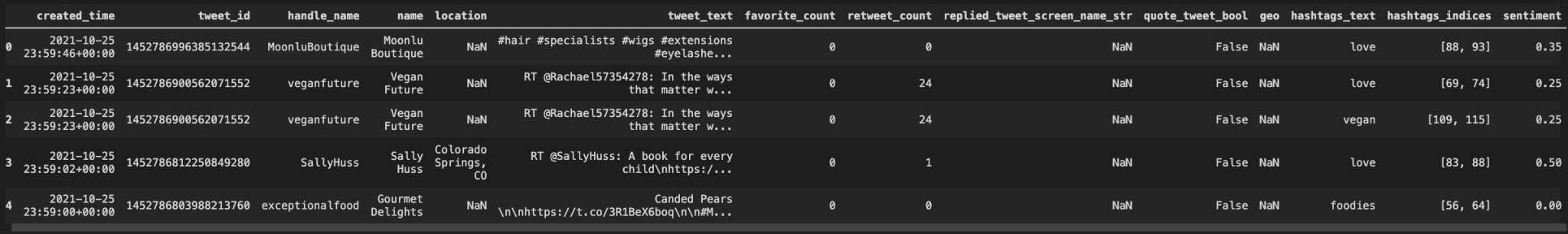
Downstream of the solution explored in our capstone project was a Dash-based UI where users can input a zip code and a food to then be presented with information on what is often found with their input food, the frequency of trigrams with this food, and a map displaying restaurants with this food item on the menu within the provided zip code. The former two insights are achieved via NLP analyses and modeling and is discussed in the Data Modeling & Visualizations and Data Preprocessing sections; data on the restaurant menus were captured by tapping into an API provided by Documenu[[15]](#footnote-14).

The user-input zip code on the UI was used as an input parameter to query Documenu’s GET Restaurants By Zip Code API[[16]](#footnote-15). A dataset containing information on the restaurants in the provided zip code was built out in-memory for each submission from the UI.

Dataset Description

*Twitter data*

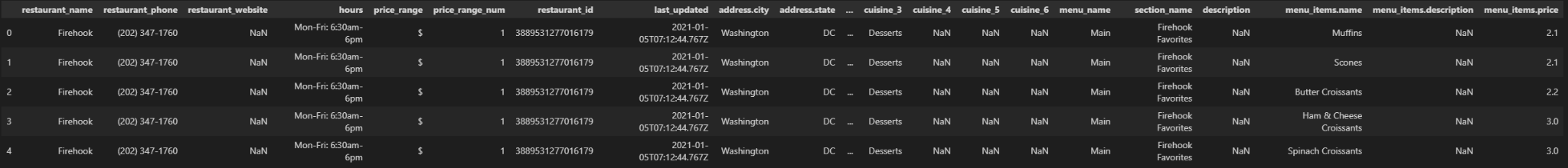
Mining Twitter data resulted in a dataset containing 3,166,787 rows, though the final dataset had just 288,366 rows across 32 columns; roughly 9% of the data mined was retained as a result of tweet redundancy and null values. The data captured contained information about the tweet - including the user information, the tweet text, tweet and user locations, hashtags, URLs, or images it contained, and the interactivity with the tweet, such as retweets and likes. Additional columns were added to capture the sentiment of the tweet text using TextBlob Sentiment Analysis[[17]](#footnote-16), and the zipcode of the tweet location using pyzipcode[[18]](#footnote-17) (**Fig. 1**)



**Fig. 1:** Sample Twitter data

*Restaurant menu data*

Restaurant data from Documenu were limited to 25 results per query when searching for menu information due to a bug in the API parser (**Supplementary Figure 1**); requesting the data without the menu would return all possible data. Nonetheless, the response from the GET Restaurants By Zip Code API returned information on the name, address, location, and cuisine(s) of the restaurant as well as the text from the entire menu, including price (**Fig. 2**).



**Fig. 2:** Sample Documenu data

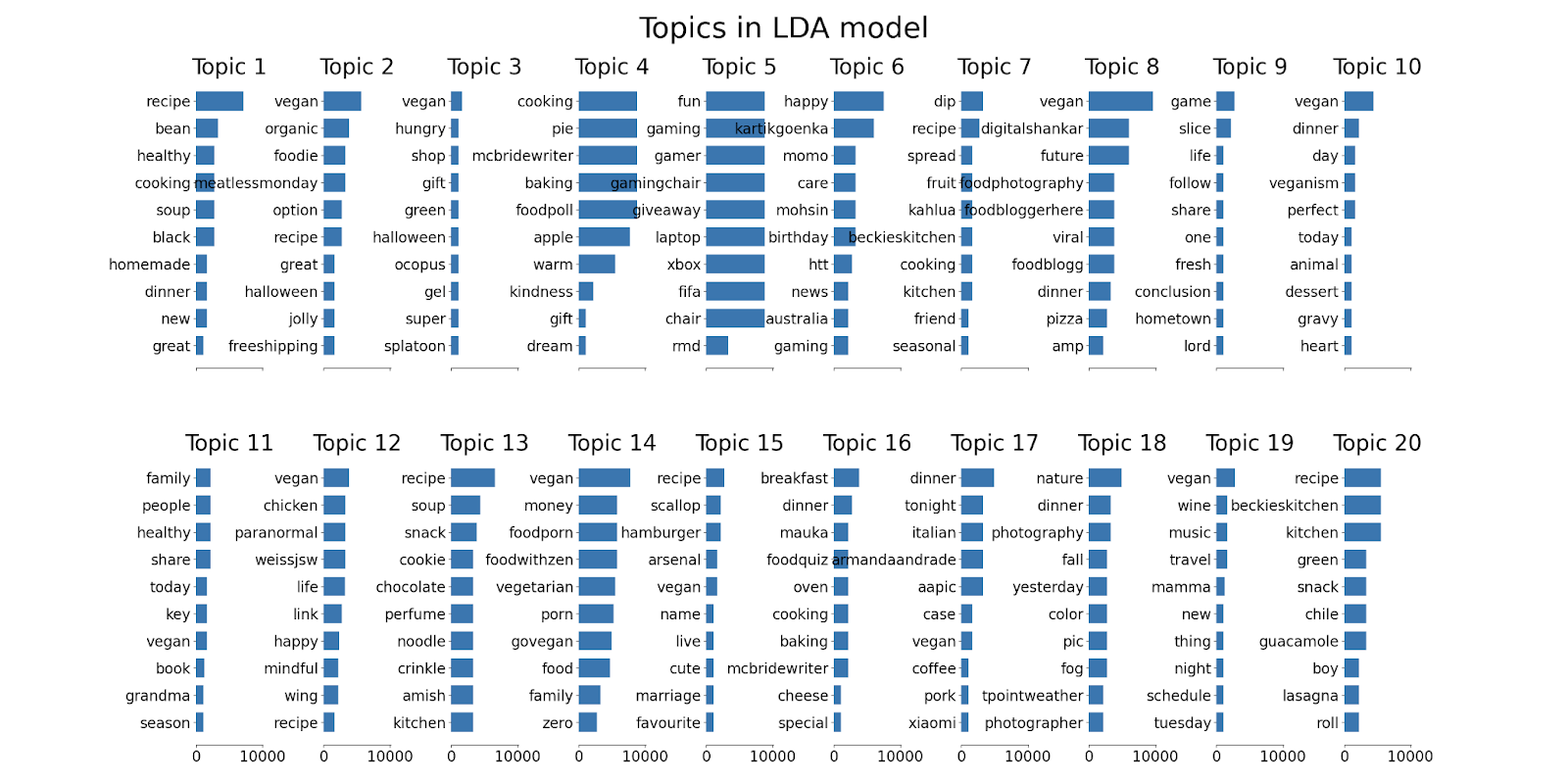
Data Preprocessing:

Once we obtained the Twitter data through their developers’ API, we focused on the text of each Tweet. We first had to process and clean Tweets’ text from expected impurities and symbols that pollute social media contents and Python libraries like NLTK[[19]](#footnote-18) and Preprocessor[[20]](#footnote-19) handled these tasks reliably.

Once the tweets had been cleaned and all text had been converted to lower-case, we used Spacy[[21]](#footnote-20) to only keep the lemmas of the nouns, proper nouns and adjectives of the text to limit our analysis to food-related terms and excluding other articles of the language not directly applicable to our project.

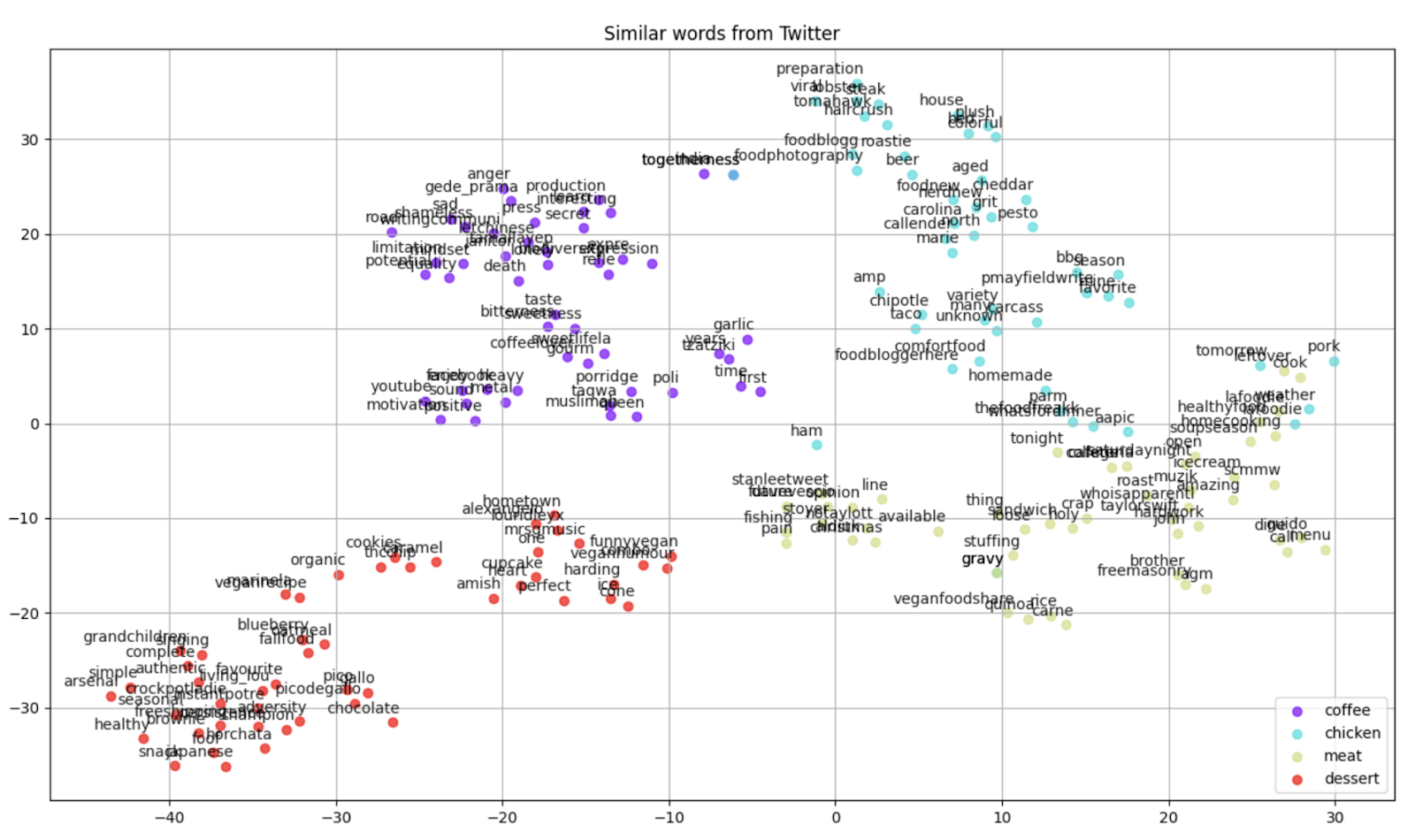
An exploratory analysis into the cleaned data consisted of inspecting the most frequent tokens and n-grams, particularly bigrams and trigrams. We assessed from these investigations that our dataset was indeed relevant and potentially useful for downstream tasks.

Furthermore, we were interested in investigating the topics that were present in this dataset. The Latent Dirichlet Allocation model[[22]](#footnote-21) is an unsupervised technique that looks at terms across documents in a corpus and identifies different distributions of terms as probable topics that are present in said corpus. A clustering algorithm, LDA expects that the user input the number of topics to search for, which can easily lead down a path of trial and error. We approached this problem using a grid-search algorithm to find the optimal number of topics for our dataset, as this approach would test LDA for a range of values and find the one that results in lower perplexity. 20 topics was found to be the optimal number to search for, though topics of this magnitude often reduce the interpretability of the output. Looking at Figure 3, the interpretability of the result is highly sacrificed by the number of topics detected, making LDA a less than desirable analytical route to pursue for our specific dataset.



**Fig. 3:** Topics for corpus identified by LDA.

Data Modeling & Visualizations:

As we trained our Twitter corpus on W2V[[23]](#footnote-22), we obtained dense vectors for each token, meaning that the model output fixed meanings for each word that was present in the tweets. For instance, “chicken” would have the same vector regardless of the sentence where it occurred. This makes analysis by similarity fairly straightforward with the embeddings obtained from W2V as every wodk has always the same embedding and can then be compared with others among the corpus (**Fig. 4**)

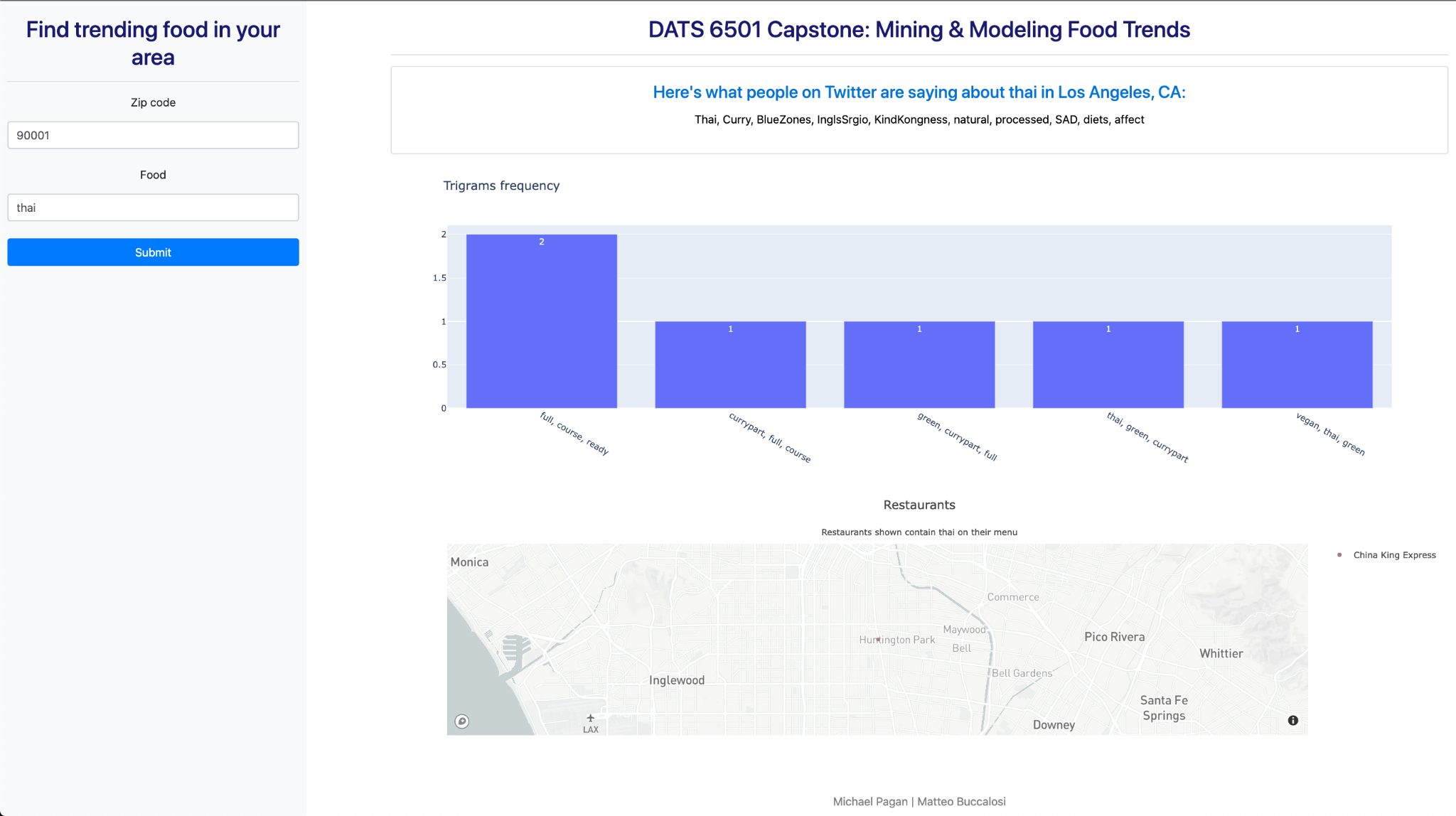
**Fig. 4:** W2V results when searching on coffee, chicken, meat, and dessert.

As we tried to improve the word embeddings to perform more accurate similarity-based analyses, we trained a BERT-based model on our corpus. BERT has revolutionized NLP since the model was published in 2018[[24]](#footnote-23), and its ability to capture context within the embeddings seemed to be useful given the linguistic facets of Tweets. However, we soon realized that word-level embeddings obtained from BERT were sparse vectors that are different for each occurence of each term in the corpus. For instance, the word “chicken” would have an entirely different vector each time it occurs in the corpus. Thus, although a vital tool for tasks like text generation and classification, word embeddings from BERT did not provide a useful base for performing downstream tasks based on similarity measures as we could not simply measure similarity across terms as each occurrence (not term) had a completely different embedding.

S-BERT[[25]](#footnote-24) is a Transformer model that can obtain contextual embeddings for sentence-long parts of text - this was directly applicable to our Tweets since we could obtain unique embeddings for each tweet in our corpus. Once the Twitter corpus was vectorized with a pre-trained S-BERT model, we obtained an embedding feature for each row representing that tweet’s text allowing us, or a user, to query the whole corpus and find the tweets that are most relevant to said query simply by calculating cosine similarity across the dataset.

**Results & Analysis**

The results derived from the NLP analysis and modeling were presented on a UI built on Dash[[26]](#footnote-25). Here, users are presented with two input fields on the left-hand side where they can enter a food item and a zip code to be presented with keywords often associated with that item, the frequency of the bigrams associated with it, and a map of restaurants that have that item on their menu (**Fig. 5**)

****

**Fig. 5:** Dash-based UI presenting the NLP and Documenu results.

Behind the scenes, this dashboard first filters the Twitter data by the provided zip code and the tweets from the resulting rows and serves as the input for the NLP analyses and modeling. Using TF-IDF[[27]](#footnote-26), which identifies the importance of single terms across a corpus, we communicated the keywords often associated within those tweets found by their search. The tweet text from the filtered dataset is then injected into the S-BERT model to calculate the embeddings for the sentences in each tweet. The filtered tweets are then used to deduce the common trigrams associated with the food input. Finally, the zip code is used to query Documenu’s GET Restaurants By Zip Code API and return the location of restaurants in the provided zipcode that have the input food item on their menu.

**Conclusion**

Conclusion

For this capstone project we scraped Twitter to create our own dataset of food-related tweets and provided a prototype product that is able to utilize unsupervised NLP approaches to gather insights from unlabeled free text data, all of which can be presented to a user on something like a Dash-based UI. The outcome of this project suggests how social media chatter can be leveraged to find popular sentiment and preferences, which has important implications when consulting business decisions on innovation as product development, especially for a company like Inspire Brands.

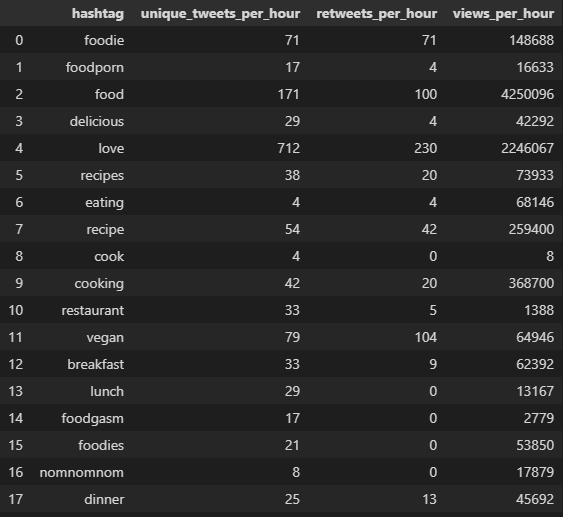
Project Limitations

The most pertinent augmentation to this capstone project would be to increase the amount of data captured: more data leads to better NLP analyses and modeling. Our data were limited by the redundancy in how the Twitter API grabs past tweets and the availability of the tweets that meet our query criteria, and the latter likely has an effect on the former - a more robust search could reduce the redundancy in the scrape. Moreover, a well-defined methodology for capturing both “trending” and “relevant” data would be a significant improvement to the methods employed in this capstone project. Finally, we would be able to show better and more numerous restaurant results if Documenu’s API returned the number of restaurants provided to their API should the aforementioned parser issue not exist.

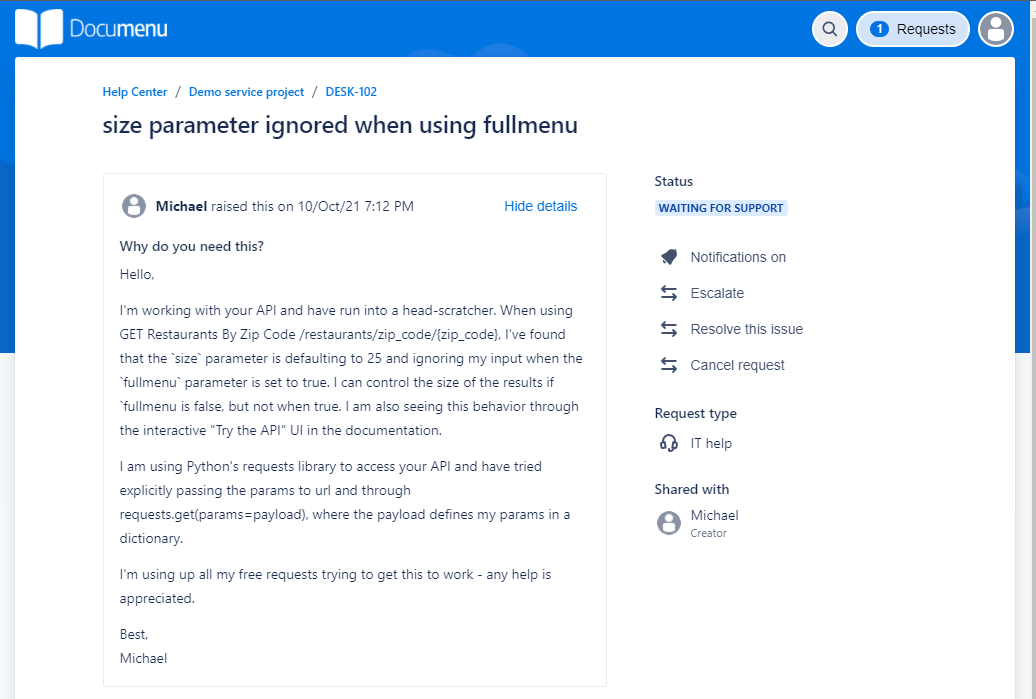
Future Research

This capstone presented results for one month’s worth of tweets aggregated into a single dataset and peeked into the results for a geographic area. One could alternatively gather meaningful insights from our methods on a temporal axis and understand the changes in trends over time. Additionally, a labeled corpus for supervised training could augment the time-series analysis, as well as the work we describe here. Further, since we are providing insights into what is essentially user preferences, developing user profiles to utilize in a recommendation model would tease out data-driven suggestions that could drive revenue by pointing users to items they may not know exist or that they are interested in, as well as serve targeted ads to those users.

**Appendix**



**Supplementary Figure 1:** Hashtag statistics obtained from RiteTag.



**Supplementary Figure 2:** JIRA ticket to address issue in Documenu’s API parser.

1. https://www.nrn.com/fast-casual/inspire-brands-ceo-paul-brown-accomplishes-historic-acquisition [↑](#footnote-ref-0)
2. https://medium.com/swlh/nlp-project-cuisine-classification-topic-modelling-cd7b4c734b19 [↑](#footnote-ref-1)
3. https://medium.com/@kellyzhangty/foodie-discovery-ed7586d384cc [↑](#footnote-ref-2)
4. https://www.tastewise.io [↑](#footnote-ref-3)
5. https://spoonshot.com [↑](#footnote-ref-4)
6. Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., & Torralba, A. (2021). RECIPE1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 187–203. [↑](#footnote-ref-5)
7. Harashima, J., Ariga, M., Murata, K., & Ioki, M. (2016). A Large-scale Recipe and Meal Data Collection as Infrastructure for Food Research. LREC. [↑](#footnote-ref-6)
8. Wiegand, M., Roth, B., & Klakow, D. (2014). Automatic food categorization from large unlabeled corpora and its impact on relation extraction. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. [↑](#footnote-ref-7)
9. Y. Pan, Q. Xu and Y. Li, "Food Recipe Alternation and Generation with Natural Language Processing Techniques," 2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW), 2020, pp. 94-97. [↑](#footnote-ref-8)
10. Annamoradnejad, I., & Habibi, J. (2019, April). A comprehensive analysis of twitter trending topics. In *2019 5th International Conference on Web Research (ICWR)* (pp. 22-27). IEEE. [↑](#footnote-ref-9)
11. https://app.ritetag.com/hashtag-suggestions [↑](#footnote-ref-10)
12. https://selenium-python.readthedocs.io/installation.html [↑](#footnote-ref-11)
13. https://www.enchantedlearning.com/wordlist/food.shtml [↑](#footnote-ref-12)
14. https://docs.tweepy.org/en/stable/ [↑](#footnote-ref-13)
15. https://documenu.com/ [↑](#footnote-ref-14)
16. https://documenu.com/docs#get\_restaurants\_by\_zip\_code [↑](#footnote-ref-15)
17. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis [↑](#footnote-ref-16)
18. https://github.com/vangheem/pyzipcode [↑](#footnote-ref-17)
19. https://www.nltk.org/ [↑](#footnote-ref-18)
20. https://github.com/s/preprocessor [↑](#footnote-ref-19)
21. https://spacy.io/ [↑](#footnote-ref-20)
22. D. Blei, A. Ng, and M. Jordan. “Latent Dirichlet allocation”. Journal of Machine Learning Research, 3:993–1022, January 2003. [↑](#footnote-ref-21)
23. https://radimrehurek.com/gensim/models/word2vec.html [↑](#footnote-ref-22)
24. https://arxiv.org/abs/1810.04805 [↑](#footnote-ref-23)
25. Reimers, N., &amp; Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese Bert-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [↑](#footnote-ref-24)
26. https://plotly.com/dash/ [↑](#footnote-ref-25)
27. https://en.wikipedia.org/wiki/Tf%E2%80%93idf [↑](#footnote-ref-26)