Baltic J. Modern Computing, Vol. 9 (2021), No. 2, 155-165
https://doi.org/10.22364/bjmc.2021.9.2.01

Utility of Large-Scale Recipe Data in Food Computing

Maija KĀLE[1], Ebenezer AGBOZO[2]

[1]Faculty of Computing, University of Latvia, 19 Raina Blvd., Riga, LV-1586, Latvia
[2]Ural Federal University, 19 Mira Str., 620002 Ekaterinburg, Russia

maijakale@gmail.com, eagbozo@urfu.ru

Abstract: This article aims to look at the recipe data analysis from a critical perspective, offering the authors' own learning experience from successes and failures of the research process. The present recipe research has been limited by the availability of data, which in the case of recipes mostly consists of texts depicting a variety of ingredients. This has contributed to a better understanding of flavour formation and nutritional value of food but has not led further to establishing a corpus of healthy and unhealthy foods. Time-related cooking aspects have remained largely out of the present research's scope due to the difficulties in obtaining immediately analyzable data. The same goes for the recipe-relate research on food texture, color and other aspects. In this research the methodology of topic modelling has been applied to analyze recipes in North American and Mexican cuisines in order to highlight the core culinary themes within these two cuisines. Potential for result analysis, as well as its limitations, are also discussed. Topic models of agglomerated data can be helpful in further multisensory research, as they provide some insights into the colour, the flavour and, potentially, the texture of certain groups of dishes. It can be combined further on with social media sentiment analysis and other research methods to better grasp the human relationship with food.

Keywords: Food Computing, Recipes, Topic Modelling, NLP, Healthy Food, Multisensory Research

1. Introduction

Recipes have long been among the most easily available massive data sources for the early analysis of food-related texts and, thus, have served as one of the building blocks in food computing, which aims to utilize large datasets for understanding the human relationship with food. Most of the analysis has focused on the number and combination of ingredients as the cooking process itself has not been that easily derivable from the textual data of recipes. Research has been focused mainly on the comparative analysis of various cuisines, national dishes and ingredient pairings to determine the flavour characteristics, looking at e.g., authentic ingredients, frequently used ingredients and alike (Ahn et al., 2011; Min et al., 2019B; Kim and Chung, 2016).

Ultimately, the research of large-scale data analysis of recipes has been based on an assumption that flavour matters and that flavour is determined by the combination of

ingredients. Ingredient pairings, flavour differences among various cuisines and the number of ingredients per dish have served as the baseline in the analytical discourse on recipes. Much of this research has been inspired by the availability of a vast amount of digital data and the willingness to utilize it. We aim to look at the utility of such an analysis from a theoretical standpoint, as well as to draw upon other research opportunities that could be derived from the current analysis of the recipe texts.

2. Recipe Data and Assumptions behind the Human Relationship with Food

Most of the recipe data analysts base their work on the assumption that recipes contain important information about the meal and that this information concerns flavour, nutritional and health aspects (Kusmierczyk and Nørvågg, 2016; Ahn et al., 2011; Min et al., 2019B; Herruzo et al., 2016). This assumption seems to be an obvious truth, however, an increasing number of studies in cognitive sciences reveals that 'everything else' pertaining to the food matters as well (Spence, 2017). In fact, the flavour or taste of a product might be subordinate to other more determining factors at play – such as context, social class and even 'food fashion'. Multisensory food research is on the rise (Obrist et al., 2019), which shows that the explanation of palatability based on a sole factor – such as flavour – is insufficient in terms of the human relationship with food.

Taking this into account, we may conclude that while recipes can be utilized, up to a certain degree, as the indicators of a food's flavour, doing so without a holistic approach that includes cognitive and social sciences as well as computing might turn out of a marginal benefit for the research community. Working with large-scale flavour-related data has proven to be complex, and most of the analytics has been done in a comparative framework looking at various cuisines in different geographical areas, since the data on flavour analysis per se does not provide any substantial clues for further understanding of the human and food relationship. A comparative geographical analysis proves, however, that there are differences between various cuisines when it comes to flavour and that there is a particular divide between the East and the West (Jurafsky, 2014). How to proceed with this knowledge in tow remains an open question though.

3. The Impossible Task of Defining Healthiness

While recipes provide us with rich amount of textual data on food ingredients, it has proven to be a difficult task to divide the foods into the categories of 'healthy' and 'unhealthy'. Recipes contain information about potentially 'good' and 'bad' ingredients – such as the amount of carbohydrates, salt and fats – and make it possible to determine the food's nutritional value, as has been proven by Kusmierczuk and Norvag (Kusmierczyk and Nørvåg, 2016). Nevertheless, when it comes to the meal as a whole it is impossible to determine if the food can be categorized as healthy or unhealthy, because those aspects are more contextual and dependent on the overall diet and lifestyle. Here the development of a personalized nutrition research could be a step in the right direction (Min et al., 2019).

Proceeding to the level of national cuisines, we see that the Mediterranean cuisine enjoys the status of a healthy cuisine. "Mediterranean diet is scientifically known as a healthy diet that helps to prevent main metabolic diseases. [..] For example, a growing

number of scientific researches has been demonstrating that olive oil, [sic] operates a crucial role on [sic] the prevention of cardiovascular and tumor-related diseases, being related with low mortality and morbidity in populations that tend to follow a Mediterranean diet" (Herruzo et al., 2016).

As can be guessed, the very definition of the Mediterranean diet heavily relies on one healthy ingredient – olive oil. Conversely, it has been difficult to determine those ingredients that harm human health as there is no single ingredient associated with specific diseases: "We attempted to see the relationships between the ingredients and diseases (diabetes, obesity, body mass index and so on). However, we found that it's not easy to identify single ingredient with strong impact on specific disease" (Jurafsky, 2014).

All in all, the definition of a healthy cuisine is not overtly specific, which makes us suggest that all recipe-related data be categorized under 'healthy food'. It turns out that home-cooked food by its very nature is healthier than the food that can be consumed in fast food restaurants – at least judging from the data showing "the correlation between the prevalence of fast food restaurants in a county to its obesity rate" (Mejova et al., 2015).

We can conclude that recipe data does not reveal much about the healthiness or unhealthiness of foods or even whole cuisines, as there is no way of meaningfully classifying recipes into 'healthy' or 'unhealthy'. Instead, the whole corpus of recipes could be viewed as a corpus of healthy food as compared to other eating contexts, such as fast food restaurants.

4. Lack of Time Perspective in Cooking Process Analysis

Deriving time-related data from recipes is difficult. We have not yet come across any large-scale data analysis of food with relation to its preparation time and e.g., the processes leading to changes in food temperature and texture. Time, however, would be an interesting aspect to examine as it might reveal something about the complexity of the whole process of food preparation. Complexity, in turn, could be viewed as a phenomenon possibly related to hedonic attitude to food (Kāle and Agbozo, 2020): "…complexity is commonly talked about as a desirable attribute of the consumer's experience of food and drink [..] while many western writers tend to want to put the complexity in the flavour, those from the East far more often situate the complexity in the art of making/preparing." (Spence, 2018).

Thus, time-related analysis of recipes could be of interest when tracing complexity in food preparation, but as the data does not provide easy access to time- and preparation-related factors we have to attribute the complexity to the food's ingredients.

Complexity of the cooking process can be well traced via the number of ingredients that a recipe contains. Curiously, the average number of ingredients per recipe turns out to be fairly constant across nations and continents.

"The histogram analysis shows that the distributions of the number of ingredients in recipes are rather similar (showing a normal distribution) across regions. [..] The average number of ingredients in a recipe ranged from 6.7~10.8" (Kim and Chung, 2016).

These findings might lead us to a broader philosophical discussion of human nature and the human relationship with food. Margot Finn develops such a discussion, looking

at theoretical considerations of human nature in her study about the role of social class in determining food taste preferences. She notes:

"Hume acknowledges that people may seem to have subjective, idiosyncratic preferences, but he argues that ultimately, we're all predisposed to enjoy the same things" (Ludington and Booker, 2019).

Potentially, the similarities in average ingredient numbers could mean that humans across nations and continents enjoy the same level of complexity, albeit enjoying different flavours, when it comes to the composition of ingredients. How complexity evolves in time is an interesting question but shall, however, be left for other future researchers to tackle, as this side-track remains out of the scope of the current research.

5. Large-Scale Data Analysis of Recipes

Taking into account the scepticism towards large-scale recipe data analysis, we want to showcase our own analysis that will illustrate both the limitations as described above, as well as potential future steps for further research and large-scale recipe data utilization.

Our first intention when looking at a particular cuisine – North American and Mexican – was to determine the number of healthy foods in each of them. We did not succeed though, as it was impossible to determine the healthiness of one particular food looking from a broader dish/meal perspective - for the reasons described above.

Our second intention was to do a comparative analysis of the topic models of various dishes "as John R. Firth famously put it in 1957, "a word is characterized by the company it keeps"" (Schöch, 2017). We aimed to compare the topic models of dishes, such as salads and desserts, and see how the "company of words they keep" differed within the framework of a cuisine – we chose the North American.

Third, we wanted to look at both cuisines, North American and Mexican, and discern the differences in terms of the ingredients used and the nouns describing them, thus trying to develop a meta-recipe for a meal and guess its alleged texture.

For the above purpose we applied the topic model analytical framework as described below.

6. Methodology

The methodology of the research is based on topic modelling, which is an unsupervised model learning a set of underlying topics. It was deemed the best choice for our study because it enables researchers to draw conclusions from word distributions in relation to topics, which is then followed by a close examination of documents on specific topics for a more in-depth analysis (Nikolenko et al., 2017). We employed the topic model known as Latent Dirichlet Allocation (LDA), which, despite its inability to model correlation between topics, is known to be a true generative probabilistic model without the drawbacks of the earlier topic models (Uys et al., 2008).

The initial dataset derived from the Yummly-66k dataset[1] (consisting of 66,615 recipes from ten cuisines on Yummly - a website generating personalized recipes and serving as a search engine for recipes) repository of cross-regional food recipes prepared by Min et al. (Min et al., 2018) was deemed suitable for answering our research

[1] http://isia.ict.ac.cn/dataset/Yummly-66K.html

questions (Min et al., 2019B). The dataset spanned ten national cuisines while our study focused on North American and Mexican cuisines. The original dataset files were in the JSON format, which upon extraction were converted to the CSV format and uploaded into the RStudio software for analysis.

The next phase involved splitting the dataset into training (20%) and testing (80%) in order to reduce the runtime while building the model. For the cleaning process (initial and main) the following steps were taken:

i. duplicate rows with repeating cuisines were removed;
ii. using a dictionary of common terms (e.g. 'bake', 'fry', 'toast', and 'whisk') compiled by the authors, kitchen terminologies and measurement units (e.g. 'ounce', 'tbsp', 'temperature', and 'pinch') were removed;
iii. unnecessary whitespaces and English stop words (e.g. 'am', 'the', 'is', 'will', 'in', 'down', and 'above') were filtered out;
iv. using the Regular Expression (regex), all numbers and punctuation marks except for hyphens were removed:
    a. [^-a-zA-Z[:space:]]+
    b. [0-9]
v. Using SQL queries, terms with less than two characters were removed.

We approached this looking at the number of nouns and adjectives, which signals the richness of the text describing the food. According to Martin & Johnson (Martin and Johnson, 2015), semantic coherence of topics is improved by focusing on a particular part-of-speech. In their own study they focused on nouns, whereas our work focuses on adjectives and their modification and the descriptive attributes of food ingredients in the cuisines of our study's choice.

The topic model function was then built based on the LDA (Latent Dirichlet Allocation) model to obtain the top ten (10) models from the corpus. LDA is a probabilistic model that explains word co-appearances in text arising from a relatively small number of possible semantic groups or topics (Ihler and Newman, 2011). LDA has been applied by researchers for classifying documents and uncovering patterns and trends – for example, in the medical field (Tran et al., 2019). A second model for extracting frequent adjectives from the cuisine corpus was built after cleaning the dataset and filtering the lemma using the part-of-speech (POS) tagging model (udpipe developed by Straka et al. (Straka et al., 2016)). In their work, Straka and Straková (Straka and Straková, 2017) intricately discuss the inner workings of the tokenization, lemmatization and POS tagging with the udpipe library. The udpipe model is used for tagging POS in relation to all tokens. Upon obtaining the results, the output was parsed into functions in order to discern the frequent adjective terms. Finally, the test dataset was run conjointly with the train dataset, in order to obtain the full picture.

7. Description of Results

The x-axis signifies the beta (β), which denotes the topic-word density/probability/concentration (i.e., distribution of words per topic) (Jelodar et al., 2019). According to Blei et al. (Blei et al., 2003): (1) A high β-value indicates that each topic is likely to contain a mixture of most of the words and not any word specifically, while a low value means that a topic may contain a mixture of just a few of the words;

(2) A high β-value also means that each topic is more likely to contain a specific word mix defined by the base measure; (3) Finally, a high β-value will lead to topics being more similar in terms of what words they contain. Beta (β) is represented by the formula:

$$\beta = P(\text{tokens} \mid \text{topics})$$

The beta value formula is representative of per-topic-per-word or per-topic-per-token probabilities from the topic model. Thus, a low beta is synonymous with the presence of fewer prevalent words (tokens).

We made analysis for 13 foods/dishes of American and 14 of Mexican cuisine, including afternoon tea, soups, main courses, snacks etc. We looked at the topic analysis in terms of adjectives and nouns. For each of the meals and word categories (nouns vs adjectives) we made a topic model analysis using ten topic models, thus following the rationale of the average amount of ingredients per recipe. All topic models for adjectives and nouns for both cuisines are available in the "Visualization" folder of our Github repository[2]. For simplicity, we used only several of the dishes with the topic models of adjectives and nouns to illustrate our research inquiries, namely, salads and desserts of North American cuisine.

Looking at the topic models associated with salads (Figure 1) and desserts (Figure 2), we can see that the adjective 'fresh' dominates regarding both dishes. Colours such as red, black and green go together in American salads – potentially indicating the colour of spices used. Desserts meanwhile tend to be described by adjectives such as 'pure', 'golden' and 'white'. Whether ingredient colours might be used in future research in order to determine the food's potential texture remains to be seen. Certainly, unlocking the potential of adjectives used in recipes could provide us with more data on the dominant colours in dishes as well as their texture, thus leading to better understanding of other factors behind a recipe.

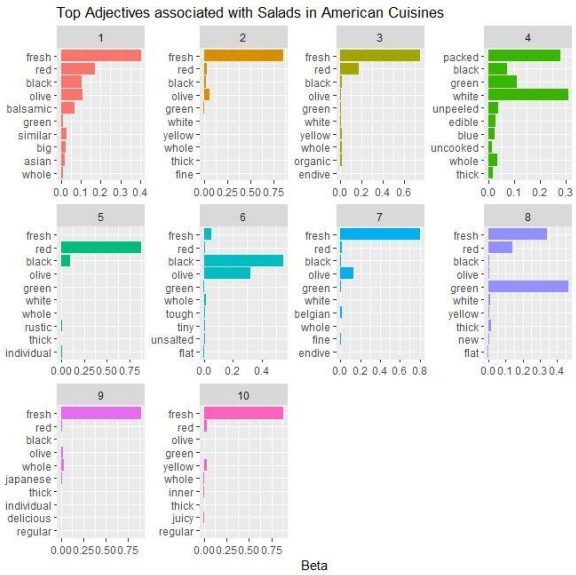

Figure 1: Topic Models for Adjectives Related to Salads in North American Cuisine

[2] https://github.com/agbozo1/foodComputing

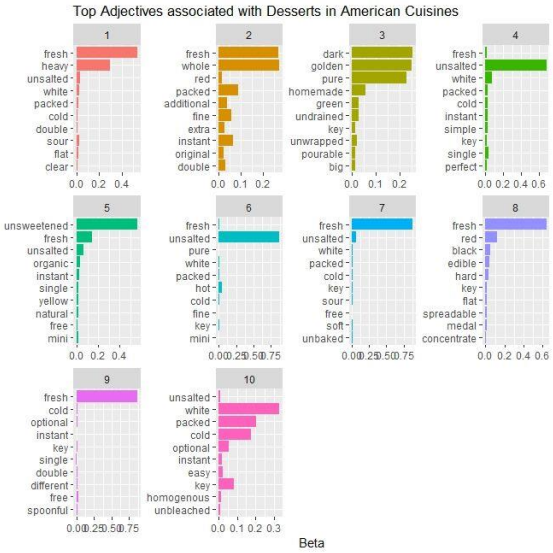

Figure 2: Topic Models for Adjectives Related to Desserts in North American Cuisine

Looking at the results generated regarding salads and desserts in American cuisine (Figure 3 and 4), the first thing to be noticed is that the noun 'salad' does not appear overly frequently and is outnumbered by the sauce ingredients such as 'vinegar', 'pepper' and 'lemon'.

Meanwhile, 'sugar', 'cream', 'vanilla', 'eggs' and 'chocolate' (not surprisingly) dominate the dessert topic models. Again, there is a potential for deriving the food's texture based on such supplements as vinegar or eggs. As Ahn et al. concludes, "there are many ingredients whose main role in a recipe may not be only flavouring but something else as well (e.g., eggs' role to ensure mechanical stability or paprika's role to add vivid colours)" (Ahn et al., 2011).

We acknowledge that visualization plays an important role in topic modelling and it can shape the way we look at the food and perceive opportunities for data analysis. Arguably, the best example e.g., in feminist scholarship in this respect is Signs@40: Feminist Scholarship through Four Decades (http://signsat40.signsjournal.org/topic-model/). For the convenience of visualization, when comparing American and Mexican cuisines we turned the topic models into wordles. As Figure 5 and Figure 6 illustrate, these wordles of American and Mexican main courses could be viewed as a kind of meta-recipe consisting purely of ingredients and leaving out time, preparation and other components. This has been the unit of analysis that most of the recipe papers have looked at.

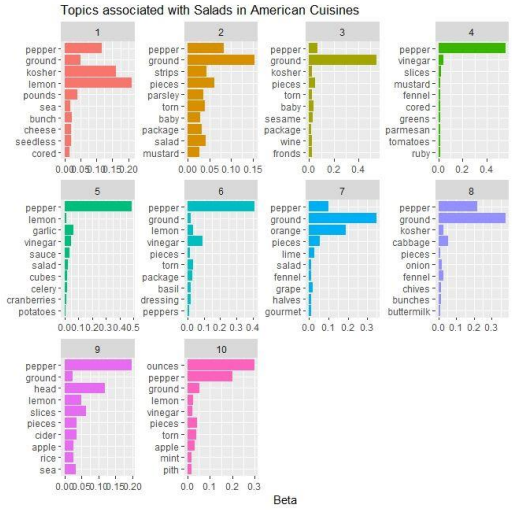

Figure 3: Topic Models for Nouns Related to Salads in North American Cuisine

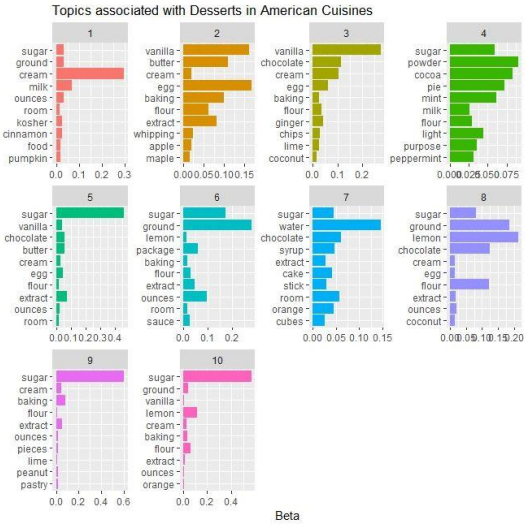

Figure 4: Topic Models for Nouns Related to Desserts in North American Cuisine

Figure 5: Wordle of Main Courses, American: Nouns

Figure 6: Wordle of Main Courses, Mexican: Nouns

What is the utility of this data? As mentioned earlier, we are not able to determine healthiness of the food. Instead, we can try to distinguish a set of flavours that shape the cuisine and its nutritional value. We can also look upon the possibilities of determining the food's texture, which would be of help in the multisensory research domain. In order to offer a full-fledged view of the texture or meta-texture of dishes in a certain cuisine, we would need to classify ingredients according to their texture first. Taking into account that texture might change during the cooking process, the analytics of time and preparation should be integrated here.

When looking at the recipe data, we should also consider the actual habits of eating the specific meals. As an example, breakfasts instead of being prepared at home might be replaced by ready-made cereals. Here, a large-scale data analysis might be useful if combined with the actual knowledge of consumption patterns regarding specific meals.

In addition to that, knowledge of certain dominating ingredients can be helpful when choosing certain keywords for social media search. Mejova et al. look at combination of various hashtags associated with #foodporn to determine healthy and unhealthy food representations in social media (Mejova et al., 2015). Knowing the most frequent ingredients of certain dishes can help refine the search of social media content related to certain dishes and cuisines.

8. Conclusions

To conclude, large-scale recipe data analysis does provide certain but limited understanding of the human relationship with food. There is a potential for large-scale recipe data to be used in combination with additional classification of ingredients (e.g., according to texture, timeline and food preparation), social media analysis (determining search inquiries for food representation and sentiments associated with it) and actual consumption habits (e.g., which meals are prepared at home and which are based on ready-made produce).

Social media analysis is still underused when it comes to food and large-scale recipe data, however, it can be useful in combination with food-related social media data. Thus, for example, insight into recipe ingredients can help form a better understanding of which hashtags should be used for social media search inquiries. Based on recipe data, one can choose the most used ingredients to search for social media response, followed by the social media sentiment analysis.

While large-scale recipe data can provide a certain added value for food computing researchers, the assumptions used for analysing human relationship with food must be

constantly refined, lest the oversimplified views on the importance of flavour and palatability become entrenched.

The field of food computing is rich and promising. The new emerging methodologies for analysing text- and image-based food data can advance into personalized food models, ultimately condensing the data for a targeted benefit of an individual's health.

References

Ahn Y.Y., Ahnert, S. E., Bagrow, J. P., Barabási. A. (2011). Flavor network and the principles of food pairing. In: *Sci Rep* **1**, 1 (December 2011), 196. DOI: https://doi.org/10.1038/srep00196

Blei, D. M., Ng, A. Y., Jordan, M.I. (2003). Latent dirichlet allocation. In: *Journal of machine Learning research*, **3**(Jan), 993-1022.

Herruzo, P., Bolaños, M., Radeva, P. (2016). Can a CNN Recognize Catalan Diet? arXiv:1607.08811 [cs] (2016), 020002. DOI: https://doi.org/10.1063/1.4964956

Ihler, A., Newman, D. (2011). Understanding errors in approximate distributed latent Dirichlet allocation. In: *IEEE Transactions on Knowledge and Data Engineering*, **24**(5), 952-960.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li Y., Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. In: *Multimedia Tools and Applications*, **78**(11), 15169-15211.

Jurafsky, D. (2014). *The Language of Food: A Linguist Reads the Menu*. W.W. Norton & Company.

Kāle, M., Agbozo, E. (2020). Tracing Complexity in Food Blogging Entries. In: *arXiv:2007.05552* [cs] (July, 2020).

Kim, K., Chung, C. (2016). Tell Me What You Eat, and I Will Tell You Where You Come From: A Data Science Approach for Global Recipe Data on the Web. In: *IEEE Access* **4**, (2016), 8199–8211. DOI: https://doi.org/10.1109/ACCESS.2016.2600699

Kusmierczyk, T., Nørvåg, K. (2016). Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ACM, Indianapolis Indiana USA, 2013–2016. DOI: https://doi.org/10.1145/2983323.2983897

Ludington, C. C., Booker, M.M. (2019). *Food Fights: How History Matters to Contemporary Food Debates.* University of North Carolina Press, Chapel Hill

Martin, F., Johnson, M. (2015). More efficient topic modelling through a noun only approach. In: *Proceedings of the Australasian Language Technology Association Workshop 2015* (December, 2015), 111-115.

Mejova, Y., Haddadi, H., Noulas, A., Weber, I. (2015). #FoodPorn: Obesity Patterns in Culinary Interactions. In: *arXiv*:*1503.01546* [cs] (March, 2015).

Min, W., Bao, B., Mei, S. Zhu, Y., Rui, Y., Jiang, S. (2018). You Are What You Eat: Exploring Rich Recipe Information for Cross-Region Food Analysis, *IEEE Transactions on Multimedia*, **20**(4). DOI: https://doi.org/10.1109/TMfinM.2017.2759499

Min, W., Jiang, S., Jain, R. (2019). Food Recommendation: Framework, Existing Solutions and Challenges. arXiv:1905.06269 [cs] (November, 2019).

Min, W., Jiang, S., Liu, L., Rui, Y., Jain, R. (2019B). A survey on food computing. In: *ACM Computing Surveys (CSUR)*, **52**(5), 92.

Nikolenko, S., Koltcov, S., Koltsova, O. (2017). Topic modelling for qualitative studies. In: *Journal of Information Science*, **43**(1), 88-102.

Obrist, M., Tu, Y., Yao, L., Velasco, C. (2019). Space Food Experiences: Designing Passenger's Eating Experiences for Future Space Travel Scenarios. In: *Front. Comput. Sci.* **1**, DOI: https://doi.org/10.3389/fcomp.2019.00003

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: *DHQ 011*, **2** (May, 2017).

Straka, M., Hajic, J., Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In: *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (May, 2016) 4290-4297.

Straka, M., Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: *Proceedings of the CoNLL* 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (August, 2017). 88-99.

Spence, C. (2017). *Gastrophysics: The new science of eating.* Penguin, London, UK.

Spence, C. (2018). Complexity on the Menu and in the Meal.In: *Foods 7*, 10 (September, 2018). DOI: https://doi.org/10.3390/foods7100158

Tran, B.X., Nghiem, S., Sahin, O., Vu, T.M., Ha, G.H., Vu, G.T., Ho, C.S. (2019). Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study. In: *Journal of medical Internet research*, **21**(11), e15511.s

Uys, J. W., Du Preez, N.D., Uys, E.W. (2008). Leveraging unstructured information using topic modelling. In: PICMET'08-2008 Portland International Conference on Management of Engineering & Technology, (July, 2008) 955-961. IEEE.

Received September 15, 2020, revised February 25, 2021, accepted April 12, 2021