Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics

Clustering and Augmenting Recipes

Team Member: Tom Rix, 3307600, M. Sc. Applied Computer Science
rix@stud.uni-heidelberg.de
Team Member: Maximilian Jalea, 3256466, M. Sc. Scientific Computing
jalea@stud.uni-heidelberg.de
Team Member: Christan Heusel, 4020794, B. Sc. Applied Computer Science
c.heusel@stud.uni-heidelberg.de
Team Member: Henrik Reinstädtler, 3307518, M. Sc. Scientific Computing
reinstaedtler@stud.uni-heidelberg.de

Find this project on GitHub:
`https://github.com/christian-heusel/ITA_WS_2020`

1 Introduction

When was the last time you looked into your recipes collection and thought: '*Oh, what a mess, why is there no structure in this collection?*' Well, most of us tend to collect interesting recipes from various different sources – may it be old recipes from grandma or fancy modern, low-carb recipes from lifestyle magazines. Usually, we file all our recipes away in a folder or even digitalize them and throw them into virtual pile of recipes.

But wouldn't it be nice, if there was a structure in our recipes collection? All pasta recipes in one chapter, all starters in a chapter separate from deserts and even a categorization into different cultural cuisines? In our project we aim to solve this everyday-life problem by applying text analytics methods to recipe datasets. For example, by clustering recipes into meaningful categories similar recipes can be detected and grouped together. One goal is to find out if recipes can be automatically separated into groups like:

- type of food: e. g. starters, mains, deserts, pastry and drinks,
- cultural origin of the cuisine: e. g. all Mediterranean dishes appear in the same cluster whereas Asian food is in another cluster far apart. Analyzing even sub-clusters would be interesting to find out how closely the different cuisines are related to each other.

Further ideas for real-world scenarios are the estimation of a healthiness score and the prediction of a dish's preparation time based on the recipe, similar to reading time estimations on websites. Additionally, analyzing the comments of recipes, if available, with sentiment analysis methods would be interesting for generating ratings automatically. Choosing a suitable recipe by interpolating between a set of given recipes in latent space might be another task that would be useful in a real-world context. As there is barely anything more fundamental than food, all text analytics that can be performed on recipe data have real-world applications.

In the end, an intelligent recipe collection shall be offered, where newly added recipes will automatically be sorted into a suitable position. Thus, one can nicely browse through all items and quickly find a suitable recipe in order to indulge oneself.

An overview over related work is given in section 2, including not only domain specific articles but also text analytics tasks in general. In section 3 our planned project is described in detail and a project roadmap as well as tasks and objectives are presented.

2 Research Topic Summary

Astonishingly, there is only a small research community which analyzes recipes, given their ubiquity and that understanding recipes is a crucial everyday skill. Perhaps, this is due to the fact that recipes themselves aren't economically interesting, but only the ingredients and advertising have potential for commercialisation.

The problems that are tackled and the issues that are tried to be solved in relation to recipes vary a lot as the text analytics tasks depend especially on the availability of additional information about the recipes. If labels for type of dish/course and cuisine are available, the dataset is suitable for classification [1, 2]. In their paper Su et al. applied associative classification and used support vector machines. However, most datasets don't provide labels for all recipes. However, performing an unsupervised clustering of recipes would still be a possibility to gain insights into the similarity and relation of recipes. In his article B. Sturm presents an examination of cuisines through unsupervised learning [3], in particular Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA). B. H. Tan performed k-means clustering [4] on the German Recipes Dataset [5]. To the best of our knowledge, there is no published and peer-reviewed paper about clustering of recipes yet.

There are several interesting publications that solve similar or related problems dealing with recipes: Shidochi et al. tried to automatically find replaceable materials in recipes considering characteristic cooking actions [6]. Therefore, a detailed analysis and understanding of the recipes was inevitable. Hanai et al. developed a method to detect spam recipes based on their similarity using a so called Surprising Degree-Recipe Frequency-Interted Ingredient Frequency (S-RT-IIF) [7]. J. Jermsurawong and N. Habash tried to extract a tree-like structure of instructions from recipes that models the dependency of steps upon each other [8]. Min et al. [9] as well as Marin et al. [10] go even further and analyse not only text data but multi-modal data including images of the final dishes. There are several papers about the conversion from images to corresponding recipes and vice-versa [10, 11]. Majumder et al. generated personalized recipe suggestions based on user preferences and selected ingredients [12].

Apart from recipes, text clustering is a common task for what different methods were developed. In their review paper N. Allahyari et al. [13] categorize these methods into hierarchical clustering, k-means clustering as well as probabilistic clustering and topic models. Applying these methods to recipe datasets in particular will be the scientific contribution of our project.

3 Project Description

Main Project Goals Our main goal is to create a automatic clustering / categorization of recipes that can be visualized as map. Additionally, it could be interesting to calculate similarity in a cluster or even summarize the given recipes, if they fit. Furthermore, we are interested in augmenting recipes with additional information like difficulty level or preparation time automatically or recommending recipes based on left over ingredients.

Text Analytics Tasks On the one hand we aim to perform a clustering task, on the other hand the augmentation is somehow a sentiment analysis and finding suitable recipes is an information retrieval task.

Pipeline The data we found is already structured, however in different ways. An overview of our planned pipeline can be found in Figure 1. Most of the sources consist of multiple recipes, nicely separated in different documents. Sometimes, the recipes contain sub-recipes, that are not linked. We might purge them from our datasets, because they increase the complexity. Conversely, short recipes might also be removed, as they do not contain enough data, e. g. we are not interested in the simple instructions on heating up a frozen pizza. This cleaning can be done heuristically based on the length of the recipes. From the recipes the ingredients and instructions need to be separated. Abbreviations, like tbsp = tablespoon, need to be normalized. After stop word removal and tokenization, a part of speech analysis will take place to later on distinguish between instructions that are usually verbs, ingredients and utensils which are nouns and adjectives to processed ingredients. It might be useful to apply some stemming and lemmatization in order to reduce the complexity of our problem.

Used Datasets There are several large datasets of recipes from different websites so that we might not have to crawl our on recipe dataset. We found the following datasets:

- Recipes1M+ dataset [10] containing 1,029,720 recipes from multiple cooking websites with corresponding analysis [11]
- German Recipes Dataset [5] containing 12,190 recipes in German language from `chefkoch.de`
- `Food.com` dataset containing 180K+ recipes with 700K+ interactions [14] with notebook on sentiment analysis, most / least favourite ingredients

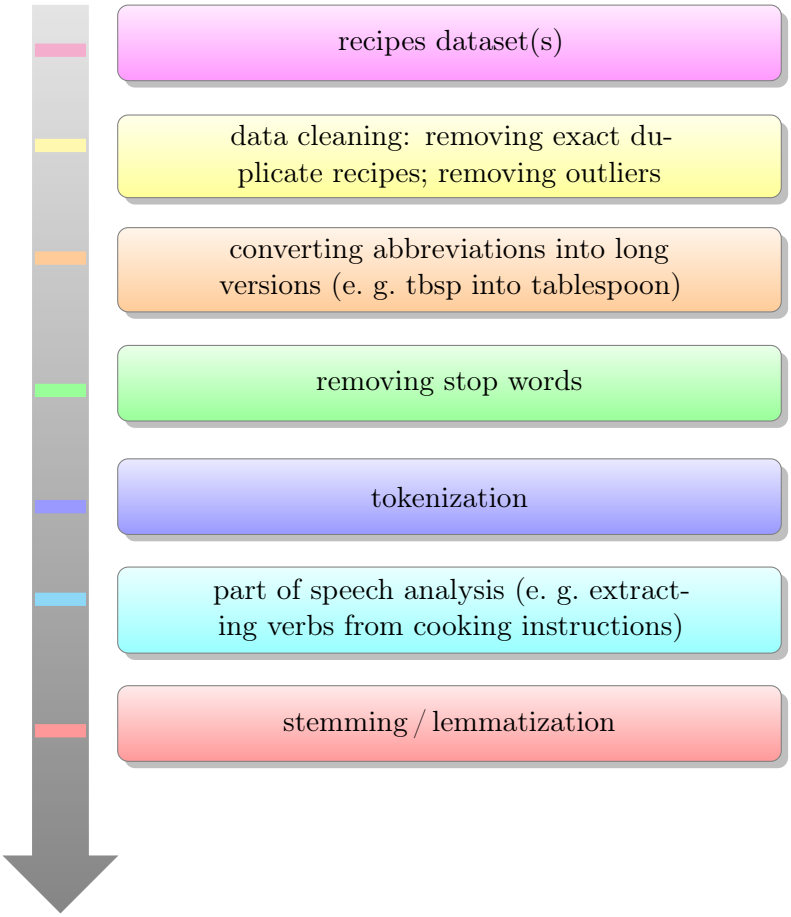

Figure 1: Pipeline: From entire recipes to pre-processed features that can be encoded into in matrices and processed for analysis purposed.

- What's cooking dataset [15] containing recipes from `yummly.com` with cultural diffusion analysis [2]
- Recipe box dataset [16] with $\tilde{1}$25,000 recipes from various food websites
- Epicurious: 20K+ recipes with Rating and Nutrition from Epicurious [17]

All of them contain for each recipe a title, a list of ingredients with measurements and preparation instructions. Some datasets include additional information like cuisine, url to website, an image, comments on the recipe, etc. We plan to compare our developed methods on several datasets to evaluate their generalization performance. So for example, training on a dataset with labeled data and then evaluating on others where no labels are available

makes sense. One difficulty might be that recipes are written in different languages. So generalization only works within the same language. Also the quality of the recipe data might vary a lot.

Evaluation & Baselines As there aren't many labels available, a manual evaluation and qualitative analysis of the results needs to be performed. For clustering one can visually see, if it worked and how the different categories are spread. The baseline to compare against would be our human knowledge about recipes.

3.1 Project roadmap

Our planned milestones are the following:

1. Data Inspection: Detailed analysis of the different datasets to get an understanding of the data quality and their distribution. Especially detecting abnormalities that need to be adjusted in consecutive work might be valuable.

2. Constructing a pre-processing pipeline as outlined above.

3. Implementing different methods to cluster recipes. In parallel, methods to predict additional information about the recipes like its cuisine association, its preparation time or healthiness score are developed.

4. Each method's performance will be evaluated. Ultimately all tasks will be combined into an intelligent recipes system.

3.2 Subtasks & Objectives

After inspecting our datasets and successfully constructing our preprocessing pipeline, we might split up and focus on different clustering methods (e. g. PCA, k–means, hierarchical clustering, probabilistic clustering) as well as topic modeling individually.

We plan to utilize an agile approach to conduct our project. By meeting regularly and presenting our results to each other we hope to obtain fruitful results by utilizing our different scientific backgrounds.

References

[1] Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang. Automatic recipe cuisine classification by ingredients. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, page 565–570, New York, NY, USA, 2014. Association for Computing Machinery. URL https://doi.org/10.1145/2638728.2641335.

[2] Alona Levy. Cultural diffusion on the what's cooking dataset. Kaggle Python Notebook. URL https://www.kaggle.com/alonalevy/cultural-diffusion-by-recipes.

[3] Ben Sturm. An examination of international cuisines through unsupervised learning. Blog Post on towards data science. URL https://towardsdatascience.com/an-examination-of-international-cuisines-through-unsupervised-learning-93c8b56d1ea0.

[4] Boon Hau Tan. Clustering recipes. Kaggle Python notebook. URL https://www.kaggle.com/tboonhau/clustering-recipes.

[5] Tobias (Sterby) Sterbak. German recipes dataset. Kaggle Dataset. URL https://www.kaggle.com/sterby/german-recipes-dataset.

[6] Yuka Shidochi, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. In *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities*, CEA '09, page 9–14, New York, NY, USA, 2009. Association for Computing Machinery. URL https://doi.org/10.1145/1630995.1630998.

[7] Shunsuke Hanai, Hidetsugu Nanba, and Akiyo Nadamoto. Clustering for closely similar recipes to extract spam recipes in user-generated recipe sites. In *Proceedings of the 17th International Conference on Information Integration and Web-Based Applications and Services*, iiWAS '15, New York, NY, USA, 2015. Association for Computing Machinery. URL https://doi.org/10.1145/2837185.2837269.

[8] Jermsak Jermsurawong and Nizar Habash. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 781–786, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1090. URL https://www.aclweb.org/anthology/D15-1090.

[9] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. A delicious recipe analysis framework for exploring multi-modal recipes with various attributes. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 402–410, New York, NY, USA, 2017. Association for Computing Machinery. URL https://doi.org/10.1145/3123266.3123272.

[10] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images, 2019.

[11] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3068–3076, 2017. doi: 10.1109/CVPR.2017.327.

[12] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5976–5982, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1613. URL https://www.aclweb.org/anthology/D19-1613.

[13] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *CoRR*, abs/1707.02919, 2017. URL http://arxiv.org/abs/1707.02919.

[14] Shuyang Li et al. Food.com recipes and interactions. Kaggle Dataset. URL https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions.

[15] Kaggle. What's cooking? Kaggle Dataset. URL https://www.kaggle.com/c/whats-cooking/overview.

[16] Ryan Lee. Recipe box. Blog Post. URL https://eightportions.com/datasets/Recipes/#fn:1.

[17] Hugo Darwood. Epicurious – recipes with rating and nutrition recipes from epicurious by rating, nutritional content, and categories. Kaggle Dataset. URL https://www.kaggle.com/hugodarwood/epirecipes.