Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics

Clustering and Enriching 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:

- types 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 automatically generating ratings. As there is barely anything more fundamental than food, all text analytics that can be performed on recipe data has real-world applications.

Could it be an idea also to do a search in the cluster over different recipes as an interpolation of them?

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.

Am Schluss sollten wir einen Ueberblick über die Struktur dieses Dokumentes geben. Also kurz zwei Saetze, welche Infos man wo findet.

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.

So far, the problems that are tackled and the issues that are tried to be solved in relation to recipes vary a lot. 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 this 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 (https://dl.acm.org/doi/pdf/10.1145/1630995.1630998). Therefore, a detailed analysis and understanding of the recipes was inevitable. 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. W. Min et al [6] as well as [7] 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 [7, 8]. Majumder et al. generated personalized recipe suggestions based on user preferences and selected ingredients [9].

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

3 Project Description

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 enriching 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 enrichment is somehow a sentiment analysis.

Pipeline The data we found is already structured in different ways. Most of the sources consists of multiple recipes, nicely seperated in different documents. Sometimes the recipes contain subrecipes, that are not linked. We might purge them from our datasets, because they increase the complexity. Conversly, short recipes might also be removed, as they do not contain enough data, e.g. we are not interested in the instructions on heating up a pizza. This cleaning can be done heuristically based on the length of the recipes. From the recipes the ingredients and instructions need to be seperated.Abbreviations,like tbsps=tablespoons, need to be normalized. After tokenizing a part of speech analysis will take place to distinguish later between instructions and adjectives to processed ingredients. It might be useful to apply some stemming and remove stop words 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 [7] with corresponding analysis [8]
- German Recipes Dataset [5]
- Food.com Dataset with Interactions [10] with notebook on sentiment analysis, most/least favourite ingredients
- What's cooking dataset `https://www.kaggle.com/c/whats-cooking/overview`
- Recipe box dataset [11]
- epirecipes

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, ... 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.

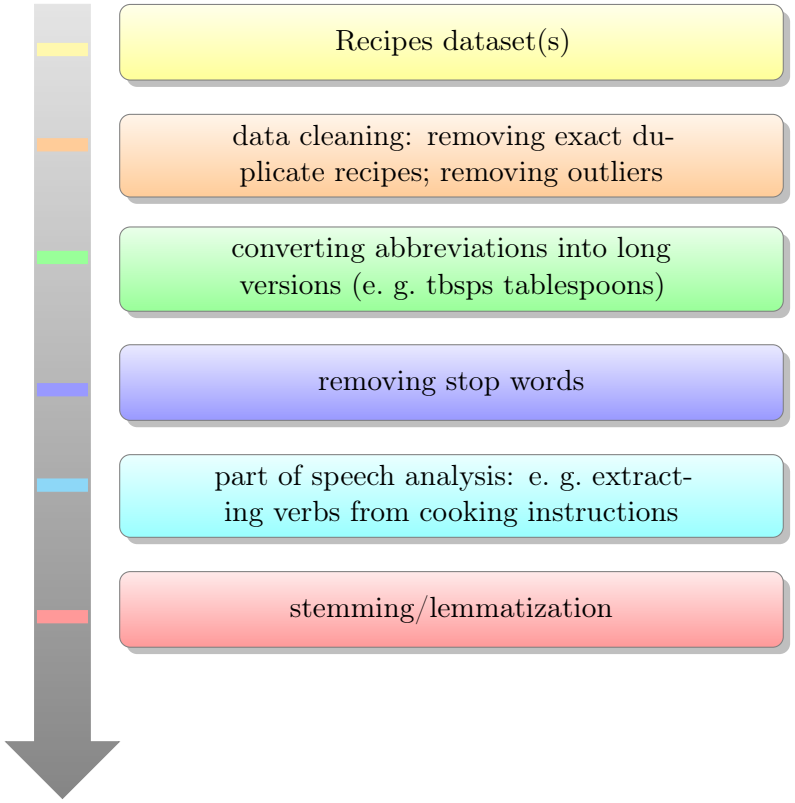

Figure 1: Pipeline: From entire recipes to features in matrices...

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 its distribution. Especially detecting abnormalities that need to be adjusted in consecutive work might be valuable.

2. Constructing 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.

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. ISBN 9781450330473. doi: 10.1145/2638728.2641335. 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-93c8b

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

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

[6] 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. ISBN 9781450349062. doi: 10.1145/3123266.3123272. URL https://doi.org/10.1145/3123266.3123272.

[7] 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.

[8] 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.

[9] 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.

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

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