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 Sciences rix@stud.uni-heidelberg.de
Team Member:	Name, Matriculation Number, Course of Study email address
Team Member:	Name, Matriculation Number, Course of Study email address
Team Member:	Very long Name, Matriculation Number Course of Study, email address

Find this project on GitHub: `www.github.io`

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, pastryand 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.

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.

2 Research Topic Summary

Astonishingly, there is a small research community which analyzes recipes. However, the problems that are tackled and the issues that are tried to be solved vary a lot.

Things to cite (datasets, applications, methods):

- recipes1M+ dataset [2] with corresponding analysis [3]
- German Recipes Dataset https://www.kaggle.com/sterby/german-recipes-dataset
- Recipe generation [1]
- Food.com Dataset with Interactions [?] with notebook on sentiment analysis, most/least favourite ingredients
- Cultural diffusion: https://www.kaggle.com/alonalevy/cultural-diffusion-by-recip on the What's cooking dataset https://www.kaggle.com/c/whats-cooking/overview
- Recipe box dataset https://eightportions.com/datasets/Recipes/#fn:1
- Clustering Recipes https://www.kaggle.com/tboonhau/clustering-recipes

definitely a more structured literature review needs to be done here

3 Project Description

Main project goals: Create a meaningful clustering/categorization of recipes and enrich plain recipes with useful additional information.

Text Analytics Tasks: Clustering, sentiment analysis, ...

Pipeline: Parsing, Tokenization, stemming, ... How to split up ingredients and instructoins. Extract verbs from instructions might be useful as they are good indicators of what is happening. In Figure 1 all steps are presented in detail.

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:

1. recipes1M+
2. German recipes dataset
3. epirecipes
4. what's cooking dataset
5. food.com recipes and interactions

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

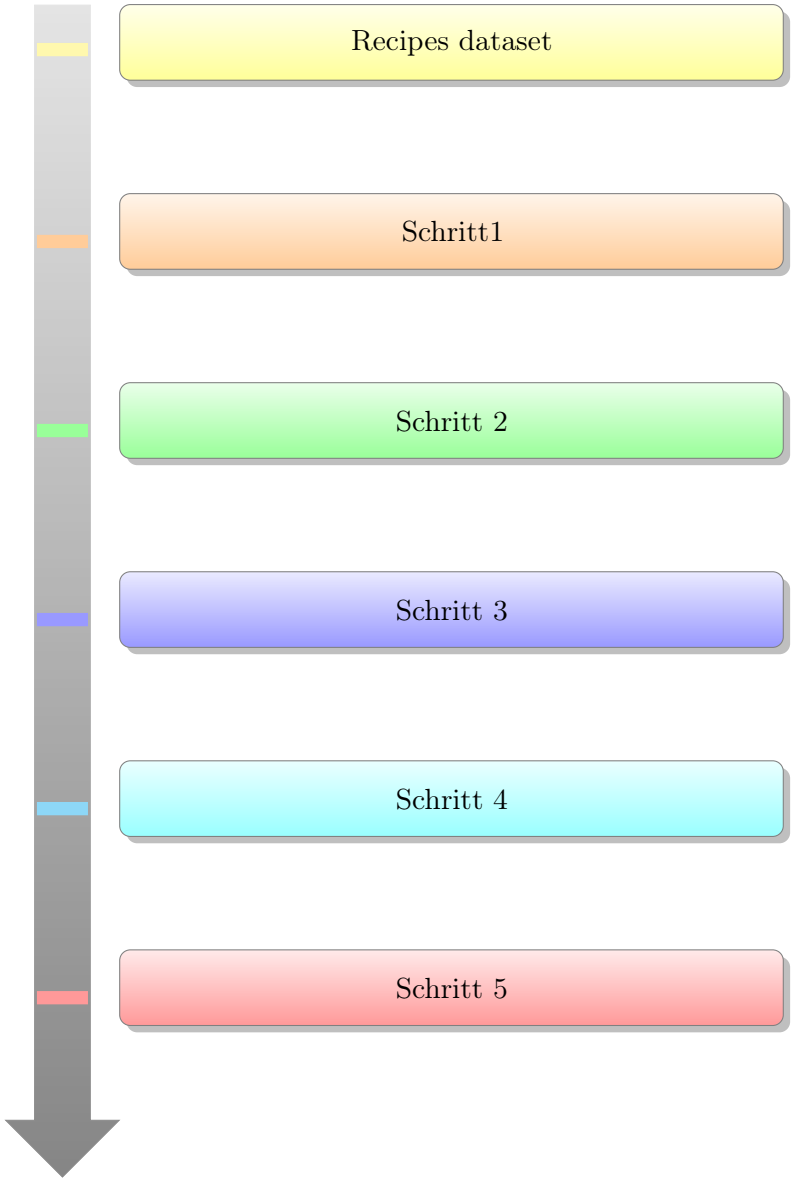

Figure 1: Pipeline: From entire recipes to features in matrices...

languages. So generalization only works within the same language. Also the quality of the recipe data might vary a lot.

Evaluation: 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.

Baselines: Do we know any baselines for our tasks?

References

[1] 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.

[2] 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.

[3] 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.