**NLP Write-up**

**Ed Lee**

DSML, Aug 2022

**Abstract**

Sometimes it becomes a chore to decide what to eat, be it cooking or eating out. In such situation, most people start off with some general, vague idea for a food. Sometimes the idea can be main ingredients (meat, veggie, etc), or a broad categories (baked, soup, fried, etc). My aim was to collect various recipes and build a solid set of topics that can characterize such idea. With the topics at hand, I can utilize it to create a recommendation system for suggesting possible foods that matches what one might be thinking to eat.

Throughout the project I have secured a dataset of recipes for various types of foods, and then processed the dataset so it can be trained via unsupervised learning method of choice. As a result, the model yielded considerably well characterized topics for the recipes, with EDA and clustering analysis suggesting further tuning is desirable in order to gain even more distinctive topics. Although I didn’t manage to implement a recommendation model, the quality of topic models leave a good hope for one in a future works.

**Design**

First, the raw dataset, obtained from kaggle, was processed into a type of term-document matrix. The data, then, was fed into a couple different dimensionality reduction methods, NMF and LDA, in order to build topic models. The resulting topics for each method were compared in order to select a topic model that is more interpretable. Then the chosen topic model was analyzed in order to further evaluate it in more detail. Throughout the modeling process hyperparameters were mainly tuned manually via inspecting the resulting topic models for each combination of the parameters.

**Data**

Dataset contains various information of recipes scraped from food.com, and was obtained from kaggle (<https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews>). Among the dataset only the names of food and their corresponding recipes were extracted and used. It contains roughly 520,000 articles of different recipes. Counting total numbers of words of the recipes yielded about 53.5 million (non-unique) words.

**Algorithms**

1. Initial dataset cleaning, feature selection, managing nulls, data cleaning on texts in order to remove unnecessary words and punctuations
2. Cleaned texts were tokenized, and then turned into a term-doc matrix via a select method
3. The term-doc was then used to build a topic model. Hyperparameters were tuned manually
4. The resulting topic model was evaluated by looking at its distributions and clusters.

**Tools**

* **Pandas, regex -** Data cleaning, data manipulation, feature engineering
* **PorterStemmer –** Text preprocessing (only used at a trial run)
* **Seaborn, pyplot, pyLDAvis, wordcloud –** for visualization
* **TF-IDF(sklearn)** – Term-Doc generation
* **NMF, LDA (sklearn)** - Building topic models
* **T-SNES** – Clustering analysis
* Other tools that were only used occasionally are not listed.

**Communication**

The result was summarized via PowerPoint. Rest of the projects (codes) will be organized and then uploaded to GitHub.