The study on distinction for Authentic and Overseas Chinese recipes

Zijie Feng - zijfe244 - 732A92

March 16, 2020

Abstract

Chinese food is an important part of Chinese culture, which influenced many other cuisines during thousands of years. Simultaneously, overseas Chinese restaurants thrive and their dishes differ from traditional Chinese recipes gradually. Based on our studies on 482 Chinese recipes, it is hard to find the difference between authentic and overseas Chinese food via some unsupervised learning skills. However, the hidden specific combination among cooking methods and raw materials can be found by such simply supervised learning models as linear SVM and two-layer neural network, which can be used for the classification for authentic and overseas Chinese recipes.

1 Introduction

Chinese food is very popular among the world, but the authentic Chinese cuisines are not as famous as its name implies. Outsids of China, fake or overseas Chinese food can be created from two backgrounds. The first is natural tramsmission. During thousands of years, many Chinese people (especially the Han nationality) immigrated to the other parts of Aisa and many foreigners came to China as well. With the cultural communication, Chinese cuisines had been spread to Japan, Korean, South-East Asia, India and etc. Such cuisines improve gradually and differ from the authentic Chinese food totally, such as Ramen (Japan), Manchow soup (India) and Pancit (Filipinos).

The second background is Contemporary immigration. From 1800s to 2000s, thousand Chinese labors immigrate to Europe, North & South America and Australia because of the civil war and weakness of ancient China. After several generations, the offsprings of these labors began to run restaurants and new fake Chinese recipes had thereby been invented. Most of such Chinese food are very sweet (ex. General Tso's Chicken) and full of oil and juice (Orange Chicken).

The recipes of dishes would vary depending on the availability of local materials and convenience for transportation, so food menu in restaurant can represent and affect the impression of local guests approximately. This report is aim to apply several machine learning and text mining skills to analyze the differences on recipes between Chinese restaurants both in China and overseas.

2 Theory

With the broadcast of Chinese food [1] [7], the studies about Chinese food are popular. However, most of them focus on such image recognitions as [2], [4], [5] and natural language questions about Chinese cuisines [6]. At present, the accuracy of image recognitions of Chinese cuisines are not very high, with around 80% correction. On contrast, the classification model about questions of Chinese cuisine preforms better, with around 96.22% correction. To be more specific, the Chinese cuisine is an art not only considering the combination of raw materials, but also the approaches for cooking. Such background provides us the possibility for high-accurate classification based on Chinese food instructions. For convenience we will focus on the classification of authentic and overseas Chinese foods based on their text recipes. Besides the knowledge we have learned from lectures in Text Mining, such other algorithms as Author-Topic Model (ATM model), K-Medoids and Support Vector Machine (SVM) would be implemented.

ATM model is an topic model algorithm which considers "author" as supervised attributes. It can provide the author preference and similarity with regard to LDA model. Differ from LDA model, we will use the whole data to imply the ATM model and thereby getting the preferences for both authentic and overseas Chinese food.

K-Medoids is another typical unsupervised learning skill. Compared with K-Means, K-Medoids is much robuster for noises. To check the quality of original data, both K-Means and K-Medoids will be used to the distinguish between recipes of Chinese and non-Chinese food.

SVM is an extended algorithm from maximum margin classifier. It implies kernel functions to classify the observations in hidden high dimension efficiently and robustly.

3 Data

There are 482 recipes totally and 241 recipes for each clusters. All authentic Chinese recipes are gathered from [8], which are written by several Chinese lived outsides China. The majority of recipes are from Sichuan cuisine (or Szechwan cuisine) and Cantonese cuisine (or Hong Kong style). In addition, the non-Chinese recipes are collected from such other websites as [9]-[14]. Their authors are non-Chinese or ethnic Chinese.

We would extract both verbs and nouns from instructions and use word vectorization when necessary. Therefore, we can analyze and create models for food verification. All the instructions will be loaded as lower letter cases, and be separated into nouns and verbs via **en_core_web_lg** from **spaCy**. Figure 1 is an instance of our processed data.

'add small pinch salt sesame oil mince beef mix set aside mix tablespoon cornstarch tablespoon water small bowl water starch cut tofu square cube bring large water boil add pinch salt slide tofu cook minute drain wok heat tablespoon oil fry mince meat crispy transfer beef leave oil fry doubanjiang minute slow fire add garlic scallion white ginger ferment black bean cook second aroma mix pepper flake add water seasoning bring boil high fire gently slide tofu cube add light soy sauce heat boil simmer minute add chop garlic green stir water starch pour half mixture simmer pot wait second pour half slightly taste tofu add pinch salt salty way feel spicy add sugar milder taste carefully broth hot point transfer seasoning stick tofu cube sprinkle szechuan peppercorn powder chop garlic green serve immediately steam rice'

Figure 1: the processed recipes of mapo tofu

Figure 2 is the data frame of original data. To simplify, we consider all the dishes that are possible in Chinese restaurants in China as 'authentic Chinese food', since the cooking methods are different even if the dish is from abroad actually. For example, Kung Pao Chicken is a traditional Chinese dish, but its recipes are different in China and America. Additionally, Chinese people might eat Almond Cookies and Egg Tart, but their recipes are already localized into Chinese flavour.

	name	instruction	label
0	mapo tofu	add a small pinch of salt and sesame oil to mi...	0
1	Chinese Cured Pork Belly	in a small pot, add rice cooking wine, soy sau...	0
2	Mung Bean Cake	pre-soak the yellow mung beans overnight. rins...	0
3	Skinny Chinese Pan-Fried Fish	cut the fish into large chunks around 3-4 cm t...	0
4	Glass Noodles Stir Fry with Shredded Cabbage	in a large bowl, soak the glass noodles with h...	0
...	...	...	...
477	Jungguk-naengmyeon	1.soak the brisket in a bowl of water and set ...	1
478	Kkanpunggi	prep the chicken combine the chicken, ginger,...	1
479	Kkanpung saeu	clean the shrimp, drain, and pat dry with pape...	1
480	Rajogi	chicken flesh eating beomurinda put the cut ch...	1
481	Udong	instructions gather all the ingredients. niku...	1

482 rows × 3 columns

Figure 2: original recipes data

3.1 Verbs

Figures 3 and 4 represent the most 25 common verbs in authentic and overseas Chinese recipes, respectively. There are 7921 verbs in authentic Chinese recipes with total 481 kinds. Compared with that, there are 6490 verbs with 426 different kinds in overseas recipes.

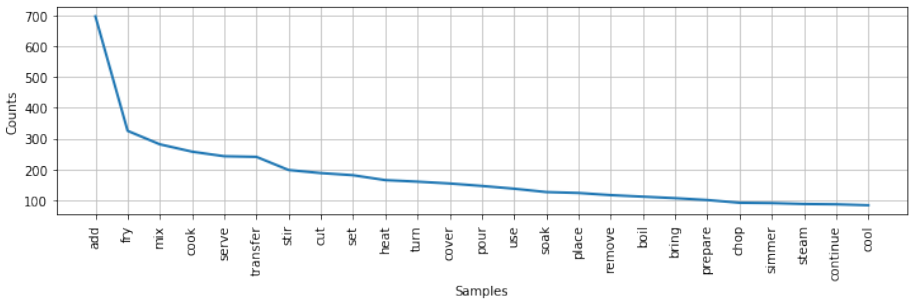

Figure 3: most 25 common verbs in authentic recipes

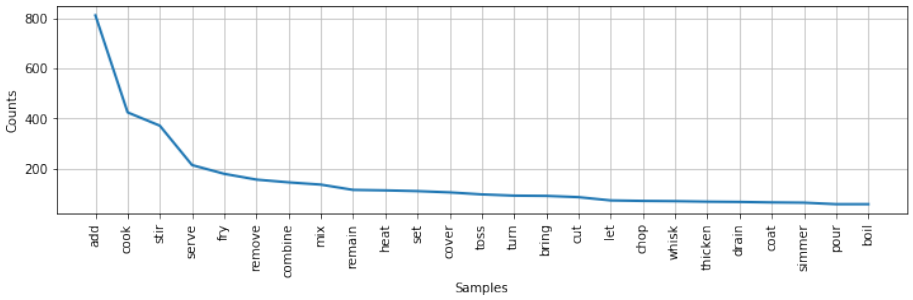

Figure 4: most 25 common verbs in overseas recipes

It is reasonable that such cooking style as *fry*, *mix*, *stir*, *cut*, *boil*, *cover* and simmer are sensible in both figures. But *steam*, *transfer* and *soak* are only in authentic recipes, and *toss*, *whisk*, *thicken* are only in overseas recipes. The verbs *transfer*, *soak* and *thicken* are normally about the situation or color of raw materials. It seems that authentic Chinese food prefers *fry*, but *stir* for the overseas food.

3.2 Nouns

Except of verbs, there are 13653 nouns with 891 kinds for authentic and 12384 nouns with 873 kinds for overseas in all recipes, respectively.

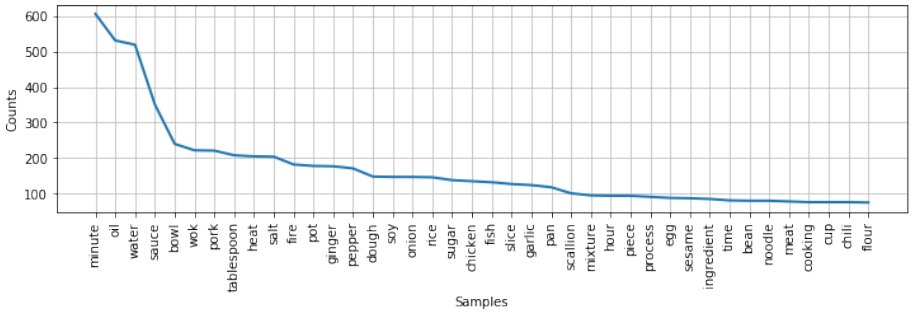

Figure 5: most 40 common nouns in authentic recipes

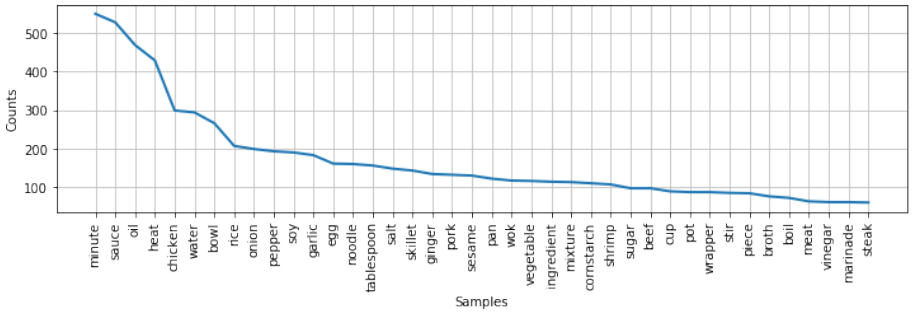

Figure 6: most 40 common nouns in overseas recipes

According to Figures 5 and 6, it is obvious that overseas recipes fancy *chicken*, which overtakes more than 100 counts in authentic recipes. Overseas recipes also like *shrimp*, *beef*, *broth* and *marinade*, but authentic recipes like *fish*, *scallion*, *chili* and *flour*. In addition, the cookers in overseas are more than in authentic ones.

4 Method

According to the detailed verbs and nouns of instructions, we firstly apply topic model (LDA and ATM Models with 100 passes) for analysis of two actual clusters. Then we will apply such unsupervised learning as K-Means with 20 iterations and K-Medoids for the vectorized data (482*3136), and evaluate them by their misclassification rates.

Finally, the data would be separated into train set (361) and validation set (121) randomly. Such two supervised learning as a SVM with linear kernel and a fully-connected Neural Network (NN) with 100 nodes in the first layer and 50 ndoes in the second layer will be applied for classification. Their results would be evaluated by both misclassification rates and ROC curves.

5 Results

5.1 Topic Model

```
[(0,
  '0.032*"add" + 0.025*"oil" + 0.021*"fry" + 0.021*"minute" + 0.018*"sauce" + 0.018*"water" + 0.016*"heat" + 0.014*"wok" + 0.013*"cook"
+ 0.012*"serve" + 0.011*"stir" + 0.011*"mix" + 0.010*"transfer" + 0.010*"tablespoon" + 0.010*"pepper" + 0.010*"salt" + 0.009*"slice" +
0.009*"ginger" + 0.009*"pork" + 0.008*"cut"'),
 (1,
  '0.024*"minute" + 0.020*"water" + 0.015*"dough" + 0.011*"add" + 0.011*"mix" + 0.011*"cover" + 0.010*"place" + 0.009*"bowl" + 0.009*"ri
ce" + 0.008*"flour" + 0.008*"small" + 0.008*"set" + 0.007*"heat" + 0.007*"ball" + 0.007*"oil" + 0.007*"use" + 0.007*"steamer" + 0.007*"m
ixture" + 0.007*"large" + 0.007*"pork"')]
```

Figure 7: LDA for authentic recipes

```
[(0,
  '0.037*"add" + 0.029*"minute" + 0.028*"heat" + 0.025*"sauce" + 0.023*"oil" + 0.023*"cook" + 0.023*"stir" + 0.018*"chicken" + 0.013*"la
rge" + 0.012*"bowl" + 0.012*"medium" + 0.011*"pepper" + 0.010*"water" + 0.010*"rice" + 0.010*"soy" + 0.010*"skillet" + 0.010*"garlic" +
0.010*"serve" + 0.009*"high" + 0.009*"onion"'),
 (1,
  '0.024*"add" + 0.015*"water" + 0.013*"sauce" + 0.010*"wrapper" + 0.009*"oil" + 0.009*"noodle" + 0.009*"shrimp" + 0.009*"fry" + 0.008
*"heat" + 0.008*"soup" + 0.008*"stir" + 0.008*"cook" + 0.008*"minute" + 0.007*"boil" + 0.007*"bowl" + 0.007*"wonton" + 0.007*"serve" +
0.007*"mix" + 0.006*"pork" + 0.006*"step"')]
```

Figure 8: LDA for overseas recipes

```
[(0,
  '0.033*"add" + 0.023*"minute" + 0.022*"heat" + 0.022*"sauce" + 0.019*"oil" + 0.019*"cook" + 0.019*"stir" + 0.013*"chicken" + 0.012*"wa
ter" + 0.011*"bowl" + 0.010*"large" + 0.009*"medium" + 0.009*"fry" + 0.009*"serve" + 0.009*"rice" + 0.009*"pepper" + 0.008*"onion" + 0.0
08*"soy" + 0.008*"pan" + 0.008*"garlic"'),
 (1,
  '0.025*"add" + 0.022*"minute" + 0.019*"oil" + 0.019*"water" + 0.015*"fry" + 0.013*"heat" + 0.013*"sauce" + 0.011*"mix" + 0.010*"stir"
+ 0.010*"cook" + 0.009*"wok" + 0.009*"transfer" + 0.009*"serve" + 0.009*"bowl" + 0.008*"small" + 0.008*"pork" + 0.008*"salt" + 0.008*"ta
blespoon" + 0.007*"place" + 0.007*"cut"')]
```

Figure 9: ATM for all recipes

5.2 K-Means and K-Medoids

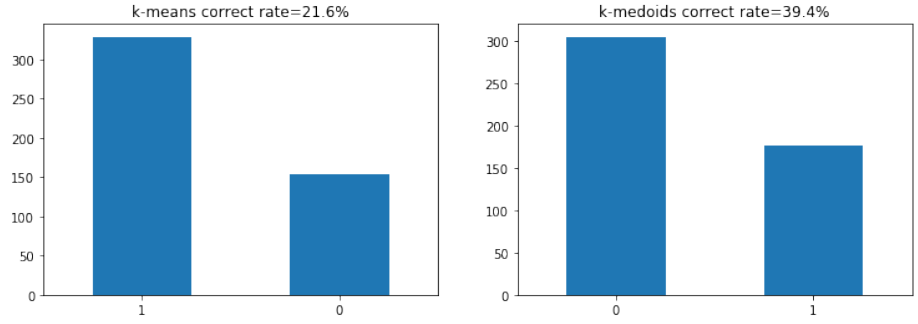

Figure 10: K-Means and K-Medoids

5.3 SVM and NN

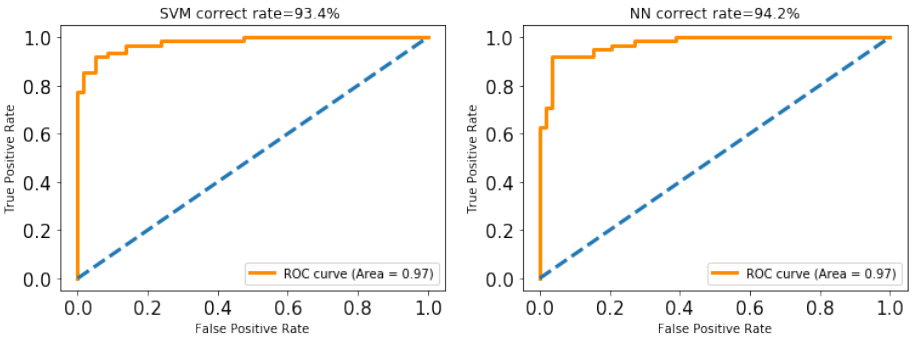

Figure 11: K-Means and K-Medoids

6 Discussion

Figure 7 shows that one typical authentic Chinese recipe might have sliced pork fried with oil, pepper, ginger, salt via wok. Using water to make source and stirring all things with. Then it might be served with rice. Another topic seems about steamed rice with pork or steamed wheaten food (noodle, baozi, etc).

Figure 8 shows an overseas recipe might have rice served with chicken, which cooked with soy sause, garlic, oil, onion and pepper. Another one is a noodle soup or a wonton soup with fried shrimp or pork.

Figure 9 provides us two topic written by two authors, authentic and overseas recipes, respectively. The first is overseas (99.993%) and the second is authentic (99.995%). It is obvious that the material and spices that overseas Chinese food prefers are *chicken* and *pepper*, *onion*, *garlic* and *soy sauce*. On contrast, we can only know that *pork* is the preferable material for authentic Chinese food. But the cooking approahes are abundant, which confirms the verb analysis for original data.

Afterwards, both performances of K-Means and K-Medoids algorithms are very weak, with only 21.6% and 39.4% correct rate respectively. Figure 10 noticed that the recipe distinguishment is not sufficient, built on combination of words only.

Last but not least, the performances of two supervised learning algorithms are quite good. The correct rates are 93.4% for SVM and 94.2% for NN respectively. Described by Figure 11, all the ROC curves cover around 97% area, which confirm the great behaviors of such two models.

7 Conclusion

Frankly speaking, we could distinguish the Chinese food barely when we collect their recipes. It is hard to find the difference among the cooking methods and raw materials when we look at the original data in general, and of course their pictures. So the results differ from our impression entirely.

In accordance with all topic models, it is obvious that authentic Chinese food prefers pork as raw material and served with rice or other flour products (noodle, baozi and wonton). It has more cooking methods than overseas Chinese food. Besides, the main materials are chicken for rice and shrimp for noodle for overseas Chinese recipes.

The reason might be that the specific combination of materials and cooking methods are changed with the improvement and broadcast of Chinese food. To be more precise, traditional Chinese chef likes to use variable spices, such as scallion, onion, pepper, ginger and garlic and their frequencies in authentic recipes are quite similar. In contrary, the Chinese chef overseas favors onion instead of various spices, since onion is more adequate and accessable than other spices. Similarly, chicken is more acceptable than pork in the whole world. For cooking methods, traditional Chinese food prefers fry in wok, which is hard to find outside of China. So pan-cooked and stirred food is more popular overseas.

After that, unsupervised learning algorithms cannot fit the data well because of the complexity of recipes. Whereas, classic supervised learning algorithms can reduce difficulty of classification efficiently, which is beyound our expectation. To some extent, it demonstrates that there are some specific combination of cooking methods and raw materials among authentic Chinese recipes. At the same time, it proves the creativity and adaptability of Chinese food as well.

8 Further Work

To find and test a better model for Chinese food distinguishment, it is reasonable and necessary to train with more recipes. There are 4 official Chinese cuisines in China. They are Chuan (Szechwan), Lu, Yue (Cantonese) and Huaiyang and represent West, North, South and East China cuisines correspondingly. In contrary, the overseas Chinese food also has various sources. Considering more recipes could improve our researches evidently.

Furthermore, recurrent neural network (RNN) would also be preferred. Such other RNNs as GRU and LSTM could consider the weights among words in different indices. It might contribute to the improvement of correction and enlighten us more probabilities in Chinese cuisines.

References

[1] Cheung, S., & Wu, D. Y. (2014). The globalisation of Chinese food. routledge.

[2] Chen, M. Y., Yang, Y. H., Ho, C. J., Wang, S. H., Liu, S. M., Chang, E., ... & Ouhyoung, M. (2012). Automatic chinese food identification and quantity estimation. In SIGGRAPH Asia 2012 Technical Briefs (pp. 1-4).

[3] Chang, R. C., Kivela, J., & Mak, A. H. (2010). Food preferences of Chinese tourists. Annals of tourism research, 37(4), 989-1011.

[4] Chen, X., Zhu, Y., Zhou, H., Diao, L., & Wang, D. (2017). Chinesefoodnet: A large-scale image dataset for chinese food recognition. arXiv preprint arXiv:1705.02743.

[5] Zhang, X. J., Lu, Y. F., & Zhang, S. H. (2016). Multi-task learning for food identification and analysis with deep convolutional neural networks. Journal of Computer Science and Technology, 31(3), 489-500.

[6] Xia, L., Teng, Z., & Ren, F. (2009). Question classification for Chinese cuisine question answering system. IEEJ transactions on electrical and electronic engineering, 4(6), 689-695.

[7] 钟树. (2007). 海外中餐发展很快 [The overseas spread of Chinese food is quite fast]. 决策与信息, (11), 69-70..

[8] China Sichuan Food Chinese Recipes and Eating Culture. Retrieved March 1, 2020. From: https://www.chinasichuanfood.com/recipe-index/

[9] 76 Chinese Food Recipes You'll Want To Make Again And Again.(2019, August 26). Retrieved March 1, 2020. From: https://www.delish.com/cooking/recipe-ideas/g3153/chinese-food-recipes/?slide=6

[10] Easy General Tso's Chicken Recipe - Dinner Then Dessert.(2017, September 30) Retrieved March 1, 2020. From: https://dinnerthendessert.com/general-tsos-chicken/

[11] 32 Chinese Takeout Dishes You Can Master at Home.(2020, February 28) Retrieved March 1, 2020. From: https://www.foodnetwork.ca/comfort-food/photos/chinese-takeout-recipes-to-make-at-home/#!chinese-tofu-sweet-sour-tofu

[12] Asian Dishes Archives - Fifteen Spatulas. Retrieved March 1, 2020. From: https://www.fifteenspatulas.com/recipes/asian-dishes/

[13] 42 Chinese Takeout Fake-Out Recipess. Retrieved March 1, 2020. From: https://www.tasteofhome.com/collection/chinese-food-recipes/

[14] Take Out Fake Out! 20 Asian Recipes You Can Make At Home. (2015, July 23) Retrieved March 1, 2020. From: https://parade.com/848366/shainawizov/take-out-fake-out-20-asian-recipes-you-can-make-at-home/