Dataset Mining to Discover Common Information Using Manual Annotation, TopMine, SegPhrase / AutoPhrase, Word2Vec

Introduction

I chose the Italian cuisine for this assignment. Single words, even the most frequent ones, do not contain useful information for dish mining and may not seem to be related to dishes. Dish names usually consist of two or more words. Therefore, I used various phrase mining tools with the purpose to compare them and, hopefully, select the most efficient one. I tried two versions of TopMine, then AutoPhrase with a refined expanded list of candidate dish names, and completed my analysis with Word2Vec using the same refined expanded list.

Step 1. Refining Manual Annotation File

In order not to lose valuable information, I very carefully revised the Italian cuisine file in manualAnnotationTask.zip by changing 1 to 0 if a phrase was not a dish (e.g. in n out, Italian cuisine, date night, strip mall, low carb, etc.) and 0 to 1 in the opposite case. I also put 0 against any food or drink items that have no association with the Italian cuisine (e.g. diet coke). I also removed common words from some of the phrases (e.g. removed “the” in “the anti pasta”, removed “and” in “and spinach salad” etc.). Then I sorted the list by 1 and 0, and removed all lines with 0 as they were irrelevant to the topic of Italian dishes. In the end I had 129 quality phrases.

Step 2. Expanding Manual Annotation File

I used free online sources of information about the Italian cuisine (including the American Italian cuisine because apparently there is a difference between the two, and some American Italian dishes were never heard of or did not gain popularity in Italy). These sources include:

- Italian-American cuisine: https://en.wikipedia.org/wiki/Italian-American_cuisine
- List of Italian dishes: https://en.wikipedia.org/wiki/List_of_Italian_dishes
- The Ultimate List of Italian Dishes: https://www.nonnabox.com/the-ultimate-list-of-italian-dishes/

I automatically parsed the above three sources with some partial manual assistance, selected multiple-word and some characteristic single-word names of dishes and their components and formed a list of them adding it to the Manual Annotation file for the Italian cuisine from Step 1. Overall, I added around 950 new lines to the list.

Step 3. TopMine 1. Ahmed El-Kishky

TopMine is an implementation of an LDA-based phrase mining algorithm in which you can set the number of topics and tune some other minor parameters, and the model will output the phrases mined for these topics [4]. This particular implementation allows you to vary such parameters as the minimum support (minimum times a phrase appears in the corpus), maximum phrase size, number of topics, Gibbs Sampling iterations (learning parameters for inference), significance threshold affecting the phrase quality which I always set to the maximum quality (5), topic model - two variants of PhraseLDA. Based on a test run, the default model 2 provided better results, so I continued to use it for the rest of the exercise. All of the above parameters are set in a separate run.sh bash script which is convenient.

In terms of the final results which are lists of phrases by topic, it is in a way similar to what we did for Task 1 with the exception that this is done for phrases and not individual words. Apart from the above criteria, you don’t have any control of how the program is run. As we will see in subsequent sections, you can provide quality phrases to some other algorithms, and the new phrases are mined based on these per-determined quality phrases. This functionality is not available in TopMine.

I have run the algorithm with 10 and 15 topics. 4 food-related topics were discovered in the first case, and 6 in the second one. The word distributions in the 4 topics were very similar to the 6 topics, therefore I have included the results of only the 15-topic model into this review. I decided to show 30 top phrases per topic because this number seems to be more representative of a topic than 10 or 20. Naturally, I included here only the topics that are related to food as the other ones are outside of the scope of this work.

Table 1. Food topics from TopMine 1 (15-topic model). Top 30 phrases, their count, total phrases in each topic (bottom row)

	Topic 3	**Topic 5**	**Topic 7**	**Topic 9**	**Topic 11**	**Topic 14**
1	Olive Garden 1802	pizza was good 1464	pasta dishes 1611	glass of wine 951	short rib 430	tomato sauce 1095
2	garlic bread 1182	pizza place 1209	cooked perfectly 712	Caesar salad 935	melt in your mouth 269	goat cheese 670
3	olive oil 1003	good pizza 1085	spaghetti and meatballs 698	wine list 914	taste buds 260	red sauce 670
4	garlic knots 971	ordered a pizza 1003	chicken parm 684	bottle of wine 748	ve eaten 226	marinara sauce 592
5	ice cream 568	thin crust 865	chicken marsala 480	caprese salad 508	wasn t bad 223	cream sauce 448
6	iced tea 536	great pizza 721	chicken parmesan 436	soup or salad 478	Caffe Boa 219	fresh ingredients 435
7	bread basket 413	Pizzeria Bianco 707	ordered the chicken 407	wine selection 465	hard to find 217	mashed potatoes 434
8	chocolate cake 345	pizza joint 621	perfectly cooked 393	house salad 410	rib eye 187	meat and cheese 422
9	bread sticks 331	pizza crust 607	deep dish 353	wines by the glass 369	ve ever tasted 173	fresh mozzarella 411
10	creme brulee 307	pizza was great 561	lobster ravioli 345	side salad 368	ve heard 156	mac and cheese 410
11	pine nuts 234	thin crust pizza 558	alfredo sauce 330	wine bar 331	mouth watering 154	Italian sausage 363
12	bread served 230	cheese pizza 552	portion size 330	red wine 312	wasn t impressed 142	tasted fresh 323
13	truffle oil 225	pizza I ve 487	husband ordered 302	osso bucco 282	didn t taste 135	parmesan cheese 319
14	crab cakes 211	margherita pizza 465	sea bass 300	chopped salad 269	heat lamp 131	meat sauce 319
15	balsamic vinegar 192	pizza we ordered 413	baked ziti 299	white wine 259	ve found 120	blue cheese 311
16	bread and olive oil 183	slice of pizza 405	al dente 278	wine pairing 246	ve experienced 114	tomato and basil 279
17	whipped cream 181	white pizza 404	eggplant parm 272	antipasto salad 245	left overs 111	ricotta cheese 279
18	bread pudding 177	pizza was delicious 391	chicken breast 267	salad dressing 231	braised short ribs 108	dipping sauce 278
19	black olives 171	pizza and wings 358	pasta sauce 267	beet salad 203	good thing 104	sauce and cheese 269
20	room for dessert 161	Humble Pie 335	generous portion 266	house wine 190	hot dog 104	cheese sauce 258
21	cotton candy 160	pepperoni pizza 320	portions are huge 263	salad or soup 187	wasn t expecting 103	mozzarella cheese 248
22	Herbs and Rye 160	pizza by the slice 308	eggplant parmesan 263	minestrone soup 183	dry aged 103	red pepper 242
23	complimentary bread 159	pizza is amazing 274	filet mignon 260	Greek salad 181	tasted good 102	cheese ravioli 238
24	flat bread 158	pizza dough 260	deep dish pizza 247	Cesar salad 168	quick bite 99	ingredients are fresh 234
25	chocolate chip 150	thinly sliced 258	ordered pasta 243	Osso Buco 167	Il Fornaio 94	beef carpaccio 234
26	save room 150	place for pizza 251	fried calamari 242	beer and wine 164	steak house 89	Italian beef 234
27	bread and butter 142	favorite pizza 250	huge portions 242	wine and beer 164	Cafe Boa 88	red onion 233
28	butternut squash 140	style pizza 247	main dish 240	drink wine 161	prime rib 84	fresh basil 229
29	vanilla ice cream 136	delicious pizza 238	side dish 238	beer selection 158	big fan 81	melt in your mouth 225
30	oil and vinegar 131	chicken wings 237	favorite dish 238	great wine	Hard Rock 79	lacked flavor 218
	Total: 882	*Total: 950*	*Total: 1550*	*Total: 1019*	*Total: 620*	*Total: 1800*

Overall, the percentage of dish names among the above topics is pretty high, and you can summarize most of them by one word/phrase. Topic 3 contains a mix of phrases related to deserts, bread, and dressings. Topic 5 is definitely pizza. The largest Topic 7 is about pasta dish entrees, often with meat/fish. Topic 11 is meat dishes / steaks, and Topic 14 can be briefly described as meats / sauces / veggies.

Step 4. TopMine 2. Another Implementation on Github

Through Google search, I found a different version of TopMine on Github [5] and decided to compare it with TopMine 1. This is also an implementation of the same algorithm as in Step 3, but I was wondering if the results will be different. In Tables 1, 2, and 3 you can see a certain similarity between the results of the two algorithms, but the word distributions do not coincide completely. You have to install pypy in order to run TopMine 2 (or change pypy to python in the code) because this fact is not mentioned on the main webpage, and the generic error message does not explicitly say that you need pypy because there is no import statement for it. In addition, TopMone 2 doesn’t have an upfront capability of changing certain algorithm parameters in a bash script as TopMine 1, but you can do it in the code. Also, TopMine 1 seems to be more of a finished product than TopMine 2, and its logging provides more information about the current processes.

The percentage of unrelated phrases for TopMine 2 is somewhat smaller; therefore, only 20 top phrases are shown per topic. Also, the word distributions for the 10-topic and 15-topic models are more different in comparison with TopMine 1, and that is why I have included the results for both models in my review in Tables 2 and 3 below.

Table 2. Food topics from TopMine 2 (10-topic model). Top 20 phrases, their count, total phrases in each topic (bottom row)

	Topic 1	Topic 4	Topic 5	Topic 9
1	olive garden 1724	marinara sauce 808	thin crust 1058	caesar salad 925
2	garlic bread 1347	tomato sauce 786	pizzeria bianco 716	caprese salad 504
3	olive oil 1138	pasta dish 785	gluten free 605	side salad 416
4	garlic knots 973	pasta dishes 729	cheese pizza 578	iced tea 389
5	goat cheese 724	spaghetti meatballs 649	thin crust pizza 505	house salad 387
6	ice cream 641	chicken parm 609	margherita pizza 475	portion size 358
7	mac cheese 436	red sauce 536	white pizza 433	soup salad 357
8	bread basket 399	chicken marsala 488	pizza crust 422	big fan 356
9	chocolate cake 357	cooked perfectly 488	pizza joint 406	huge fan 314
10	fresh mozzarella 356	meat sauce 480	pizza places 380	italian sausage 305
11	bread sticks 339	perfectly cooked 462	love pizza 372	chopped salad 274
12	creme brulee 275	chicken parmesan 441	humble pie 339	portions huge 255
13	tomato sauce 267	cream sauce 389	pepperoni pizza 338	portion sizes 250
14	blue cheese 267	cooked perfection 353	deep dish 300	antipasto salad 243
15	fresh basil 251	alfredo sauce 348	pizza wings 291	home made 234
16	melt mouth 249	mashed potatoes 343	pizza hut 270	italian beef 214
17	ricotta cheese 229	al dente 314	osso bucco 266	beet salad 207
18	truffle oil 227	sea bass 309	slice pizza 258	meatball sandwich 201
19	parmesan cheese 223	lobster ravioli 306	style pizza 240	pasta dishes 197
20	balsamic vinegar 219	ordered chicken 303	fresh ingredients 239	house made 194
	Total: 465	*Total: 981*	*Total: 836*	*Total: 752*

Topic 1 is mainly bread and cheese. Topic 4 is pasta and meats that come with it. Topic 5 is pizza, and topic 9 is mainly salads.

Table 3. Food topics from TopMine 2 (15-topic model). Top 20 phrases, their count, total phrases in each topic (bottom row)

	Topic 1	Topic 3	Topic 5	Topic 9	Topic 10	Topic 11
1	ice cream 641	pasta dishes 889	wine list 1123	olive garden 1724	mashed potatoes 343	thin crust 1058
2	iced tea 389	tomato sauce 875	glass wine 762	garlic bread 1343	cooked perfectly 268	pizzeria bianco 716
3	chocolate cake 357	marinara sauce 834	bottle wine 654	olive oil 1138	filet mignon 259	cheese pizza 578
4	sea bass 309	pasta dish 785	wine selection 497	garlic knots 973	perfectly cooked 249	thin crust pizza 505
5	creme brulee 275	chicken parm 688	wine bar 299	caesar salad 919	melt mouth 249	margherita pizza 475
6	short rib 275	red sauce 676	red wine 291	goat cheese 724	medium rare 238	white pizza 433
7	panna cotta 225	spaghetti meatballs 672	white wine 253	caprese salad 511	beef carpaccio 219	pizza crust 422
8	room dessert 174	chicken marsala 488	house wine 226	mac cheese 436	pork chop 208	pizza joint 401
9	whipped cream 169	meat sauce 484	glasses wine 213	fresh mozzarella 418	pork belly 205	pizza places 378
10	cotton candy 166	cream sauce 472	beer selection 212	bread basket 399	italian beef 203	humble pie 339
11	ice tea 160	chicken parmesan 434	pine nuts 192	side salad 370	cooked perfection 202	pepperoni pizza 338
12	vanilla ice cream 140	alfredo sauce 348	bottles wine 185	house salad 364	italian sausage 196	love pizza 293
13	peanut butter 131	ordered chicken 341	wine glass 184	bread sticks 352	lamb chops 196	pizza wings 291
14	save room 110	al dente 314	red devil 172	blue cheese 297	melted mouth 177	deep dish 282
15	saut ed 104	lobster ravioli 313	beer wine 155	chopped salad 277	mouth watering 166	pizza hut 270
16	italian ice 103	fried calamari 292	great wine 150	soup salad 265	french fries 164	slice pizza 258
17	chocolate sauce 103	husband ordered 292	wine beer 140	fresh basil 260	red pepper 150	style pizza 241
18	chocolate chip 99	eggplant parmesan 280	pinot noir 137	fresh ingredients 250	roasted red peppers 141	pizza dough 238
19	salted caramel 93	eggplant parm 264	wine pairing 119	antipasto salad 244	green beans 141	metro pizza 232
20	save room dessert 92	fettuccine alfredo 250	wines glass 117	parmesan cheese 239	sonny boy 140	deep dish pizza 228
	Total: 482	*Total: 1140*	*Total: 465*	*Total: 981*	*Total: 836*	*Total: 752*

Topic 1 is distinctively deserts. Topic 3 is pasta dishes with meats. Topic 9 is bread and cheese. Topic 10 is meat dishes with vegetables, and Topic 11 is pizza.

In general, you can mine new dish names from the results offered by this algorithm, but it doesn't seem to be as flexible as the next two ones.

Step 5. SegPhrase / AutoPhrase

The SegPhrase Girthub repository [1] currently suggests using a significantly more effective tool called AutoPhrase [2]. This tool allows you to incorporate domain-specific knowledge bases by adding this information to a certain file or by completely replacing this file with your own. Therefore, I was able to use my manual annotation files. For comparison, I made the following five AutoPhrase runs on the Italian cuisine reviews file:

a) straight out of box without my manual annotation;

b) with the refined basic manual annotation file (from Step 1 above) + AutoPhrase's quality phrases;

c) with the expanded manual annotation file (from Step 2 above) + AutoPhrase's quality phrases;

d) with only my expanded manual annotation file (from Step 2) in the AutoPhrase's wiki_quality.txt, but I kept the original AutoPhrase's wiki_all.txt;

e) with only my expanded manual annotation file (from Step 2) in both of AutoPhrase's wiki_quality.txt and wiki_all.txt files.

When using AutoPhrase, I got this error: “POS file doesn't have enough POS tags”. I learned from the error forum in the developer’s Github repository [3] that this happens because the POS tagger complains about the presence of special (non-ASCII) characters in the input file. According to the same forum, it is not an easy task to clean the input file and AutoPhrase needs to be finalized or one should connect a different POS tagger. I didn’t do this as I had limited time for this assignment, but this is something to consider in the future. It can further improve the quality of the final results. Table 4 shows the results of these five runs.

Table 4. AutoPhrase Gradual Improvement Steps. Top 20 phrases for each step

Rank	No Annotation	Step 1 (Revised manualAnnotation Task.zip) + AutoPhrase Default Phrases	Step 2 (All My Annotations) + AutoPhrase Default Phrases	Wiki_quality.txt = Only My Annotations (Step 2) + wiki_all.txt by AutoPhrase	Both wiki_quality.txt and wiki_all.txt = Only My Annotations (Step 2)
	1	**2**	**3**	**4**	**5**
1	sea urchin	treasure island	il posto	caesar salad	lamb chops
2	panda express	arnold palmer	ahi tuna	sea urchin	pork chop
3	san gennaro	blue moon	mamma mia	clam chowder	osso bucco
4	bone marrow	casa di amore	peter piper	lamb chops	au jus
5	zuppa toscana	au jus	caesars palace	pinot grigio	sea bass
6	sauvignon blanc	filet mignon	chili flakes	goat cheese	crab cakes
7	puff pastry	pinot noir	west coast	minestrone soup	chocolate mousse
8	ginger ale	tivoli village	romano's macaroni grill	blood orange	pinot noir
9	san tan	frank lloyd wright	il bosco	ground beef	filet mignon
10	peter piper	wolfgang puck's	fra diavolo	sea salt	panna cotta
11	porta bella	farmers market	egg plant	au jus	pinot grigio
12	planet hollywood	romano's macaroni grill	tivoli village	chicken parmesan	eggplant parm
13	clam chowder	santa monica	treasure island	chocolate mousse	baked ziti
14	striped bass	mama mia's	baked ziti	chocolate cake	ice tea
15	san daniele	maine lobster	tuna tartare	eggplant parm	french fries
16	fountain hills	sea bass	taco bell	bone marrow	pork chops
17	creme brûlée	iced tea	puff pastry	fettuccine alfredo	black pepper
18	le cirque	cotton candy	mama mia	lobster tail	fettuccine alfredo
19	maine lobster	le cirque	corned beef	angel hair	red pepper
20	tivoli village	chilean sea bass	heirloom tomatoes	pork loin	chicken parm

As you can see for yourself, providing my manual annotation file improves the quality of mined dish names significantly. AutoPhrase recommends to keep its own quality phrases in the wiki_quality.txt and wiki_all.txt files, but for the purpose of this exercise it is evident that using only my quality phrases in these two files, without the AutoPhrase’s quality phrases, leads to superior results, and the mined phrases are a lot more related to dish names. So, I am not sure why they have this recommendation.

The total number of mined phrases was approximately around 100,000 for each case, but the their quality (relation to dish names) deteriorated significantly as you go below a certain accuracy. The accuracies in the end of these lists were on the order of less than one percent, and the phrases had to relation to dish names, whatsoever. I assumed that this algorithm assessed every possible phrase. A relatively good dish name mining quality was observed approximately above the accuracy of 80% which corresponded to 5200 phrases in Column 4, and 4800 phrases in Column 5, but by far not all of these phrases were related to dish names.

I wrote a small script that compares these lists with the full annotated file (from Step 2 above), and the above top 20 results in Column 1 contain 15 new mined phrases (not from the annotated file), Column 2 - 15, Column 3 - 17, Column 4 - 3, Column 5 - 7. Of course, you have to remember that Columns 1, 2, and 3 contain a varying number of phrases that are not really dish names. And as for Columns 4 and 5, the percentage of new dish names among the first 20 items may not seem very high, but potentially there are thousands of mined dish names, and further analysis is needed to estimate the overall efficiency of this method for which one would have to remove manually all non-dish names which would take a very large amount of time quite beyond the time that I had for the submission of this task, so naturally I didn't do it.

Step 6. Word2vec

In order to use the word2vec model more efficiently, I removed stopwords, removed non-words using regular expressions, used nltk.word_tokenize() to tokenize every word in each sentence/phrase. Also, I used gensim.Phrases to get phrases from the Italian cuisine reviews, trained a word2vec model, picked phrases from the manual annotation file (Step 2) and tried searching for similar dishes using the `similar_by_word` function, but many words/phrases were not in the trained model's vocabulary which generated an error. Solution - I filtered unseen words out of the manual annotation file according to the method proposed in [7].

I used the Step 2 annotated phrases, and although many of them came from sources different than Yelp, I still managed to have about 400 phrases left after filtering (which is more than 129 phrases from Step 1). Then I used the `similar_by_word` function again for every remaining phrase, knowing that there will be no error message as all of them have been confirmed to be present in the word2vec model. I also had to delete all sauces from the Step 2 file and rerun this process as the sauces were confusing the word2vec model, and it was outputting a lot of them while they are not really dishes.

After some experimenting, I used topn=10 as the top number of similar words for each quality phrase just because it seemed to be reasonable as the greater this number is, the more non-dish names you may get. Once I removed all the duplicates, the total number of new mined phrases became approximately 900. After examining them, I noticed that they there are a lot more of dish names among them compared with when I was reviewing the AutoPhrase results. So, the algorithm seems to provide more quality results. This number possibly could have been greater, had topn been more than 10. Below you can see the top 20 results. When compared to the full annotated list (from Step 2), this word2vec model managed to mine 12 new dish names out of the top 20 items shown in Table 5 which is higher than the percentage of new dish names in the top 20 results from AutoPhrase Columns 4 and 5 (Table 4).

Table 5. Word2Vec similar_by_word Results. Top 20 Phrases

Word2Vec Results			
1	chicken marsala	11	goat cheese
2	chicken parmesan	12	sea bass
3	spaghetti meatballs	13	meat sauce
4	thin crust	14	ice cream
5	chicken parm	15	caramelized onions
6	creme brulee	16	fettuccine alfredo
7	chocolate cake	17	pasta dish
8	tomato sauce	18	red peppers
9	chicken parmigiana	19	short ribs

10	baked ziti	20	garlic knots

Total: 852

Conclusion

Although I gave a good try to TopMine trying to use a different number of topics and even two different implementations, out of the four methods analyzed, the methods that offer the best efficiency and flexibility are AutoPhrase and Word2vec. This has to do with the fact that you can use your current expertise in the field being analyzed in the form of quality phrases in manual annotation files (both algorithms) or similar textual material (Word2Vec). Further analysis is needed to compare the results provided by the two algorithms for which I would have to manually remove all non-dish names and evaluate their overall efficiency, but this requires a significant amount of time which does not fit the timeframe for completing this task.

The general impression is such that AutoPhrase mined thousands of phrases, but they contain a lot of non-related examples, especially when the accuracy goes below 80%. The Word2Vec is characterized by better accuracy, although the number of discovered dishes depends on the size of the sample data set (quality phrases). If it is large and the words of which the quality phrases are comprise are present in the model, you can manage to mine a sufficient number of new dishes with a much better accuracy, but keep in mind that you have to invest some time into collecting and preparing these quality phrases.

Another interesting algorithm to analyze would be the TensorFlow version of Word2Vec [9]

References

1. SegPhrase: https://github.com/shangjingbo1226/SegPhrase
2. AutoPhrase: https://github.com/shangjingbo1226/AutoPhrase
3. AutoPhrase forum: https://github.com/shangjingbo1226/AutoPhrase/issues/15
4. Topmine: http://web.engr.illinois.edu/~elkishk2/
5. Topmine: https://github.com/anirudyd/topmine
6. Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015
7. Gensim Word2vec: https://radimrehurek.com/gensim/models/word2vec.html
8. Similar_by_word: https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.WordEmbeddingsKeyedVectors.similar_by_word
9. TensorFlow Word2vec: https://www.tensorflow.org/tutorials/word2vec
10. Processing unseen words in Word2vec: https://github.com/RaRe-Technologies/gensim/issues/310