Automated Cuisine Classification of Recipes

STEVEN JORGENSEN AND THRILOK NAGAMALLA

Outline

- Objective
- Previous Work
- Data
- Features
- Classifiers
- Results
- Conclusions and Future work
- References

Objective

- Motivation
  - Sorting large sets of untagged recipes is time consuming
  - Given an unknown recipe, determine the most probable cuisine type
  - Help prevent misclassification of recipes

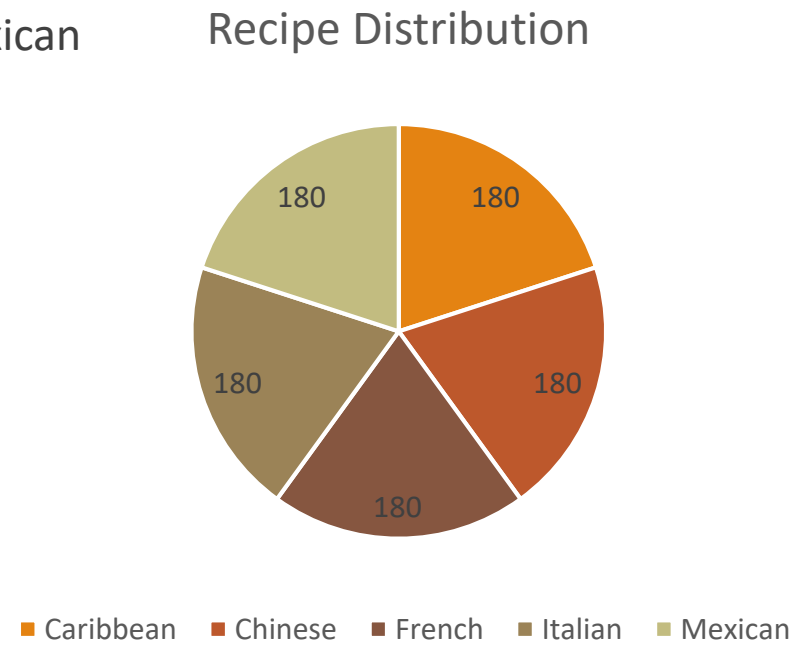
- Goal
  - Automatic cuisine classification of 5 different cuisines
  - Examine performance of different features in the recipes
  - Find which words provide the most information

Previous work

- 2014 – Rishikesh Ghewari, Sunil Raiyani
  - Used ingredients to predict cuisine type
  - Used number of signature ingredients for each cuisine, presence of ingredients as features
- 2015 – Juhi Naik, Vinaya Polamreddi
  - Used set of ingredients to determine what part of the world the dish was from
  - Also used data to automatically generate a recipe

Data

- Generated own dataset using Python module **Beautiful Soup 4**
  - Recipes pulled from www.allrecipes.com
  - Used ingredient and cooking instructions
  - 5 cuisine types – Caribbean, Chinese, French, Italian, Mexican
- 900 recipes total – 180 per cuisine type
- Used 6 fold cross-validation
  - Uniform distribution of recipes per fold
  - Baseline accuracy of 20%
- Each recipe was POS tagged
  - Used NLTK POS tagger
  - Tags used to extract features



Features used

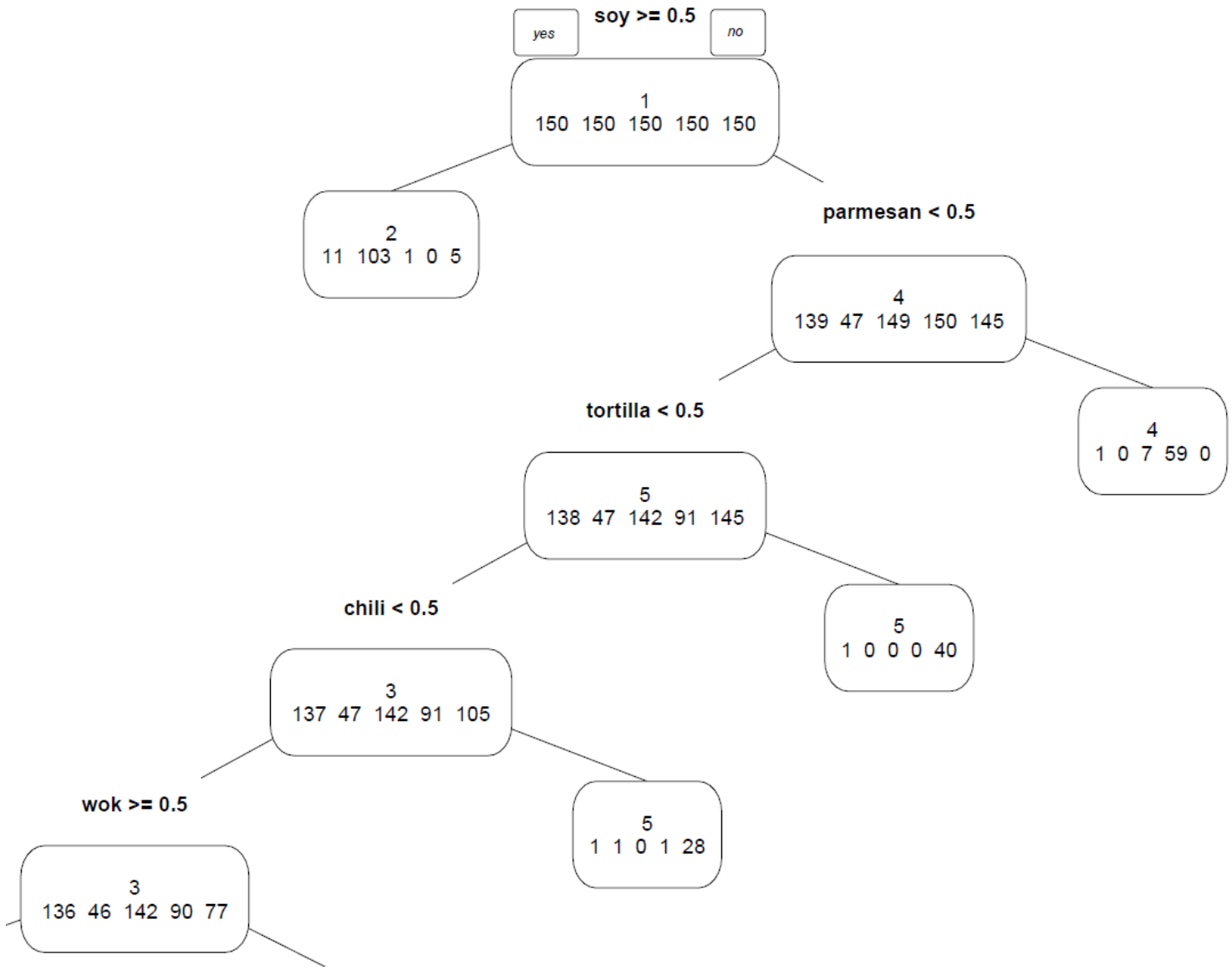
- Unigram
- Bigrams
- Nouns
- Verbs
- Nouns and Verbs
- Ingredients
- Cooking Actions

Classifiers

- Naïve Bayes (NB) – Sklearn implementation
- K Nearest Neighbors (KNN) – Implemented by us
- Support Vector Machine (SVM) – Sklearn implementation
- Recursive Partitioning (Rpart) – Implemented by us

Recursive Partitioning

- Decision tree algorithm
- Finds predictors which would divide the dataset in most informative way forming each node
- Divides until it can not partition further and hence forming leaves



Caribbean, Chinese, French, Italian, Mexican

Results: Unigrams Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	67.22%	58.89%	52.78%	65.56%
Chinese	87.78%	80.55%	66.11%	81.67%
French	63.33%	77.78%	48.33%	68.89%
Mexican	56.67%	76.67%	57.22%	52.22%
Italian	67.22%	56.67%	57.22%	63.33%
Overall	**68.44%**	**70.12**	**56.33%**	**66.33%**

Results: Bigrams Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	75.56%	73.34%	-	72.78%
Chinese	91.11%	75.56%	-	91.67%
French	77.22%	62.23%	-	76.67%
Mexican	67.22%	64.45%	-	66.67%
Italian	70.56%	38.34%	-	70.00%
Overall	**76.33%**	**62.78%**	**-**	**75.56%**

Results: Nouns Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	73.89%	64.44%	56.67%	71.11%
Chinese	87.78%	80.56%	66.67%	86.67%
French	71.67%	67.23%	50.00%	71.67%
Mexican	61.67%	56.12%	56.67%	59.44%
Italian	63.33%	57.23%	49.44%	61.67%
Overall	**71.67%**	**65.12%**	**55.89%**	**70.11%**

Results: Verbs Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	47.22%	50.55%	41.67%	51.67%
Chinese	69.44%	67.23%	46.11%	62.22%
French	65.56%	36.12%	41.11%	66.67%
Mexican	48.89%	42.23%	39.44%	45.00%
Italian	52.22%	41.12%	49.44%	49.44%
Overall	**56.67%**	**47.45%**	**43.56%**	**55.00%**

Results: Nouns and Verbs Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	72.78%	62.78%	50.00%	70.00%
Chinese	87.22%	81.12%	66.11%	86.67%
French	72.22%	62.78%	52.22%	72.78%
Mexican	59.44%	54.45%	53.33%	56.67%
Italian	63.33%	58.89%	57.78%	62.22%
Overall	**71.00%**	**64.00%**	**55.89%**	**69.67%**

Results: Ingredients Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	64.44%	62.78%	52.22%	66.11%
Chinese	81.11%	75.00%	61.67%	80.56%
French	64.44%	68.34%	54.44%	62.78%
Mexican	53.33%	63.89%	58.33%	52.78%
Italian	62.22%	56.67%	52.22%	61.67%
Overall	**65.11%**	**65.34%**	**55.78%**	**64.78%**

Results: Cooking Actions Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Caribbean	42.78%	56.66%	43.89%	40.00%
Chinese	58.33%	23.33%	43.89%	55.00%
French	45.00%	46.67%	41.11%	48.89%
Mexican	35.56%	53.34%	41.67%	37.78%
Italian	32.78%	56.67%	38.89%	36.67%
Overall	**42.89%**	**43.83%**	**41.89%**	**43.67%**

Results: Overall Accuracy

Feature	Naïve Bayes	Recursive Partitioning	KNN	Support Vector Machine
Unigrams	68.44%	70.12%	56.33%	66.33%
Bigrams	76.33%	62.78%	-	75.56%
Nouns	71.67%	65.12%	55.89%	70.11%
Verbs	56.67%	47.45%	43.56%	55.00%
Nouns and Verbs	71.00%	65.12%	55.89%	69.67%
Ingredients	65.11%	65.34%	55.78%	64.78%
Cooking Actions	42.89%	43.89%	41.89%	43.67%

Most Distinguishing Words

Caribbean	Chinese	French	Italian	Mexican
Lime	Soy	Butter	Cheese	Cumin
Juice	Wok	Bake	Parmesan	Tortilla
Blender	Sesame	Oven	Grate	Corn
Allspice	Ginger	Preheat	Olive	Chili
Cinnamon	Cornstarch	Melted	Basil	Cilantro

Most Distinguishing Actions

Caribbean	Chinese	French	Italian	Mexican
Salt	Mince	Bake	Grate	Shred
Refrigerate	Beat	Melt	Salt	Blend
Leave	Fry	Slice	Spread	Chop
Scoop	Boil	Sprinkle	Taste	Roll
Strain	Mix	Form	Arrange	Puree

Conclusions

- Using Bi-grams as features with Naïve Bayes algorithm gives the maximum accuracy
- Support Vector Machine gives the next best results with Bi-grams
- The feature “Cooking actions” gave the minimum accuracy suggesting that all there is a good overlap of these in all the cuisine types
- While Naïve Bayes performed well with Nouns and Verbs, Recursive partitioning did well with ingredients and unigram

Future Work

Extract	Extract important cooking procedures to help readers in understanding the recipe
Expand	Expand the number of cuisines used
Dish	Dish type classification

References and Tools

- Su, Han, et al. ”Automatic recipe cuisine classification by ingredients.” Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM, 2014.
- Naik, Polamreddi “Cuisine Classification and Recipe Generation.” http://cs229.stanford.edu/proj2015/233_report.pdf
- Scikit Learn (Sklearn) http://scikit-learn.org
- Natural Language Toolkit (NLTK) http://www.nltk.org/
- Beautiful Soup 4 https://pypi.python.org/pypi/beautifulsoup4
- Data gathered from www.allrecipes.org
- “R” programming language. Libraries: e1071, rpart, NLP and wordnet.
- WordNet to extract the cooking related words

THANK YOU!!!
QUESTIONS???