Computational Analysis of Food Using Distributional Semantics

Nicolai Ruhnau

February 14 2019

Motivation

- cultural importance of food:
  - identity
  - business
  - politics
- crowd-sourced online recipes as a new data source for distributional semantics:
  - first overview of "Food Computing"[30]
  - new largest 1 million recipe dataset[24]
  - recipe style translation[18, 32]

Problem Statement

- overview and definition attempts[30] could be developed further
- interesting food use cases of distributional semantics evaluated mostly manually [18, 32] - measure predictive power of word representations instead
- food topic model parameters could be discussed in more detail (cf. [17, 21, 31])

Contributions

- comprehensive literature and datasets overview
- qualitative and quantitative evaluation using distributional semantics methods:
  - ingredient type prediction
  - cuisine region prediction
  - food topic model parameters

Ingredient Word Representations

- general domain (cf. about doubts [42, p. 2]):
  - Google News Word2Vec[28]
  - Wikipedia fastText[8]
- 1 Million Recipes dataset[24]
  - Word2Vec[24]
  - modified "im2recipe trijoint embedding"[24]
  - fastText embedding

Prediction of 14 Ingredient Types

	Ingredient Type	Number of occurrences	Rounded relative Size
1	plant_derivative	424	27.7%
2	plant	313	20.4%
3	fruit	186	12.1%
4	vegetable	104	6.7%
5	herb	90	5.8%
6	flower	66	4.3%
7	meat	57	3.7%
8	fish/seafood	56	3.6%
9	spice	55	3.5%
10	alcoholic_beverage	50	3.2%
11	dairy	39	2.5%
12	cereal/crop	39	2.5%
13	nut/seed/pulse	33	2.1%
14	animal	18	1.1%
	Total number of labeled ingredients	1530	100%

- linear classifier
- stratified train and test dataset split
- hyperparameter tuning grid search with 10-fold cross-validation

Prediction of 4 Cuisine Regions (cf. [12, p.18-19])

	Merged Cuisine	Recipes	Rel. Size	Old Cuisine Labels
1	Western	44814	79.3%	NorthAm., WesternEur., NorthernEur., EasternEur.
2	Southern	8094	14.3%	African, LatinAmerican, MiddleEastern, Southern
3	Eastern	2969	5.2%	EastAsian, SoutheastAsian
4	SouthAsian	621	1%	SouthAsian
	Ing. lists	56498	100%	

recipe representations:

- ingredient embedding averages
  - simple ingredient vector sum
  - averages with relative ingredient frequency
- sparse relative ingredient frequency (term frequency inverse document frequency, TF-IDF)

Topic Modeling with Latent Dirichlet Allocation

	Cuisine label	Number of ingredient lists	Relative Size
1	NorthAmerican	41524	73.4%
2	SouthernEuropean	4180	7.3%
3	LatinAmerican	2917	5.1%
4	WesternEuropean	2659	4.7%
5	EastAsian	2512	4.4%
6	MiddleEastern	645	1.1%
7	SouthAsian	621	1.0%
8	SoutheastAsian	457	0.8%
9	EasternEuropean	381	0.6%
10	African	352	0.6%
11	NorthernEuropean	250	0.4%
	Total number of labeled ingredient lists with a unique vocabulary size of	56498 381	100%

- train models with different β parameter
- examine
  - word saliency[11]
  - word clouds
  - document-topic heatmaps

Classifier Results: Predicting 14 Ingredient Types (Test Dataset)

Embedding	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
googlenews	0.720930	0.720930	0.720930	0.736172	0.715873	0.678733
wiki_fasttext	0.746575	0.746575	0.746575	0.718039	0.696419	0.691779
im2rec_joint_null	0.189189	0.189189	0.189189	0.121795	0.230769	0.153846
im2rec_joint_avg	0.216216	0.216216	0.216216	0.133333	0.269231	0.172161
im2rec_base	0.675676	0.675676	0.675676	0.630952	0.711111	0.636735
im2rec_fasttext	0.767123	0.767123	0.767123	0.770681	0.678732	**0.698164**

Classifier Results: Predicting 4 Merged Cuisine Regions (Test Dataset)

Recipe Representation	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
tfidf	0.788598	0.788598	0.788598	0.564138	0.643382	0.591236
googlenews-sum	0.799575	0.799575	0.799575	0.558444	0.662334	0.584284
googlenews-tfidf	0.789837	0.789837	0.789837	0.548688	0.644094	0.570077
googlenews-concatenated	0.789483	0.789483	0.789483	0.544536	0.664274	0.578127
wiki_fasttext-sum	0.790545	0.790545	0.790545	0.546277	0.654734	0.566756
wiki_fasttext-tfidf	0.792139	0.792139	0.792139	0.575980	0.642186	0.601319
wiki_fasttext-concatenated	0.790722	0.790722	0.790722	0.584665	0.639792	**0.606772**
im2rec_joint_null-sum	0.795821	0.795821	0.795821	0.574389	0.611854	0.581965
im2rec_joint_null-tfidf	0.796706	0.796706	0.796706	0.564317	0.630353	0.579569
im2rec_joint_null-concat	0.786081	0.786081	0.786081	0.545738	0.646852	0.576712
im2rec_joint_avg-sum	0.803613	0.803613	0.803613	0.594594	0.594923	0.586057
im2rec_joint_avg-tfidf	0.794404	0.794404	0.794404	0.560378	0.629900	0.575944
im2rec_joint_avg-concat	0.790331	0.790331	0.790331	0.567292	0.655449	0.601776
im2rec_base-sum	0.782894	0.782894	0.782894	0.648952	0.539961	0.566245
im2rec_base-tfidf	0.797769	0.797769	0.797769	0.573206	0.659130	0.604602
im2rec_base-concatenated	0.792810	0.792810	0.792810	0.569377	0.617839	0.582932
im2rec_fasttext-sum	0.788952	0.788952	0.788952	0.568797	0.569571	0.562836
im2rec_fasttext-tfidf	0.797273	0.797273	0.797273	0.549765	0.605417	0.546812
im2rec_fasttext-concat	0.784348	0.784348	0.784348	0.550376	0.666538	0.586922

Word Cloud Topic №28

Conclusions

- practical methodology for ingredient type and cuisine region prediction using distributional semantics:
  - general domain embeddings very competitive (cf. [42, p. 2])
  - n-gram embeddings favored
  - "im2recipe"-embedding modification approach not better than baseline
  - TF-IDF performed better than expected
  - provided research overview facilitates advancement
    - other datasets, p.e. about flavors
- food topic model parameter review approach
  - latent food topics close to reality
- quantitative and qualitative evaluation procedure
  - enables intuitive data exploration and interpretation

Thank you!

Do you have any questions?

References

[1] Yong-Yeol Ahn et al. "Flavor Network and the Principles of Food Pairing". en. In: *Scientific Reports* 1 (Dec. 2011), p. 196. ISSN: 2045-2322. DOI: 10.1038/srep00196. URL: https://www.nature.com/articles/srep00196 (visited on 12/10/2018).

[2] W. E. Arnoldi. "The principle of minimized iterations in the solution of the matrix eigenvalue problem". en. In: *Quarterly of Applied Mathematics* 9.1 (1951), pp. 17–29. ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/42792. URL: https://www.ams.org/home/page/ (visited on 02/08/2019).

References (cont.)

[3] R. Arun et al. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations". en. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Mohammed J. Zaki et al. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 391–402. ISBN: 978-3-642-13657-3.

[4] David Binkley et al. "Understanding LDA in Source Code Analysis". In: *Proceedings of the 22Nd International Conference on Program Comprehension*. ICPC 2014. event-place: Hyderabad, India. New York, NY, USA: ACM, 2014, pp. 26–36. ISBN: 978-1-4503-2879-1. DOI: 10.1145/2597008.2597150. URL: http://doi.acm.org/10.1145/2597008.2597150 (visited on 02/08/2019).

References (cont.)

[5] Bkkbrad. *Latent Dirichlet allocation diagram in plate notation*. Feb. 2008. URL: https://commons.wikimedia.org/wiki/File:Latent_Dirichlet_allocation.svg (visited on 01/22/2019).

[6] David M. Blei and John D. Lafferty. "A correlated topic model of Science". In: *The Annals of Applied Statistics* 1.1 (June 2007). arXiv: 0708.3601, pp. 17–35. ISSN: 1932-6157. DOI: 10.1214/07-AOAS114. URL: http://arxiv.org/abs/0708.3601 (visited on 02/10/2019).

References (cont.)

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937 (visited on 01/13/2019).

[8] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *arXiv:1607.04606 [cs]* (July 2016). arXiv: 1607.04606. URL: http://arxiv.org/abs/1607.04606 (visited on 01/13/2019).

[9] Wray Buntine. "Estimating Likelihoods for Topic Models". en. In: *Advances in Machine Learning*. Ed. by Zhi-Hua Zhou and Takashi Washio. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 51–64. ISBN: 978-3-642-05224-8.

References (cont.)

[10] Juan Cao et al. "A Density-based Method for Adaptive LDA Model Selection". In: *Neurocomput.* 72.7-9 (Mar. 2009), pp. 1775–1781. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2008.06.011. URL: http://dx.doi.org/10.1016/j.neucom.2008.06.011 (visited on 01/13/2019).

[11] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. "Termite: Visualization Techniques for Assessing Textual Topic Models". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. AVI '12. New York, NY, USA: ACM, 2012, pp. 74–77. ISBN: 978-1-4503-1287-5. DOI: 10.1145/2254556.2254572. URL: http://doi.acm.org/10.1145/2254556.2254572 (visited on 01/16/2019).

References (cont.)

[12] Marlies De Clercq. *Prediction of ingredient combinations using machine learning techniques*. 2014. URL: https://lib.ugent.be/fulltxt/RUG01/002/166/653/RUG01-002166653_2014_0001_AC.pdf.

[13] Empetrisor. *English: A panel illustrating probability density functions of a few Dirichlet distributions over a 2-simplex, for the following alpha vectors (clockwise, starting from the upper left corner): (1.3, 1.3, 1.3), (3,3,3), (7,7,7), (2,6,11), (14, 9, 5), (6,2,6)*. July 2016. URL: https://commons.wikimedia.org/wiki/File:Dirichlet-3d-panel.png (visited on 02/08/2019).

References (cont.)

[14] David J. Galas et al. "Expansion of the Kullback-Leibler Divergence, and a new class of information metrics". In: *arXiv:1702.00033 [cs, math, q-bio]* (Jan. 2017). arXiv: 1702.00033. URL: http://arxiv.org/abs/1702.00033 (visited on 02/08/2019).

[15] Yoav Goldberg. "A Primer on Neural Network Models for Natural Language Processing". In: *arXiv:1510.00726 [cs]* (Oct. 2015). arXiv: 1510.00726. URL: http://arxiv.org/abs/1510.00726 (visited on 01/21/2019).

References (cont.)

[16] Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: *arXiv:1402.3722 [cs, stat]* (Feb. 2014). arXiv: 1402.3722. URL: http://arxiv.org/abs/1402.3722 (visited on 01/21/2019).

[17] Thomas L. Griffiths and Mark Steyvers. "Finding scientific topics". en. In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (Apr. 2004), pp. 5228–5235. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0307752101. URL: https://www.pnas.org/content/101/suppl_1/5228 (visited on 01/13/2019).

References (cont.)

[18] Masahiro Kazama et al. "A Neural Network System for Transformation of Regional Cuisine Style". English. In: *Frontiers in ICT* 5 (2018). ISSN: 2297-198X. DOI: 10.3389/fict.2018.00014. URL: https://www.frontiersin.org/articles/10.3389/fict.2018.00014/full (visited on 12/10/2018).

[19] Kmhkmh. *English: euclidean distance illustration*. Mar. 2018. URL: https://commons.wikimedia.org/wiki/File:Euclidean_distance_2d.svg (visited on 02/08/2019).

References (cont.)

[20] Markus Konrad. *Topic Model Evaluation in Python with tmtoolkit*. en-US. Nov. 2017. URL: https://datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/ (visited on 01/16/2019).

[21] Tomasz Kusmierczyk and Kjetil Nørv\a ag. "Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. New York, NY, USA: ACM, 2016, pp. 2013–2016. ISBN: 978-1-4503-4073-1. DOI: 10.1145/2983323.2983897. URL: http://doi.acm.org/10.1145/2983323.2983897 (visited on 01/14/2019).

References (cont.)

[22] Larhmam. *English: Maximum-margin hyperplane and margin for an SVM trained on two classes. Samples on margins are called support vectors.* Oct. 2018. URL: https://commons.wikimedia.org/wiki/File:SVM_margin.png (visited on 02/10/2019).

[23] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605. ISSN: ISSN 1533-7928. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html (visited on 01/14/2019).

References (cont.)

[24] Javier Marin et al. "Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images". In: *arXiv:1810.06553 [cs]* (Oct. 2018). arXiv: 1810.06553. URL: http://arxiv.org/abs/1810.06553 (visited on 01/14/2019).

[25] MathWorks Documentation R2018b. *t-SNE - MATLAB & Simulink*. 2018. URL: https://www.mathworks.com/help/stats/t-sne.html (visited on 02/08/2019).

[26] Matt Burton. *Topic Modeling for JDH*. May 2013. URL: http://mcburton.net/blog/joy-of-tm/ (visited on 02/11/2019).

References (cont.)

[27] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. event-place: Lake Tahoe, Nevada. USA: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959 (visited on 02/10/2019).

[28] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv:1301.3781 [cs]* (Jan. 2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781 (visited on 01/13/2019).

References (cont.)

[29] David Mimno et al. "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 262–272. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462 (visited on 01/13/2019).

[30] Weiqing Min et al. "A Survey on Food Computing". In: *arXiv:1808.07202 [cs]* (Aug. 2018). arXiv: 1808.07202. URL: http://arxiv.org/abs/1808.07202 (visited on 12/10/2018).

References (cont.)

[31] Weiqing Min et al. "You Are What You Eat: Exploring Rich Recipe Information for Cross-Region Food Analysis". In: *IEEE Transactions on Multimedia* 20 (2018), pp. 950–964. DOI: 10.1109/TMM.2017.2759499.

[32] Kensuke Nobumoto et al. "Multilingualization of Restaurant Menu by Analogical Description". In: *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in Conjunction with The 2017 International Joint Conference on Artificial Intelligence*. CEA2017. New York, NY, USA: ACM, 2017, pp. 13–18. ISBN: 978-1-4503-5267-3. DOI: 10.1145/3106668.3106671. URL: http://doi.acm.org/10.1145/3106668.3106671 (visited on 12/14/2018).

References (cont.)

[33] Qef. *The logistic sigmoid function*. July 2008. URL: https://commons.wikimedia.org/wiki/File:Logistic-curve.svg (visited on 01/21/2019).

[34] Alfréd Rényi. "On Measures of Entropy and Information". EN. In: The Regents of the University of California, 1961. URL: https://projecteuclid.org/euclid.bsmsp/1200512181 (visited on 02/08/2019).

[35] Xin Rong. "word2vec Parameter Learning Explained". In: *arXiv:1411.2738 [cs]* (Nov. 2014). arXiv: 1411.2738. URL: http://arxiv.org/abs/1411.2738 (visited on 01/21/2019).

References (cont.)

[36] Saravanan Thirumuruganathan. *Detecting Mixtures of Genres in Movie Dialogues*. en. Jan. 2012. URL: https://saravananthirumuruganathan.wordpress.com/2012/01/10/detecting-mixtures-of-genres-in-movie-dialogues/ (visited on 02/08/2019).

[37] Hinrich Schütze. "Word Space". In: *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, 1993, pp. 895–902.

[38] C. E. Shannon. "A Mathematical Theory of Communication". en. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. ISSN: 1538-7305. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x (visited on 02/08/2019).

References (cont.)

[39] Shuai. *Topic Modeling and t-SNE Visualization*. Dec. 2016. URL: https://shuaiw.github.io/2016/12/22/topic-modeling-and-tsne-visualzation.html (visited on 01/22/2019).

[40] Skbkekas. *English: Plot of the probability mass function for the Poisson distribution*. Feb. 2010. URL: https://commons.wikimedia.org/wiki/File:Poisson_pmf.svg (visited on 02/08/2019).

References (cont.)

[41] Hanna M. Wallach et al. "Evaluation Methods for Topic Models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. event-place: Montreal, Quebec, Canada. New York, NY, USA: ACM, 2009, pp. 1105–1112. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553515. URL: http://doi.acm.org/10.1145/1553374.1553515 (visited on 02/10/2019).

[42] Michael Wiegand, Benjamin Roth, and Dietrich Klakow. "Knowledge Acquisition with Natural Language Processing in the Food Domain : Potential and Challenges". In: 2012.

References (cont.)

[43] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. "Recent Advances of Large-Scale Linear Classification". In: *Proceedings of the IEEE* 100 (2012), pp. 2584–2603. DOI: 10.1109/JPROC.2012.2188013.

Python Libraries used

- scikit-learn (train-test-split, classifiers and results)
- tmtoolkit (topic model statistics & visualizations[20])
- pandas Dataframe (table like format)
- joblib (compression to disk)
- MulticoreTSNE (fast t-SNE[23] computation [39, cf.])

Publicly available Textual Datasets Summary

Dataset	Year	# Recipes	# Images	# Ingredients	Format
Ahn Flavours	2011	56 K, only ingr.	No	1384	5* preprocessed .tsv
Recipe1M	2017	1.029 M	887 K	16 K	2* preprocessed/1* raw .json; .jpg
Yummly66K	2017	66 K	66 K	2416	10* .json files, 9* preprocessed
RecipeQA	2018	1.9 K, 36K Q&A	25 K	Yes	1963* raw .json and .jpg
Meal Master	2013	158 K	-	Yes	.mmf, i.e. structured but raw .txt
RecipeRescuer	2016	93 K	>160 K	Yes	1* somewhat parsed .json
FoodRepo.org	2018	-	37 K	37 K products	37 K .json files via convenient API
BBC.co.uk Food	2016	74 K	74 K	Yes	folders with raw .html and .jpg
FoodOn Ontology	2018	-	-	9 K products	A single .owl file
FooDB	2018	-	Many	722	20* raw/preprocessed .csv or SQL
Wiki Food/Bev.	2018	Many Hundreds	Many	Sometimes	raw .xml/SQL
Wiki Books	2016	2.6 K	Many	Yes	raw .xml/SQL

Publicly available Textual Datasets Metadata

Dataset	Textual Metadata	Data/Metadata Origin
Ahn Flavours	Ingr.-to-type, ingr.-to-chemical-flavor-compounds, ingr.-list-to-cuisine	Fenaroli's Handbook of Flavor Ingredients; 3 popular cooking websites
Recipe1M	Recipe: title and instructions	$<$24 popular cooking websites
Yummly66K	Recipe: title, instructions, cuisine, flavors (some missing), course	yummly66K cooking website; the website computed flavors itself
RecipeQA	Recipe: title, detailed step-by-step instructions with image link	instructables.com website
Meal Master	Recipe: course, occasion, cuisine, diet, techniques and more	Anonymous users or commercial entities since at least 1998
RecipeRescuer	website address each recipe was crawled from	Many different cooking websites; cleaned version of open recipes
FoodRepo.org	Product: origin, energy, fat, carbohydr., sugars, fiber, protein, salt	Coop, Migros, Lidl (.de, .fr, .ita)
BBC.co.uk Food	Recipe: course, occasion, cuisine, diet, dish, season, chef	BBC, also from TV shows
FoodOn Ontology	Product: very detailed taxonomy	International group of researchers funded in part by Canada
FooDB	Ingr.: flavor and other chem. compounds, nutrients, descriptions, types	Extensive and ongoing literature research funded by Canada
Wiki Food/Bev.	Recipe: foods, beverages, techniques, dining, etc.; dishes by ingredient	Wikipedia Authors
Wiki Books	Recipe: cuisine, diet, dish, technique, food group, price	Wikibooks Authors

Ingredient Occurrences

	Ingredient Type	Number of occurrences	Rounded relative Size
1	plant_derivative	424	27.7%
2	plant	313	20.4%
3	fruit	186	12.1%
4	vegetable	104	6.7%
5	herb	90	5.8%
6	flower	66	4.3%
7	meat	57	3.7%
8	fish/seafood	56	3.6%
9	spice	55	3.5%
10	alcoholic_beverage	50	3.2%
11	dairy	39	2.5%
12	cereal/crop	39	2.5%
13	nut/seed/pulse	33	2.1%
14	animal	18	1.1%
	Total number of labeled ingredients	1530	100%

Cuisine Occurrences

	Cuisine label	Number of ingredient lists	Relative Size
1	NorthAmerican	41524	73.4%
2	SouthernEuropean	4180	7.3%
3	LatinAmerican	2917	5.1%
4	WesternEuropean	2659	4.7%
5	EastAsian	2512	4.4%
6	MiddleEastern	645	1.1%
7	SouthAsian	621	1.0%
8	SoutheastAsian	457	0.8%
9	EasternEuropean	381	0.6%
10	African	352	0.6%
11	NorthernEuropean	250	0.4%
	Total number of labeled ingredient lists	56498	100%
	with a unique vocabulary size of	381	

Hierarchical Clustering of Cuisines via Ingredients (cf. [12, p.18-19])

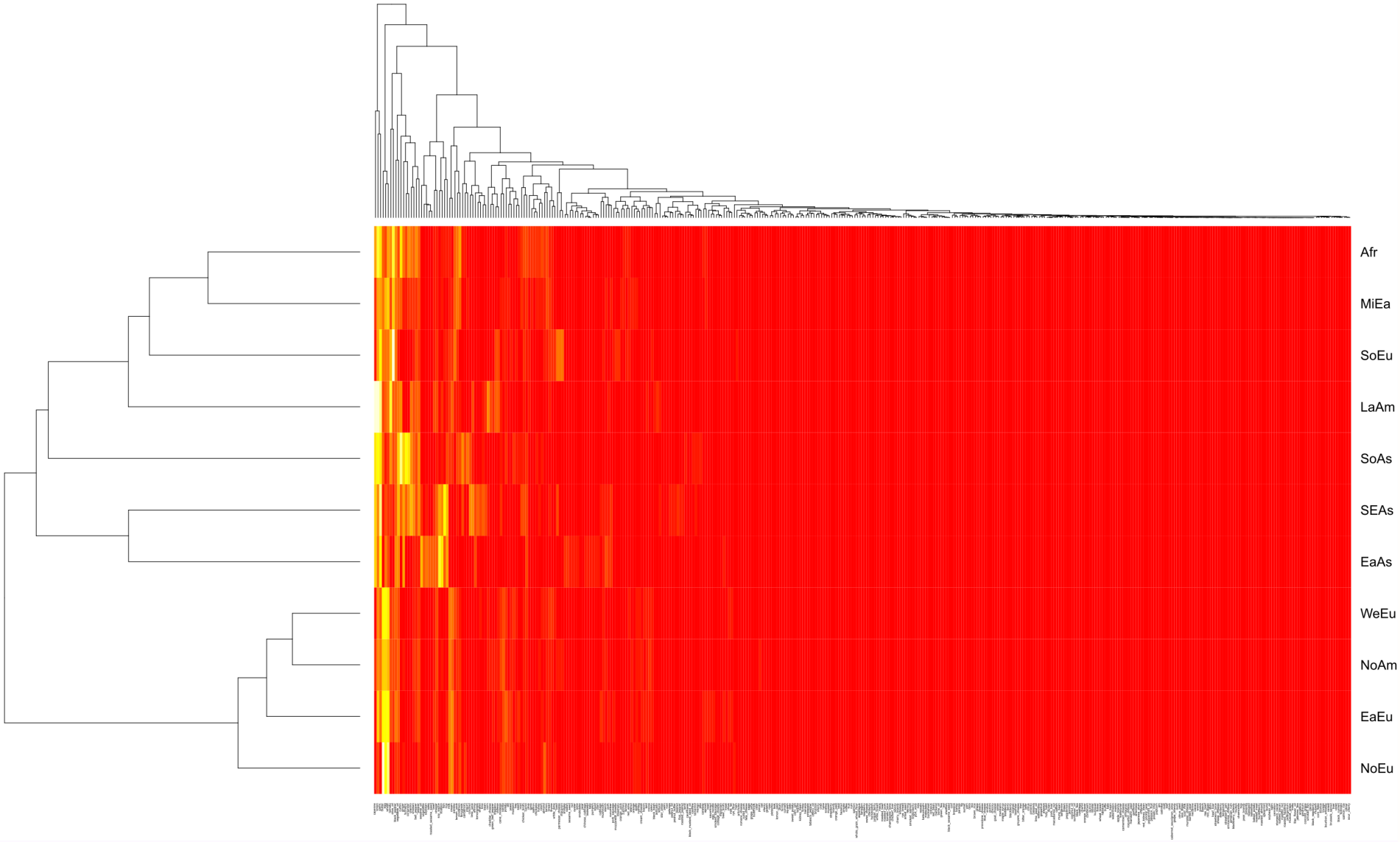


Figure 4.2: Hierarchical clustering of the origin of the recipes based on the ingre-

Merged Cuisine Occurrences

	Merged Cuisine	Recipes	Rel. Size	Old Cuisine Labels
1	Western	44814	79.3%	NorthAm., WesternEur., NorthernEur., EasternEur.
2	Southern	8094	14.3%	African, LatinAmerican, MiddleEastern, Southern
3	Eastern	2969	5.2%	EastAsian, SoutheastAsian
4	SouthAsian	621	1%	SouthAsian
	Ing. lists	56498	100%	

Ingredient Linear SVC Classifier

- 'C': [1000, 500, 200, 100, 50, 20, 10, 5, 2, 1, 0.2, 0.5, 0.01, 0.02, 0.05, 0.001]
- 'penalty': ['l2'] # Default
- 'dual': [False, True] # Default=False
- 'loss': ['squared_hinge'] # Default
- 'max_iter': [3000] # Default=1000
- 'tol': [1e-06, 1e-04] # Default=1e-4
- 'class_weight': ['balanced', None] # Default=balanced
- 'multi_class': ['ovr', 'crammer_singer'] # Default=ovr
- 'random_state': [0]

Merged Cuisine Classifiers

A "LinearSVC" classifier:

- 'C': [500, 200, 100, 50, 20, 10, 5, 2, 1, 0.5, 0.01, 0.05]
- 'dual': [False, True] # Default=False
- 'class_weight': ['balanced'] # Default
- 'random_state': [0]
- 'max_iter': [1000] # Default

The "LogisticRegression" classifier:

- 'C': [500, 200, 100, 50, 20, 10, 5, 2, 1, 0.5, 0.01, 0.05]
- 'dual': [False, True]
- 'multi_class': ['auto'] # Default=ovr
- 'random_state': [0]
- 'max_iter': [500] # Default=100

Ingredient Prediction Tuned Hyperparameters

Embedding	C	class_weight	dual	max_iter	multi_class	tol	minutes	X_cvtrain	X_test Shape
googlenews	0.02	balanced	False	3000	ovr	1e-06	7	(456, 300)	(43, 300)
wiki_fasttext	0.05	balanced	False	3000	ovr	1e-06	41	(1384, 300)	(146, 300)
im2rec_joint_null	1	None	False	3000	crammer_s	1e-06	20	(339, 1024)	(37, 1024)
im2rec_joint_avg	0.01	None	False	3000	crammer_s	0.0001	20	(339, 1024)	(37, 1024)
im2rec_base	0.05	balanced	False	3000	ovr	1e-06	6	(339, 300)	(37, 300)
im2rec_fasttext	0.05	None	False	3000	ovr	1e-06	27	(1384, 300)	(146, 300)

Ingredient Prediction Results

Train Dataset

Embedding	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
googlenews	0.684211	0.684211	0.684211	0.624208	0.662960	0.636028
wiki_fasttext	0.721821	0.721821	0.721821	0.659976	0.694756	0.673757
im2rec_joint_null	0.300885	0.300885	0.300885	0.199645	0.240986	0.214285
im2rec_joint_avg	0.300885	0.300885	0.300885	0.200172	0.235458	0.213121
im2rec_base	0.772861	0.772861	0.772861	0.687432	0.709105	0.695454
im2rec_fasttext	0.768786	0.768786	0.768786	0.747358	0.700814	**0.718177**

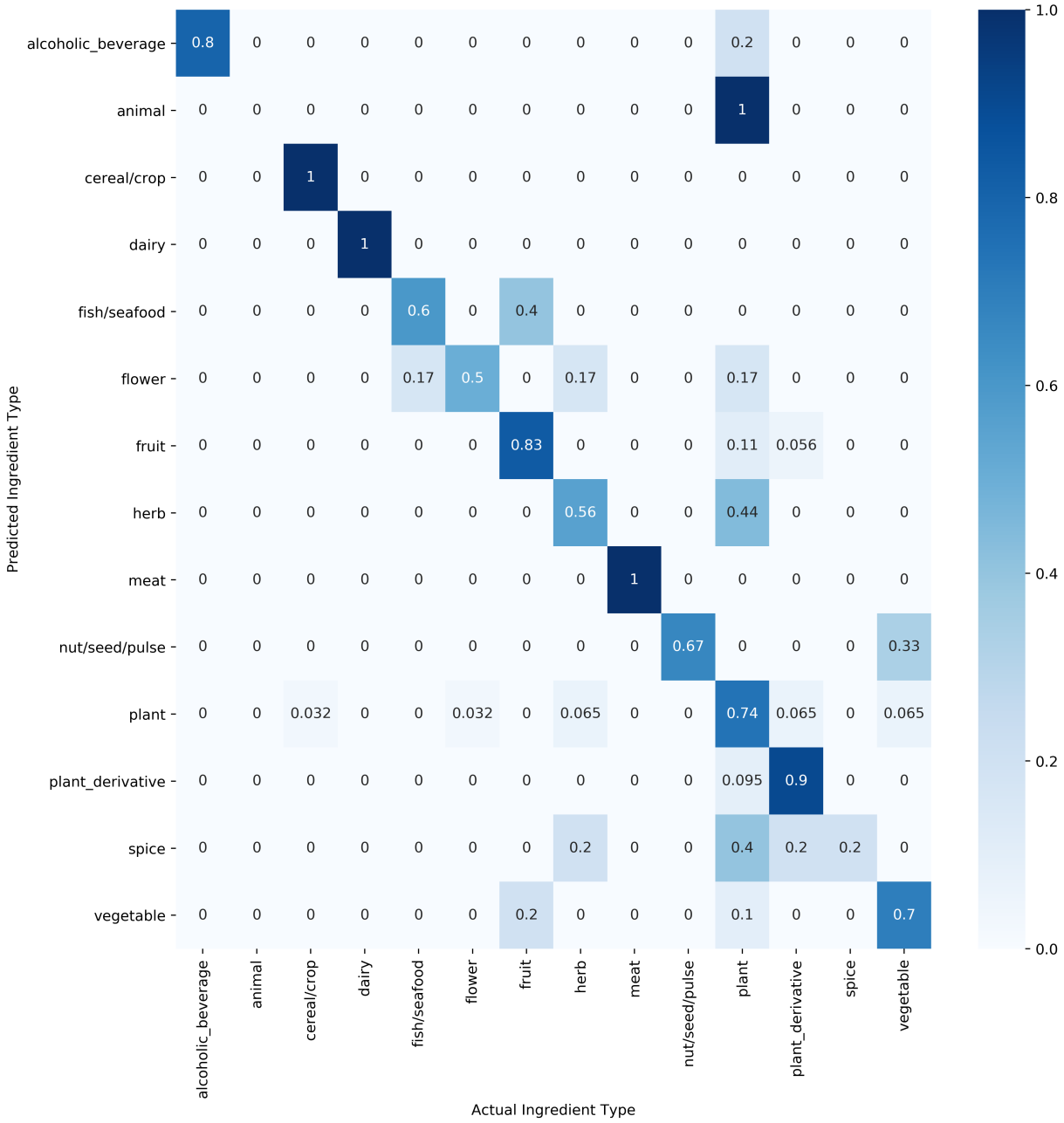
Test Dataset

Embedding	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
googlenews	0.720930	0.720930	0.720930	0.736172	0.715873	0.678733
wiki_fasttext	0.746575	0.746575	0.746575	0.718039	0.696419	0.691779
im2rec_joint_null	0.189189	0.189189	0.189189	0.121795	0.230769	0.153846
im2rec_joint_avg	0.216216	0.216216	0.216216	0.133333	0.269231	0.172161
im2rec_base	0.675676	0.675676	0.675676	0.630952	0.711111	0.636735
im2rec_fasttext	0.767123	0.767123	0.767123	0.770681	0.678732	**0.698164**

Ingredient Prediction Detailed Results Test Dataset im2rec_fasttext

im2rec_fasttext	precision	recall	macro-f1-score	support
alcoholic_beverage	1.000000	0.800000	0.888889	5.0
animal	0.000000	0.000000	0.000000	1.0
cereal/crop	0.750000	1.000000	0.857143	3.0
dairy	1.000000	1.000000	1.000000	3.0
fish/seafood	0.750000	0.600000	0.666667	5.0
flower	0.750000	0.500000	0.600000	6.0
fruit	0.789474	0.833333	0.810811	18.0
herb	0.555556	0.555556	0.555556	9.0
meat	1.000000	1.000000	1.000000	5.0
nut/seed/pulse	1.000000	0.666667	0.800000	3.0
plant	0.589744	0.741935	0.657143	31.0
plant_derivative	0.904762	0.904762	0.904762	42.0
spice	1.000000	0.200000	0.333333	5.0
vegetable	0.700000	0.700000	0.700000	10.0
micro avg	0.767123	0.767123	0.767123	146.0
macro avg	0.770681	0.678732	0.698164	146.0

Ing. Pr. Conf. Mat. Test im2rec_fasttext



Merged Cuisine Prediction Tuned Hyperparameters

Recipe Embedding	C	class_weight	dual	max_iter	minutes	X_cvtrain	X_test Shape
tfidf	100	balanced	True	1000	6	(50850, 381)	(5648, 381)
googlenews-sum	0.1	balanced	True	1000	188	(50828, 300)	(5648, 300)
googlenews-tfidf	10	balanced	True	1000	84	(50828, 300)	(5648, 300)
googlenews-concat	5	balanced	True	1000	72	(50828, 614)	(5648, 614)
wiki_fasttext-sum	0.1	balanced	True	1000	137	(50850, 300)	(5648, 300)
wiki_fasttext-tfidf	10	balanced	True	1000	78	(50850, 300)	(5648, 300)
wiki_fasttext-concat	10	balanced	True	1000	78	(50850, 681)	(5648, 681)
im2rec_joint_null-sum	5	balanced	True	1000	388	(50836, 1024)	(5647, 1024)
im2rec_joint_null-tfidf	100	balanced	True	1000	184	(50836, 1024)	(5647, 1024)
im2rec_joint_null-concat	50	balanced	True	1000	148	(50836, 1326)	(5647, 1326)
im2rec_joint_avg-sum	10	balanced	True	1000	394	(50836, 1024)	(5647, 1024)
im2rec_joint_avg-tfidf	100	balanced	True	1000	184	(50836, 1024)	(5647, 1024)
im2rec_joint_avg-concat	100	balanced	True	1000	150	(50836, 1326)	(5647, 1326)
im2rec_base-sum	2	balanced	True	1000	149	(50836, 300)	(5647, 300)
im2rec_base-tfidf	50	balanced	True	1000	80	(50836, 300)	(5647, 300)
im2rec_base-concat	50	balanced	True	1000	69	(50836, 602)	(5647, 602)
im2rec_fasttext-sum	1	balanced	True	1000	138	(50850, 300)	(5648, 300)
im2rec_fasttext-tfidf	50	balanced	True	1000	76	(50850, 300)	(5648, 300)
im2rec_fasttext-concat	10	balanced	True	1000	73	(50850, 681)	(5648, 681)

Merged Cuisine Prediction Results Train Dataset

Recipe Embedding	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
tfidf	0.815477	0.815477	0.815477	0.618125	0.737758	0.663677
googlenews-sum	0.815417	0.815417	0.815417	0.612687	0.738439	0.657196
googlenews-tfidf	0.810675	0.810675	0.810675	0.607128	0.733855	0.655117
googlenews-concat	0.807822	0.807822	0.807822	0.593693	0.740496	0.644301
wiki_fasttext-sum	0.818230	0.818230	0.818230	0.615099	0.730458	0.655571
wiki_fasttext-tfidf	0.816559	0.816559	0.816559	0.626505	0.724543	0.666916
wiki_fasttext-concat	0.814612	0.814612	0.814612	0.629419	0.726221	**0.669161**
im2rec_joint_null-sum	0.818967	0.818967	0.818967	0.627987	0.727637	0.667610
im2rec_joint_null-tfidf	0.813282	0.813282	0.813282	0.615648	0.727338	0.659470
im2rec_joint_null-concat	0.812495	0.812495	0.812495	0.608471	0.737542	0.656126
im2rec_joint_avg-sum	0.823098	0.823098	0.823098	0.654947	0.705273	0.677936
im2rec_joint_avg-tfidf	0.815485	0.815485	0.815485	0.623481	0.715941	0.662140
im2rec_joint_avg-concat	0.811610	0.811610	0.811610	0.618483	0.723870	0.661543
im2rec_base-sum	0.819419	0.819419	0.819419	0.660020	0.686968	0.672851
im2rec_base-tfidf	0.815092	0.815092	0.815092	0.627617	0.704635	0.660922
im2rec_base-concat	0.812574	0.812574	0.812574	0.622402	0.699331	0.656055
im2rec_fasttext-sum	0.828594	0.828594	0.828594	0.663399	0.693225	0.676819
im2rec_fasttext-tfidf	0.822321	0.822321	0.822321	0.641468	0.699732	0.666786
im2rec_fasttext-concat	0.813117	0.813117	0.813117	0.611035	0.748647	0.660656

Merged Cuisine Prediction Results Test Dataset

Recipe Representation	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
tfidf	0.788598	0.788598	0.788598	0.564138	0.643382	0.591236
googlenews-sum	0.799575	0.799575	0.799575	0.558444	0.662334	0.584284
googlenews-tfidf	0.789837	0.789837	0.789837	0.548688	0.644094	0.570077
googlenews-concatenated	0.789483	0.789483	0.789483	0.544536	0.664274	0.578127
wiki_fasttext-sum	0.790545	0.790545	0.790545	0.546277	0.654734	0.566756
wiki_fasttext-tfidf	0.792139	0.792139	0.792139	0.575980	0.642186	0.601319
wiki_fasttext-concatenated	0.790722	0.790722	0.790722	0.584665	0.639792	**0.606772**
im2rec_joint_null-sum	0.795821	0.795821	0.795821	0.574389	0.611854	0.581965
im2rec_joint_null-tfidf	0.796706	0.796706	0.796706	0.564317	0.630353	0.579569
im2rec_joint_null-concat	0.786081	0.786081	0.786081	0.545738	0.646852	0.576712
im2rec_joint_avg-sum	0.803613	0.803613	0.803613	0.594594	0.594923	0.586057
im2rec_joint_avg-tfidf	0.794404	0.794404	0.794404	0.560378	0.629900	0.575944
im2rec_joint_avg-concat	0.790331	0.790331	0.790331	0.567292	0.655449	0.601776
im2rec_base-sum	0.782894	0.782894	0.782894	0.648952	0.539961	0.566245
im2rec_base-tfidf	0.797769	0.797769	0.797769	0.573206	0.659130	0.604602
im2rec_base-concatenated	0.792810	0.792810	0.792810	0.569377	0.617839	0.582932
im2rec_fasttext-sum	0.788952	0.788952	0.788952	0.568797	0.569571	0.562836
im2rec_fasttext-tfidf	0.797273	0.797273	0.797273	0.549765	0.605417	0.546812
im2rec_fasttext-concat	0.784348	0.784348	0.784348	0.550376	0.666538	0.586922

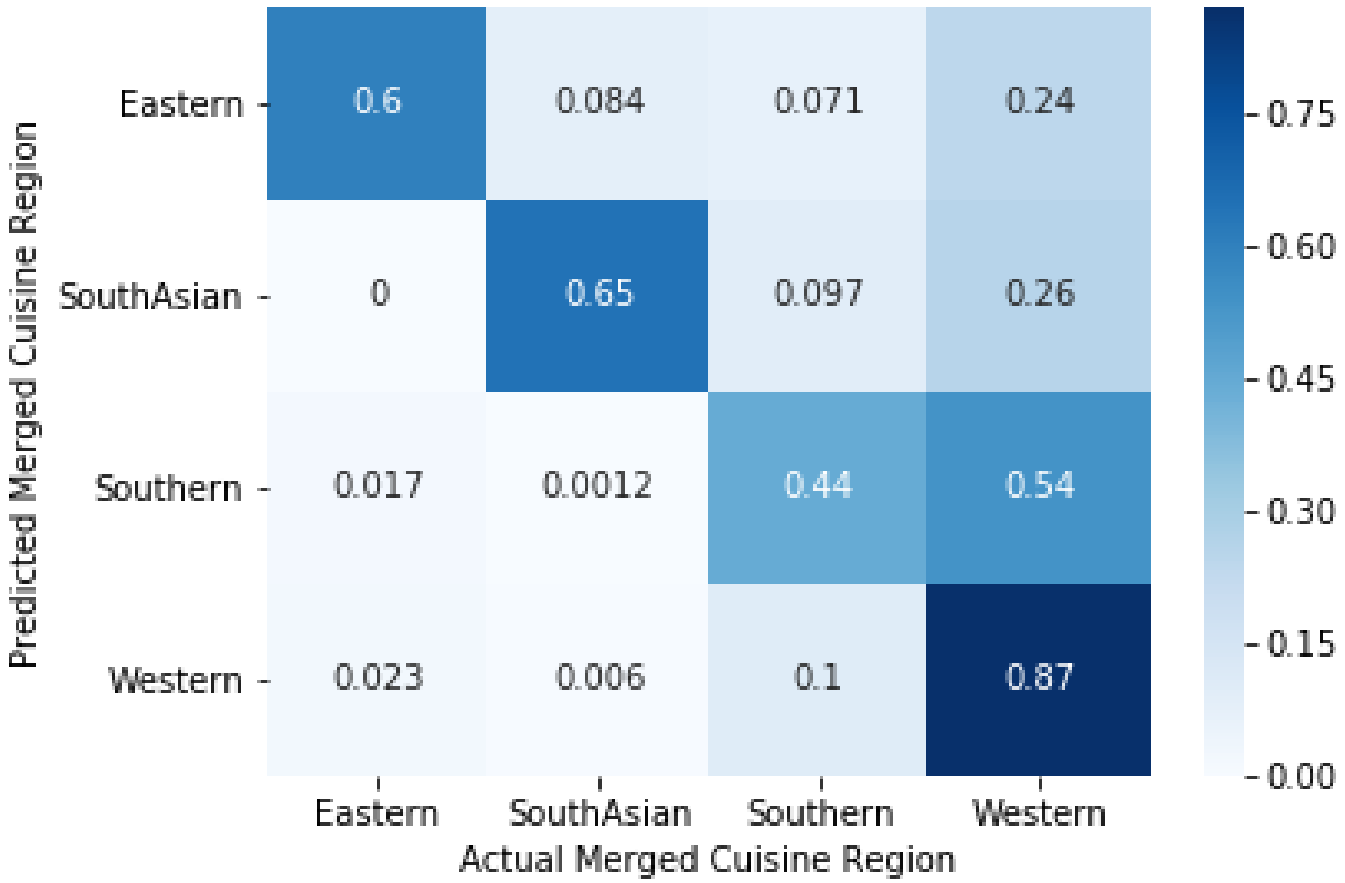
Merged Cuisine Prediction Logistic Regression Results Test Dataset

Embedding	micro-p	micro-r	micro-f1	macro-p	macro-r	macro-f1
tfidf	0.785373	0.785373	0.785373	0.549910	0.637981	0.576688
googlenews-sum	0.802231	0.802231	0.802231	0.573912	0.636614	0.591161
googlenews-tfidf	0.793201	0.793201	0.793201	0.570464	0.618766	0.585025
googlenews-concat	0.793732	0.793732	0.793732	0.571193	0.623605	0.588029
wiki_fasttext-sum	0.799929	0.799929	0.799929	0.596474	0.594779	0.589547
wiki_fasttext-tfidf	0.788421	0.788421	0.788421	0.563295	0.644124	0.594336
wiki_fasttext-concat	0.785234	0.785234	0.785234	0.558820	0.611980	0.576083
im2rec_joint_null-sum	0.798123	0.798123	0.798123	0.593311	0.614017	**0.598902**
im2rec_joint_null-tfidf	0.787498	0.787498	0.787498	0.552860	0.626504	0.570799
im2rec_joint_null-concat	0.785727	0.785727	0.785727	0.551843	0.634149	0.576974
im2rec_joint_avg-sum	0.797415	0.797415	0.797415	0.577387	0.614329	0.587515
im2rec_joint_avg-tfidf	0.786967	0.786967	0.786967	0.552370	0.626083	0.570618
im2rec_joint_avg-concat	0.786081	0.786081	0.786081	0.552880	0.634260	0.577528
im2rec_base-sum	0.797769	0.797769	0.797769	0.583129	0.627068	0.596599
im2rec_base-tfidf	0.785904	0.785904	0.785904	0.554424	0.636600	0.580314
im2rec_base-concat	0.786435	0.786435	0.786435	0.553228	0.635219	0.578506
im2rec_fasttext-sum	0.786296	0.786296	0.786296	0.577318	0.609649	0.589316
im2rec_fasttext-tfidf	0.780984	0.780984	0.780984	0.561476	0.629079	0.583537
im2rec_fasttext-concat	0.780453	0.780453	0.780453	0.561889	0.627584	0.583192

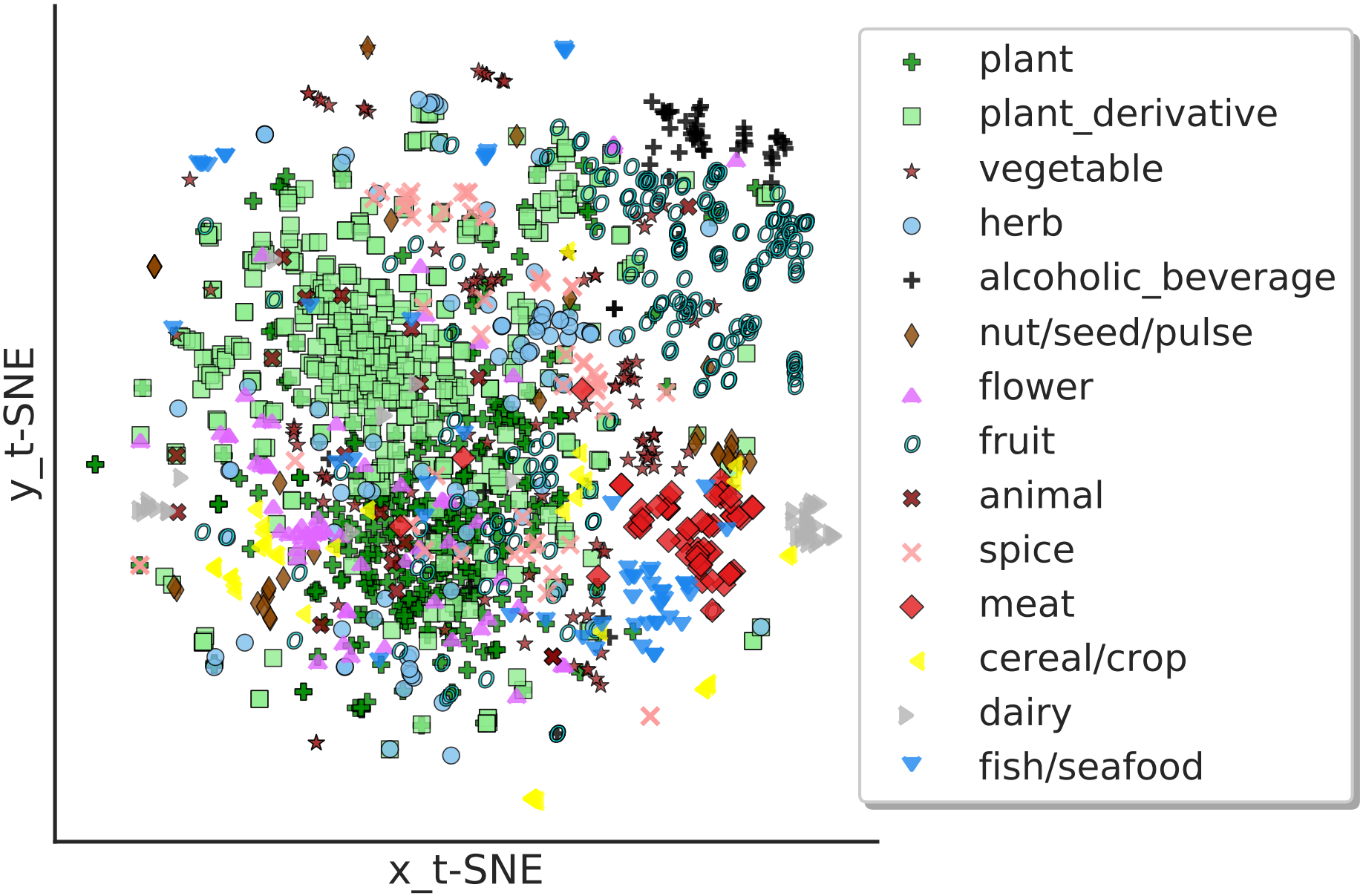
Merged Cuisine Prediction Detailed Results Test Dataset wiki_fasttext_concat

wiki_fasttext-concat	precision	recall	macro-f1-score	support
Eastern	0.603390	0.601351	0.602369	296.0
SouthAsia	0.430108	0.645161	0.516129	62.0
Southern	0.423529	0.444994	0.433996	809.0
Western	0.881633	0.867663	0.874592	4481.0
micro avg	0.790722	0.790722	0.790722	5648.0
macro avg	0.584665	0.639792	0.606772	5648.0

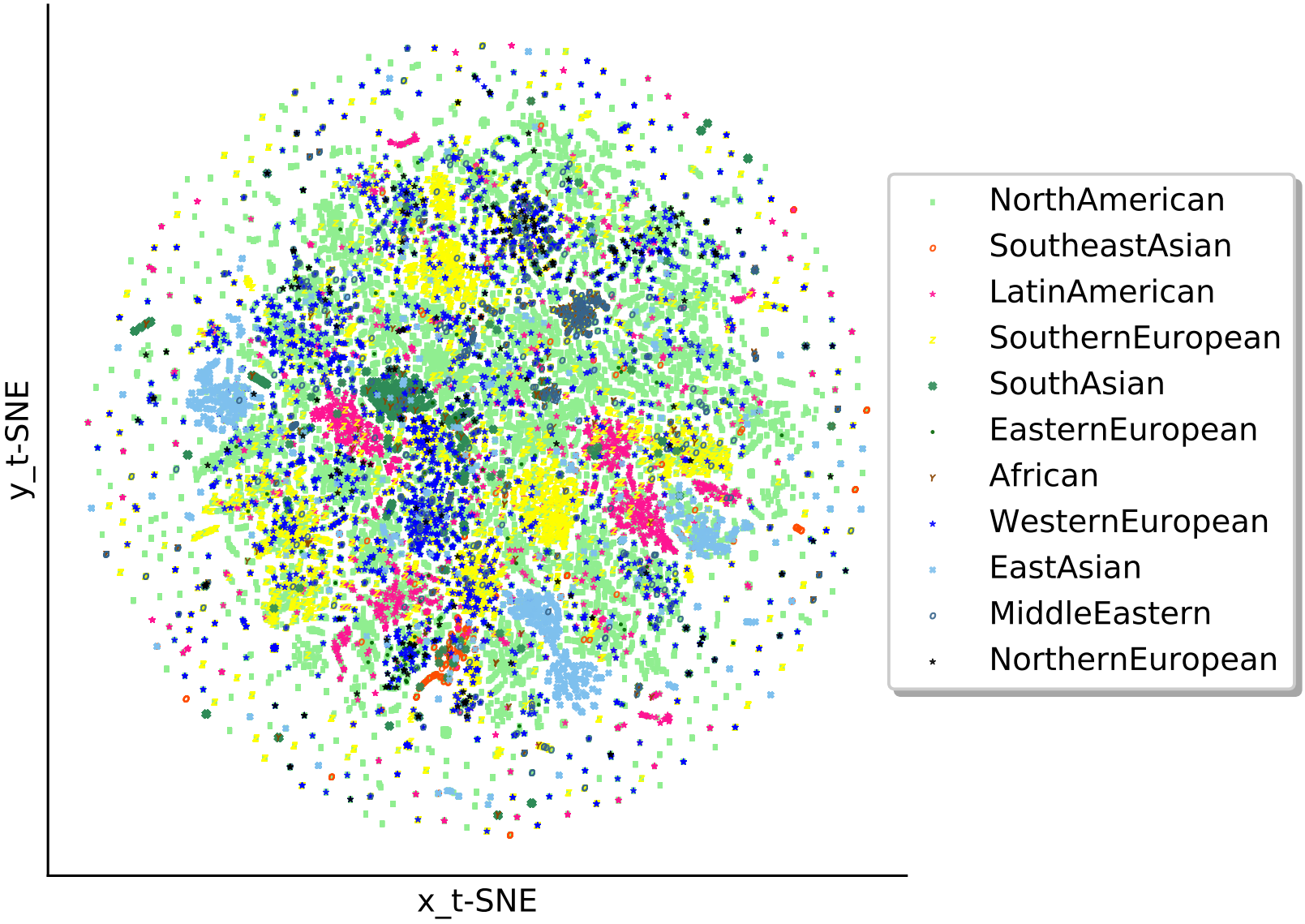
Merged Cuisine Prediction Confusion Matrix Test Dataset wiki_fasttext_concat



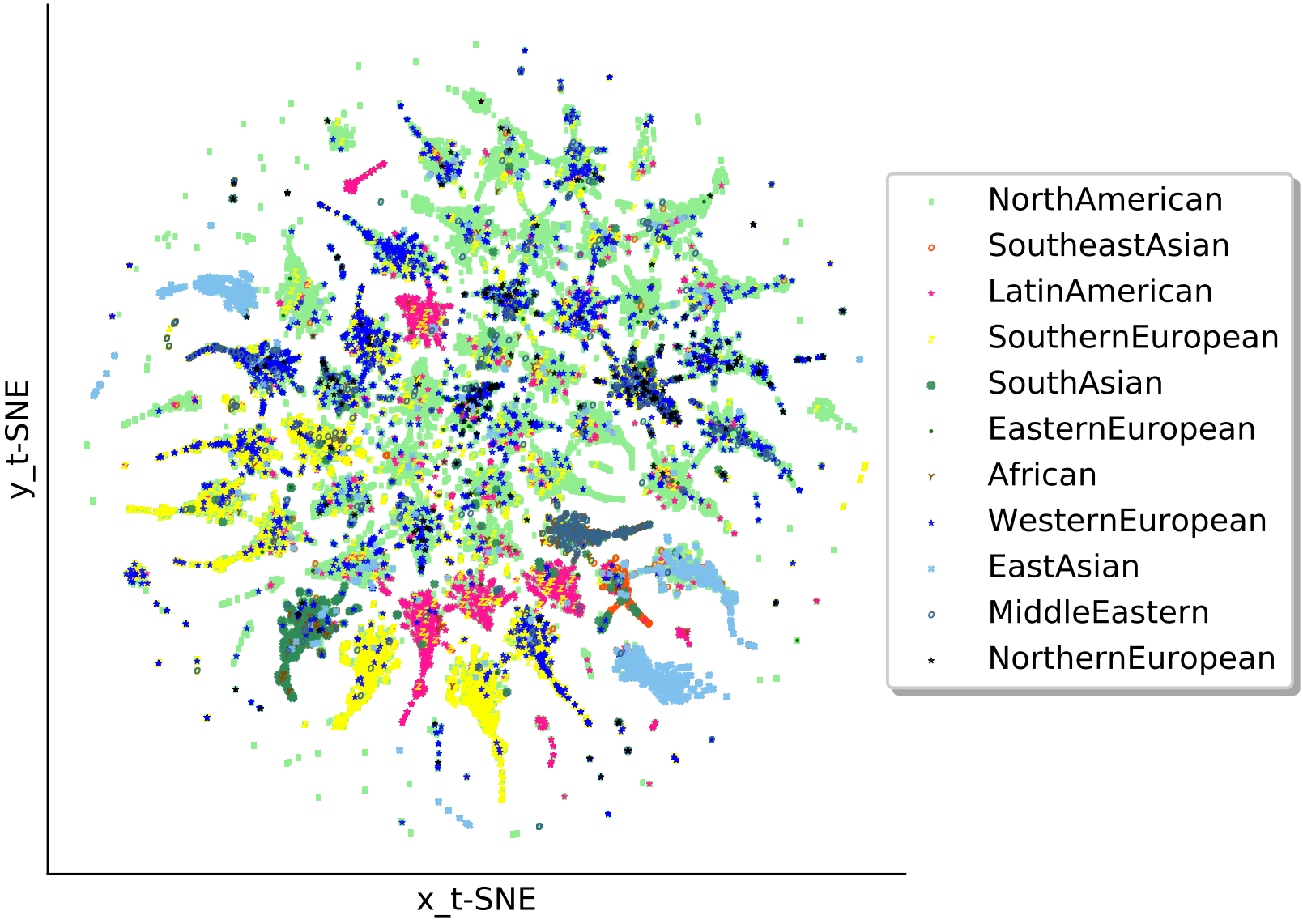
14 Ingredient Types (im2rec_fasttext, p.=40)



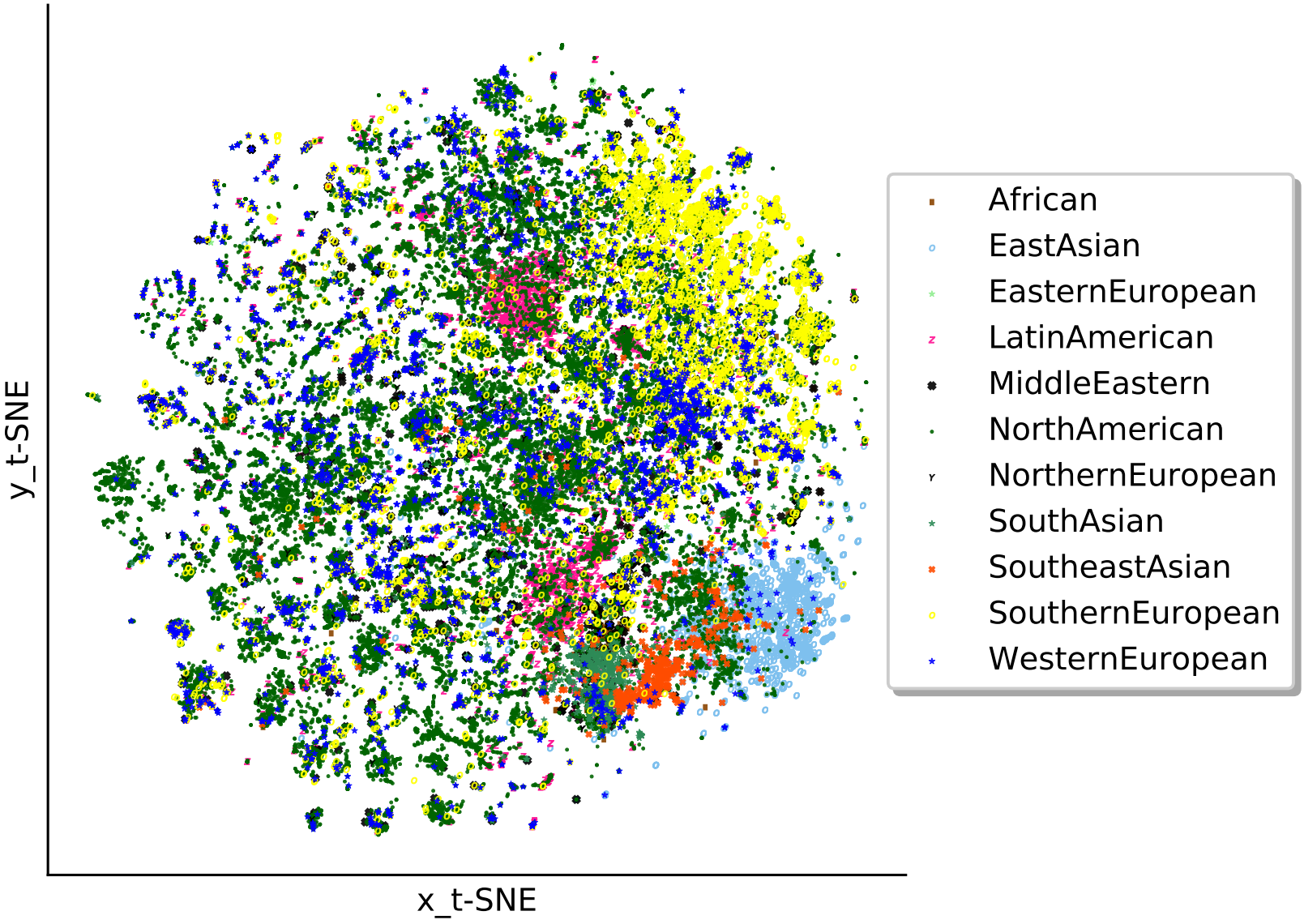
11 Cuisines Topic Model №1 (perplexity 5)



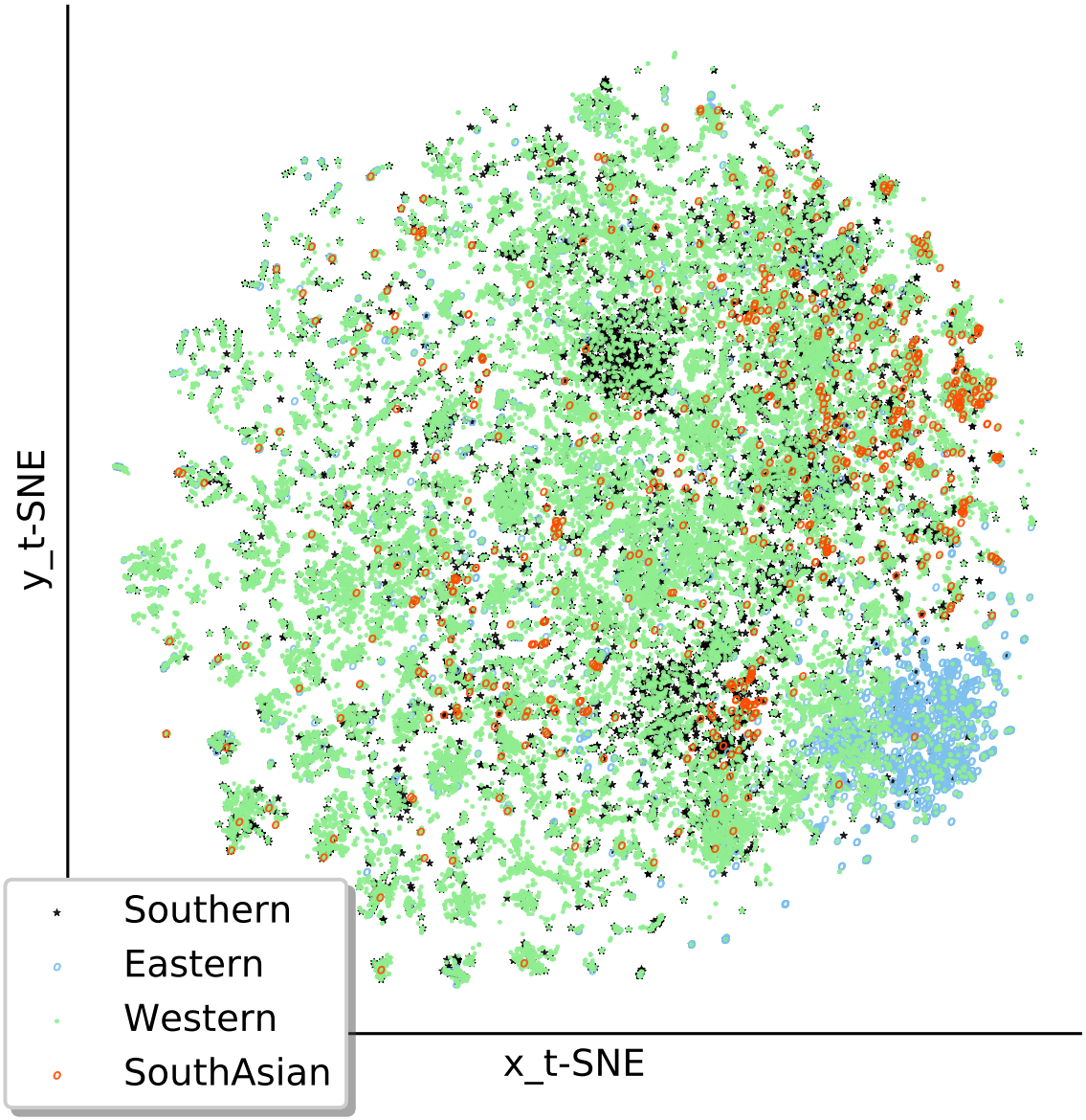
11 Cuisines Topic Model №1 (perplexity 50)



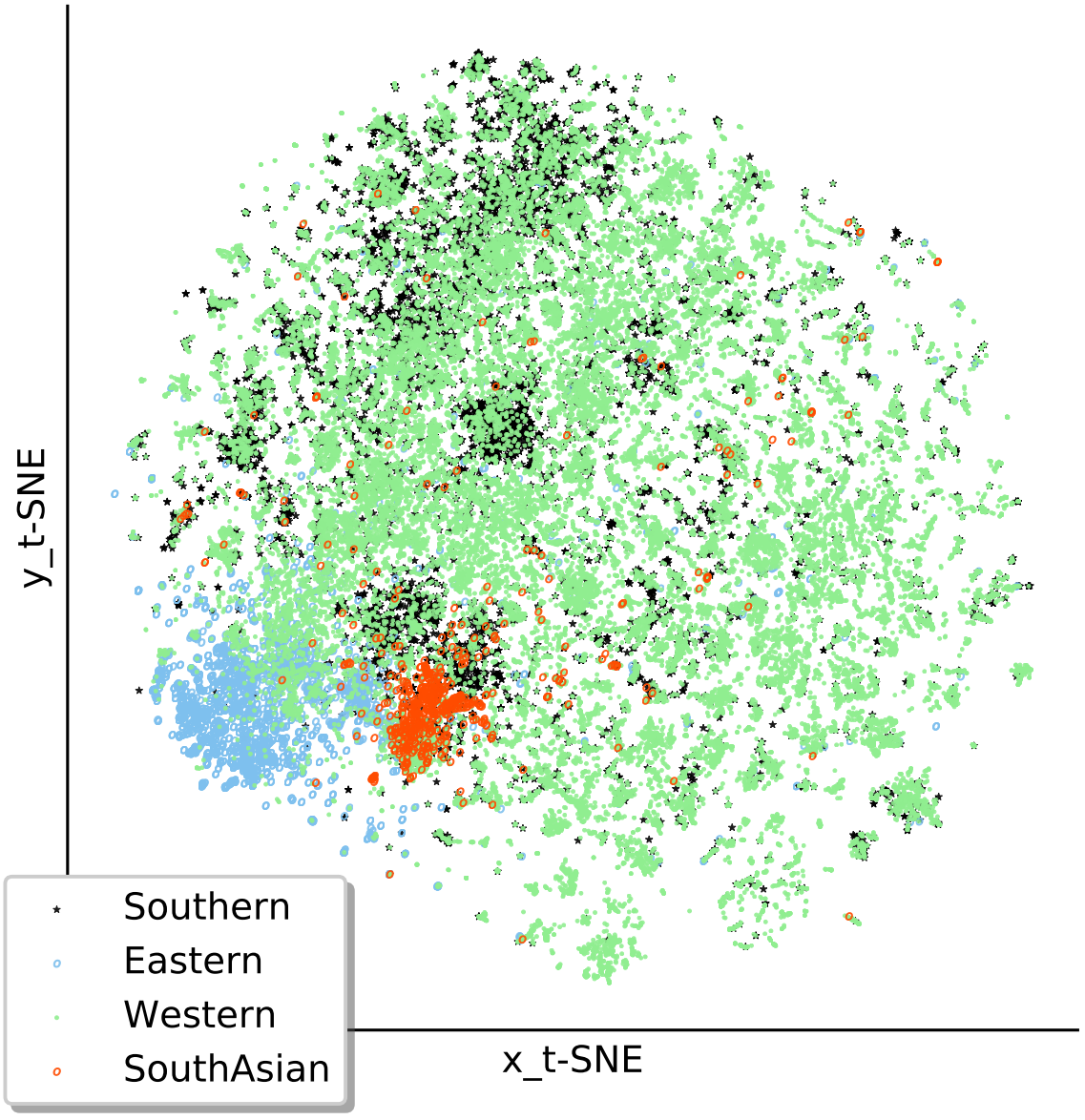
t-SNE Merged Cuisines Visualization TF-IDF Embedding (perplexity 50)



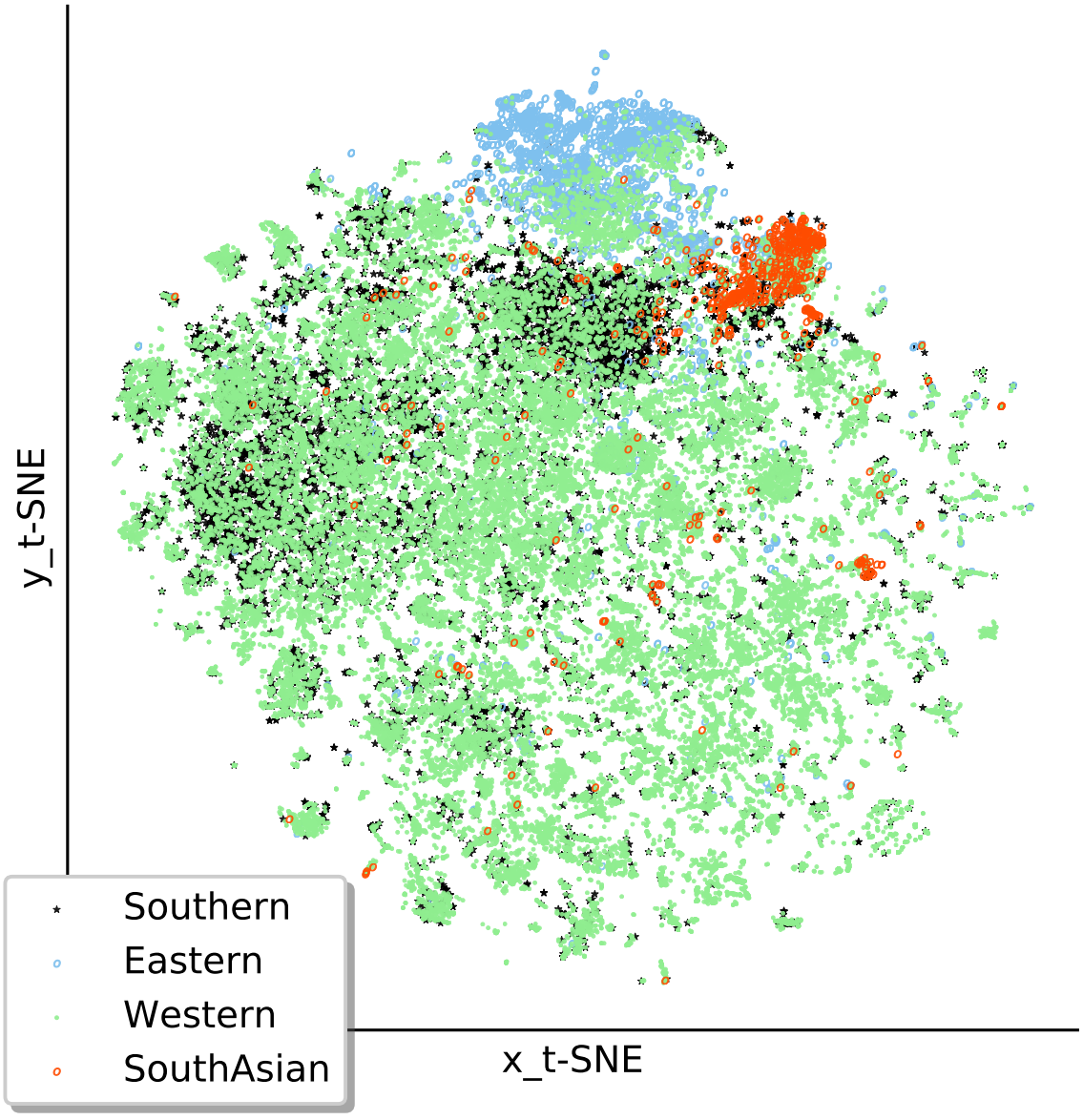
t-SNE Merged Cuisines Visualization TF-IDF Embedding (perplexity 50)



t-SNE Merged Cuisines Visualization TF-IDF Embedding (perplexity 50)



t-SNE Merged Cuisines Visualization
wiki_fasttext_concat Embedding (perplexity 50)



Google News and Wikipedia Corpus

- Google News
  - corpus size: around 100 billion words[1]
  - vocabulary size: 3 million words[2]
- Wikipedia
  - corpus size: 2-3 billion words[3]
  - vocabulary size: 1 million words[4]

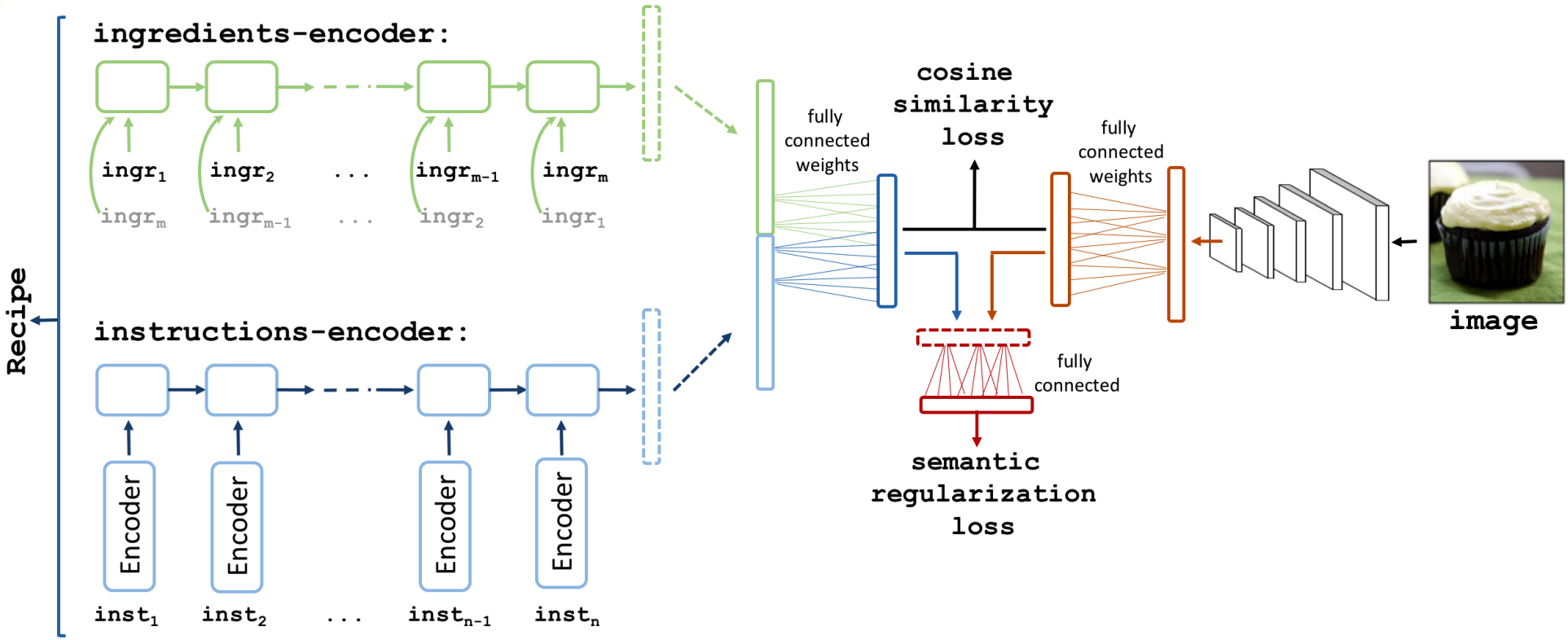
[1] mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/

[2] code.google.com/archive/p/word2vec/

[3] en.wikipedia.org/w/index.php?title=Wikipedia:Size_comparisons&oldid=707103846

[4] fasttext.cc/docs/en/english-vectors.html

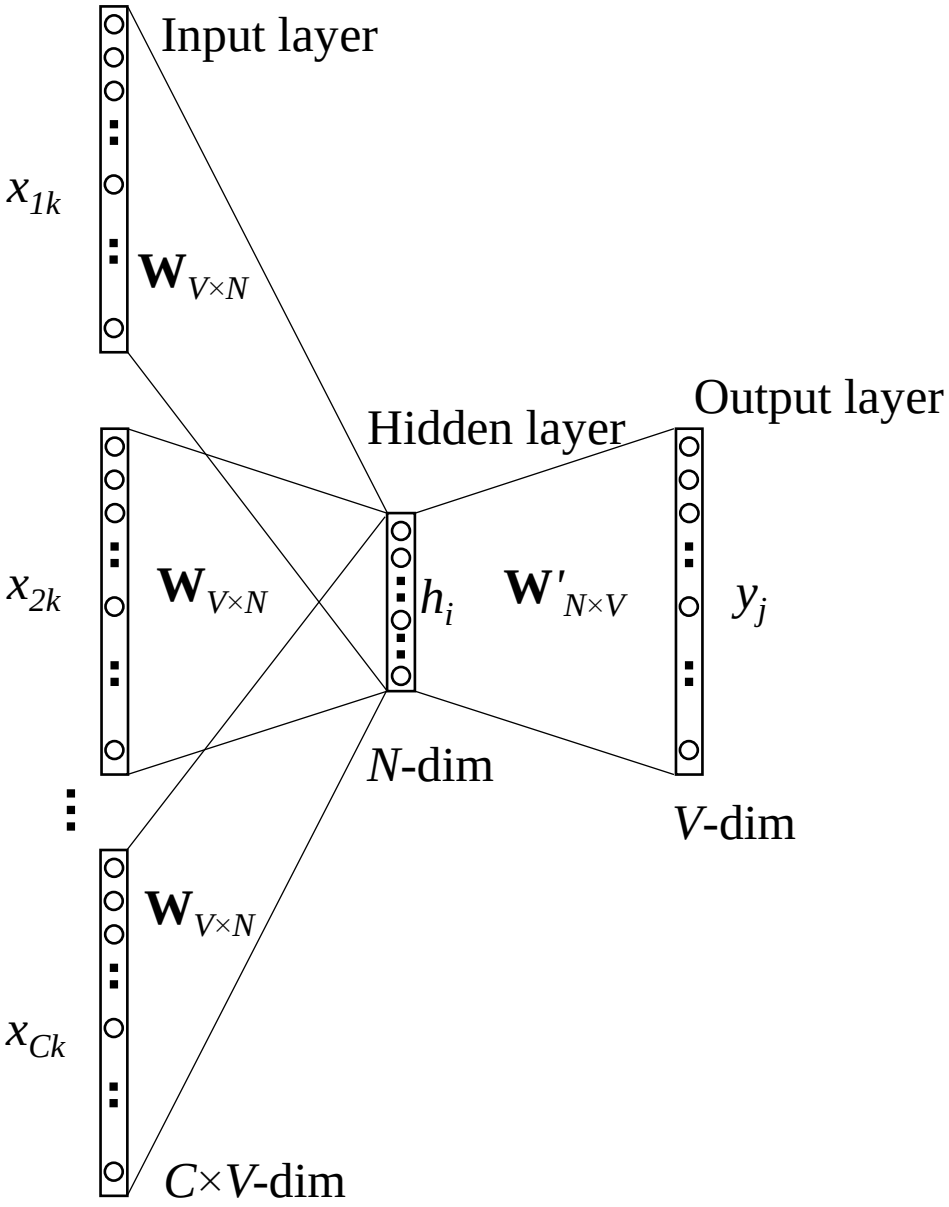
im2recipe Model [24]



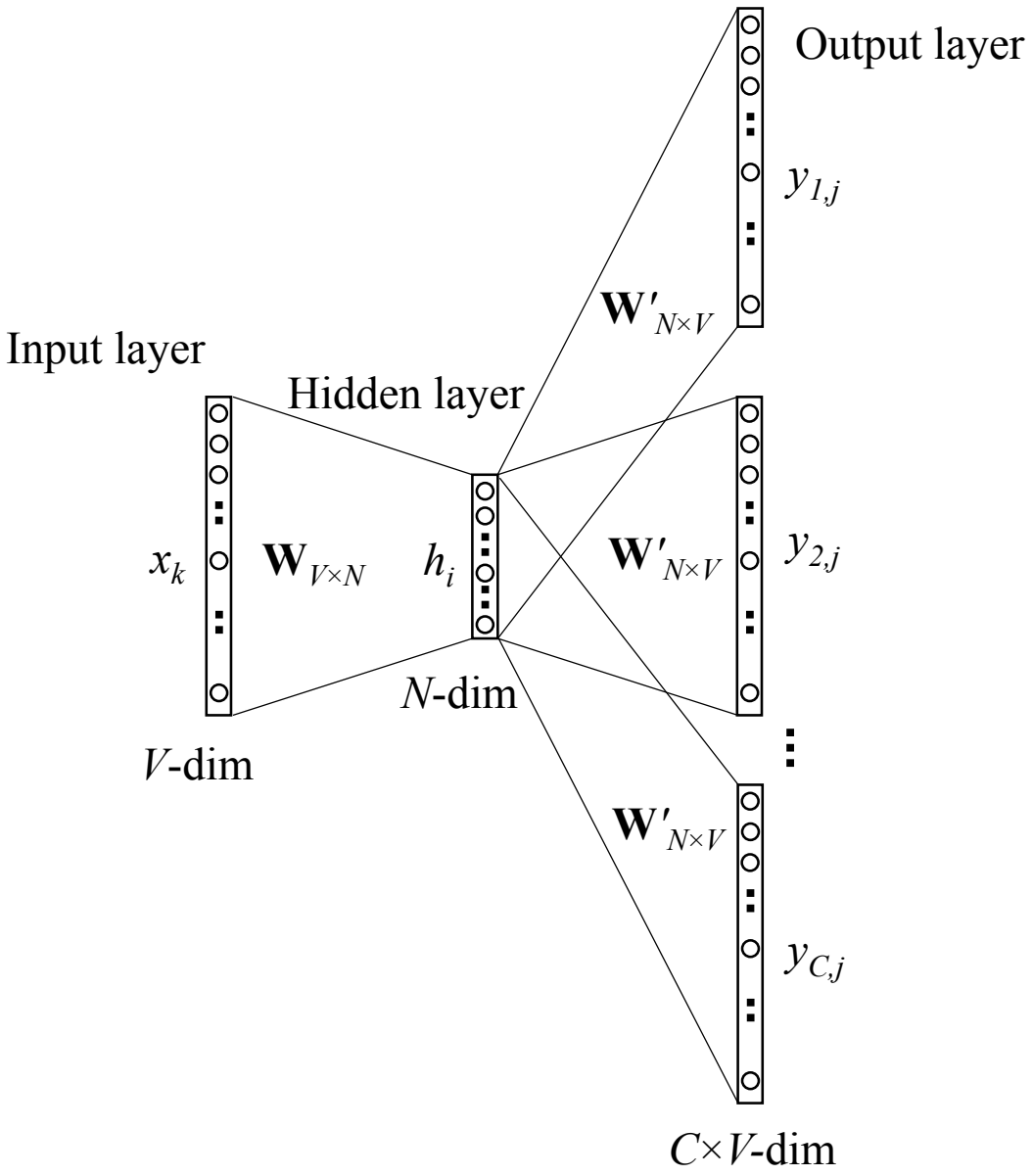
Continuous Bag-of-words And Skip-gram Model[28]

- the two main architectural parts of the dense "Word2Vec" embedding
- continuous bag-of-words
  - predicts word probability based on context words
  - vectors are array of real numbers, in contrast to the standard bag-of-words model as array of count frequencies
- skip-gram
  - assigns probabilities to context words for a given word
  - analogous to "n"-gram language models assigning probabilities to "n" token long word sequences

Continuous Bag-of-words Model[35]



Skip-gram Model[35]

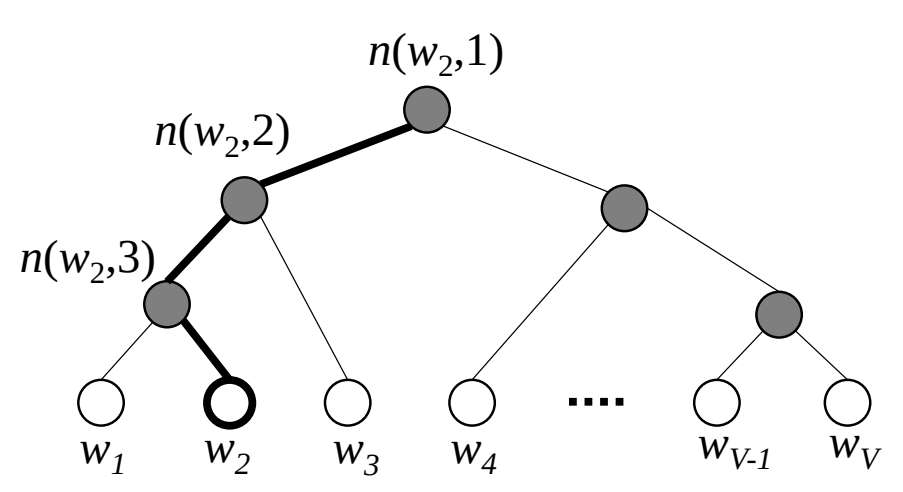


CBOW and Skip-gram Model[35]

In the case of "Skip-gram", given a set D of all word and context pairs extracted from the text, here the formula to maximize the probabilities $p(c|w)$ (probability of the context word c given word w) by adjusting the parameters θ of $p(c|w;\theta)$ ([16, p. 1]):

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta) \tag{8.1}$$

Hierarchical Softmax[28], figure by [35]



- efficient softmax computation using Huffman binary tree
- words are leaves
- instead of computing softmax over all word probabilities, follow probability mass path

Negative Sampling[27, 35, cf. p. 13]

- instead of updating all output vectors, update only a sample
- negative and positive examples: words outside or within context

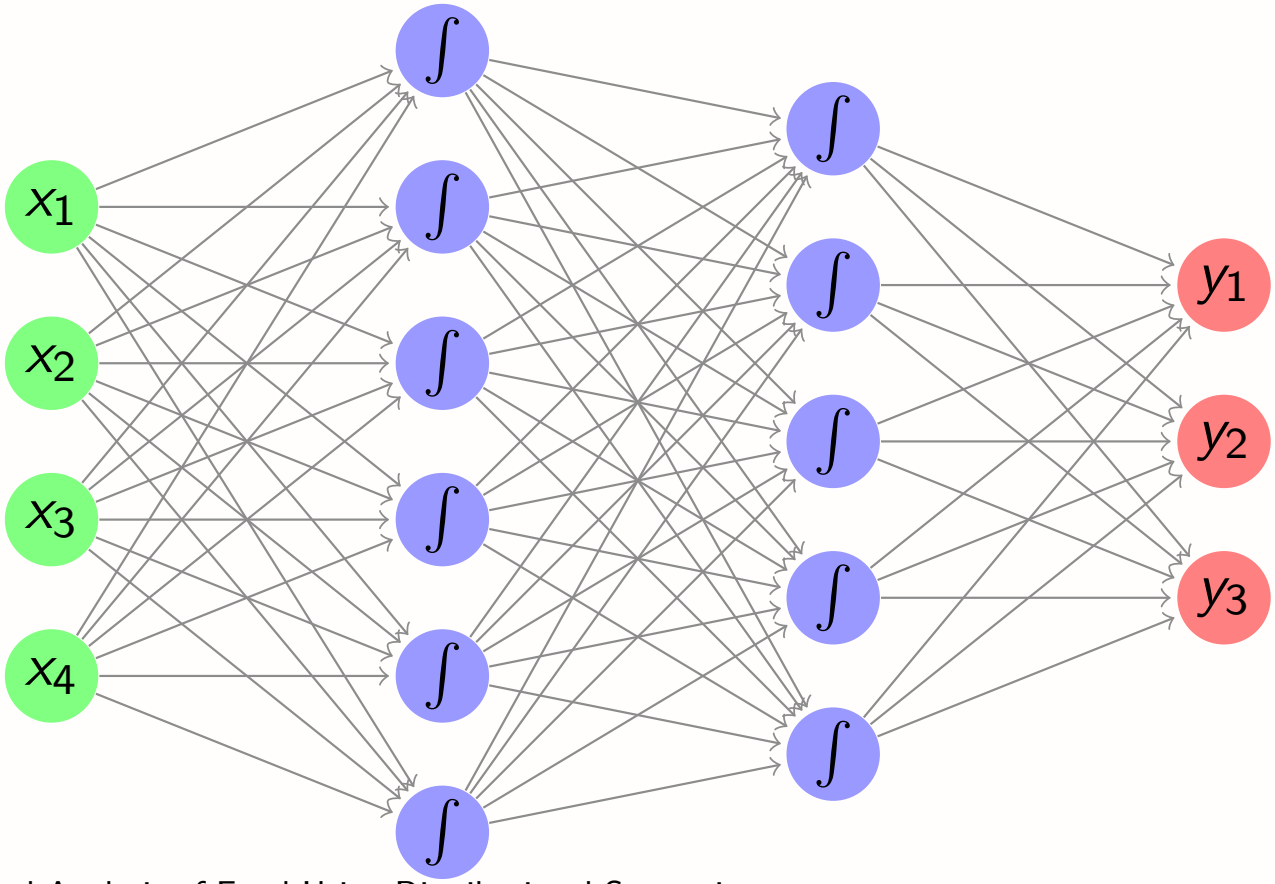
Feed-forward Neural Network With Two Hidden Layers[15, p. 13]

$$
\begin{aligned}
NN_{MLP2}(\mathbf{x}) &= \mathbf{y} \\
\mathbf{h^1} &= g^1(\mathbf{x}\mathbf{W^1} + \mathbf{b^1}) \\
\mathbf{h^2} &= g^2(\mathbf{h^1}\mathbf{W^2} + \mathbf{b^2}) \\
\mathbf{y} &= \mathbf{h^2}\mathbf{W^3}
\end{aligned}
\tag{8.2}
$$

- $\mathbf{y}$ is the output layer, $\mathbf{h^1}$ and $\mathbf{h^2}$ are the two so called hidden layers.
- $\mathbf{W}$ and $\mathbf{b}$ == matrix and bias weights of the linear input transformation
- g is the non-linear so called *activation function*)

Feed-forward Neural Network With Two Hidden Layers[15, fig. 2]

Input layer Hidden layer Hidden layer Output layer



fastText Subword-Embedding[8]

- even rare and unknown words but with known *n*-grams can be embedded; requires less corpus data to train
- words are represented by a bag of character *n*-grams of size G (cf. [37])
- $\mathcal{G}_w \subset \{1, \ldots, G\}$: the set of *n*-grams appearing in a given word w
- each *n*-gram g assigned a vector representation $\mathbf{z}_g$
- word embedding equals sum of hashed *n*-gram vectors

$$s(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_c. \tag{8.3}$$

Linear and Non-Linear Classifiers: Optimization Problem[43]

Decision function d:

- $d(\mathbf{x}) \equiv [w]^T \phi(\mathbf{x}) + b$
  - w, b = trainable weight & bias == hyperplane parameters, p.e. 1-D lines in 2-D space
  - ϕ = feature space mapping function
  - for linear classifiers $\phi(x) = x$
  - non-linear *phi* is higher dimensional

Loss function (hinge loss)[5]:

- $\ell(y) = max(0, 1 - t * y)$
  - intended output $t = \pm 1$
  - prediction y

[1] https://en.wikipedia.org/wiki/Hinge_loss

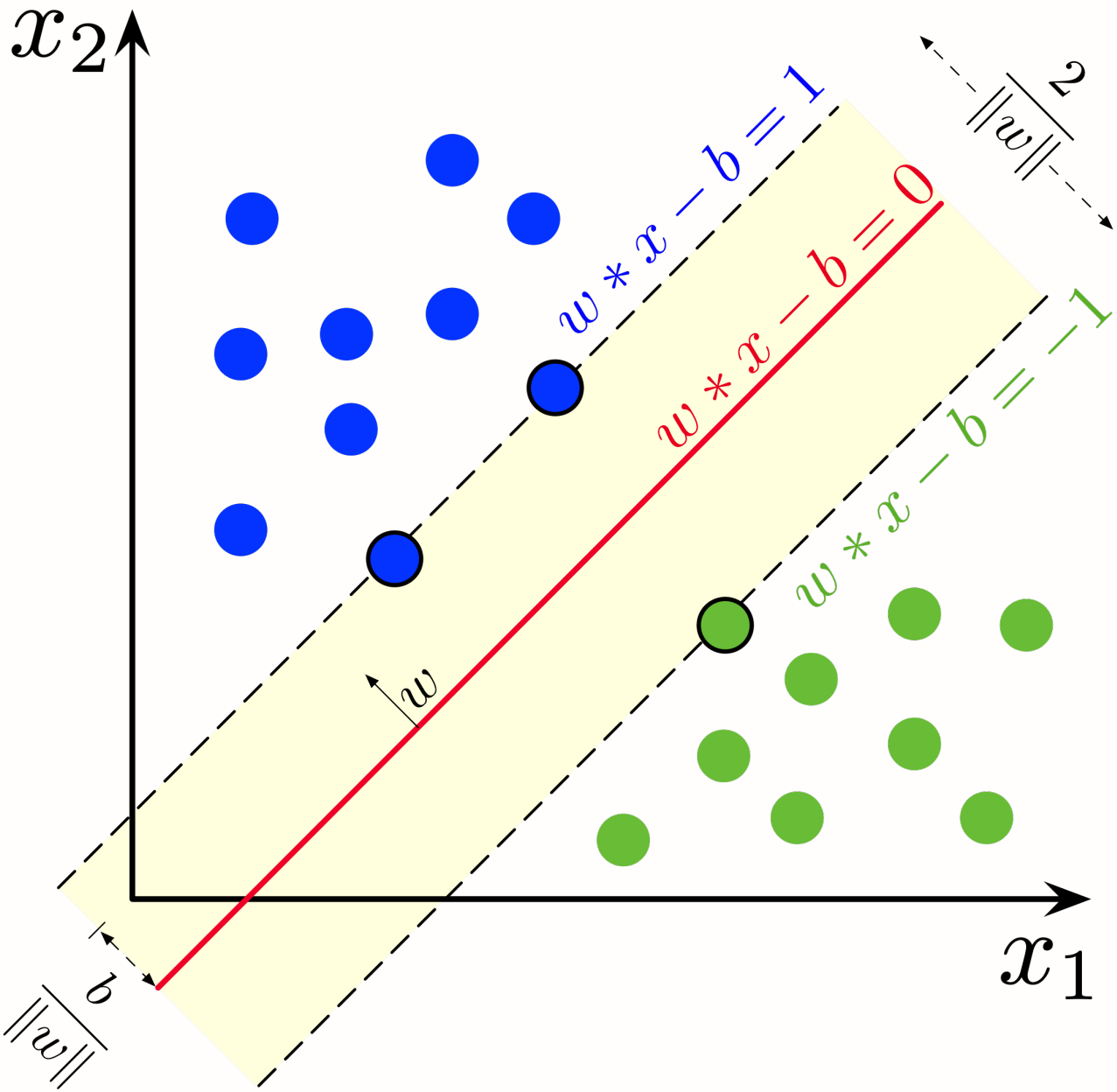
Linear and Non-Linear Classifiers: Optimization Problem[7], [43] (cont.)

$$C \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i) + \Omega(w)$$

- C = regularization parameter to emphasize prediction errors
- $\mathcal{L}$ = loss function with model parameters
- Ω = penalty function, i.e. linear L1 or quadratic L2
- Lagrangian "dual" optimiziation
  - better when n_samples $>$ n_features[6]

[1] scikit-learn.org/stable/auto_examples/svm/plot_svm_scale_c.html
[2] ...org/stable/modules/generated/sklearn.svm.LinearSVC.html

Support Vector Machine Hyperplane[22]



Topic Model Parameters[7, 4, p. 28]

- Dirichlet priors α and β influencing distributions
- probability distributions sum up to 1
- estimated using Bayesian statistics: the belief or expectation of states measured in probability
- α
  - the larger, the more significant topics per document[4, cf. p. 28]
- β
  - the larger, the more words per topic[4, cf. p. 28]
  - the smaller, the broader the topics[4, cf. p. 28]
  - "With scientific documents, a large value of β will lead the model to find a relatively small number of topics, perhaps at the level of scientific disciplines, whereas smaller values of β will produce more topics that address specific areas of research."[17, p. 5231]

Top Ten Words of Topic Number 1-10

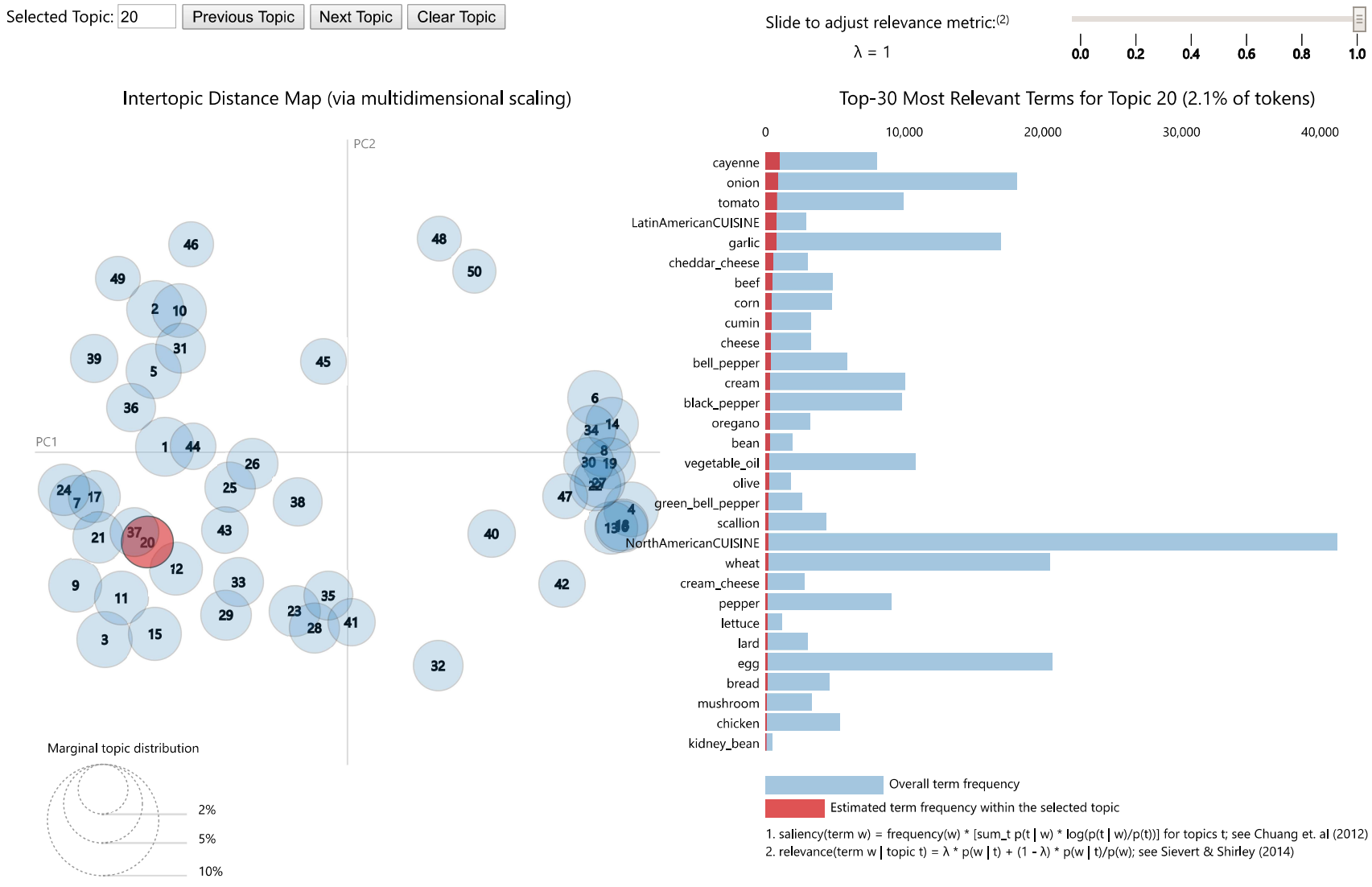
Topic Model №1 (α=1/k, β=0.1, 50 Topics)

№	Top ten words
1	butter NorthAmericanCUISINE olive_oil garlic chicken_broth mushroom parmesan_cheese cream white_wine SouthernEuropeanCUISINE
2	NorthAmericanCUISINE wheat egg butter vanilla cocoa cane_molasses milk lard walnut
3	onion black_pepper carrot garlic beef thyme WesternEuropeanCUISINE NorthAmericanCUISINE butter olive_oil
4	cayenne cilantro lime_juice LatinAmericanCUISINE onion garlic tomato olive_oil NorthAmericanCUISINE avocado
5	egg wheat butter milk almond vanilla SouthernEuropeanCUISINE WesternEuropeanCUISINE cream cinnamon
6	NorthAmericanCUISINE wheat egg cinnamon vanilla vegetable_oil butter cane_molasses walnut oat
7	garlic SoutheastAsianCUISINE cayenne fish lime_juice cilantro vegetable_oil ginger rice soy_sauce
8	cayenne onion tomato LatinAmericanCUISINE garlic cheddar_cheese beef corn cumin cheese
9	wheat NorthAmericanCUISINE yeast egg butter milk vegetable_oil whole_grain_wheat_flour honey seed
10	vinegar cane_molasses NorthAmericanCUISINE onion tamarind black_pepper mustard garlic vegetable_oil egg

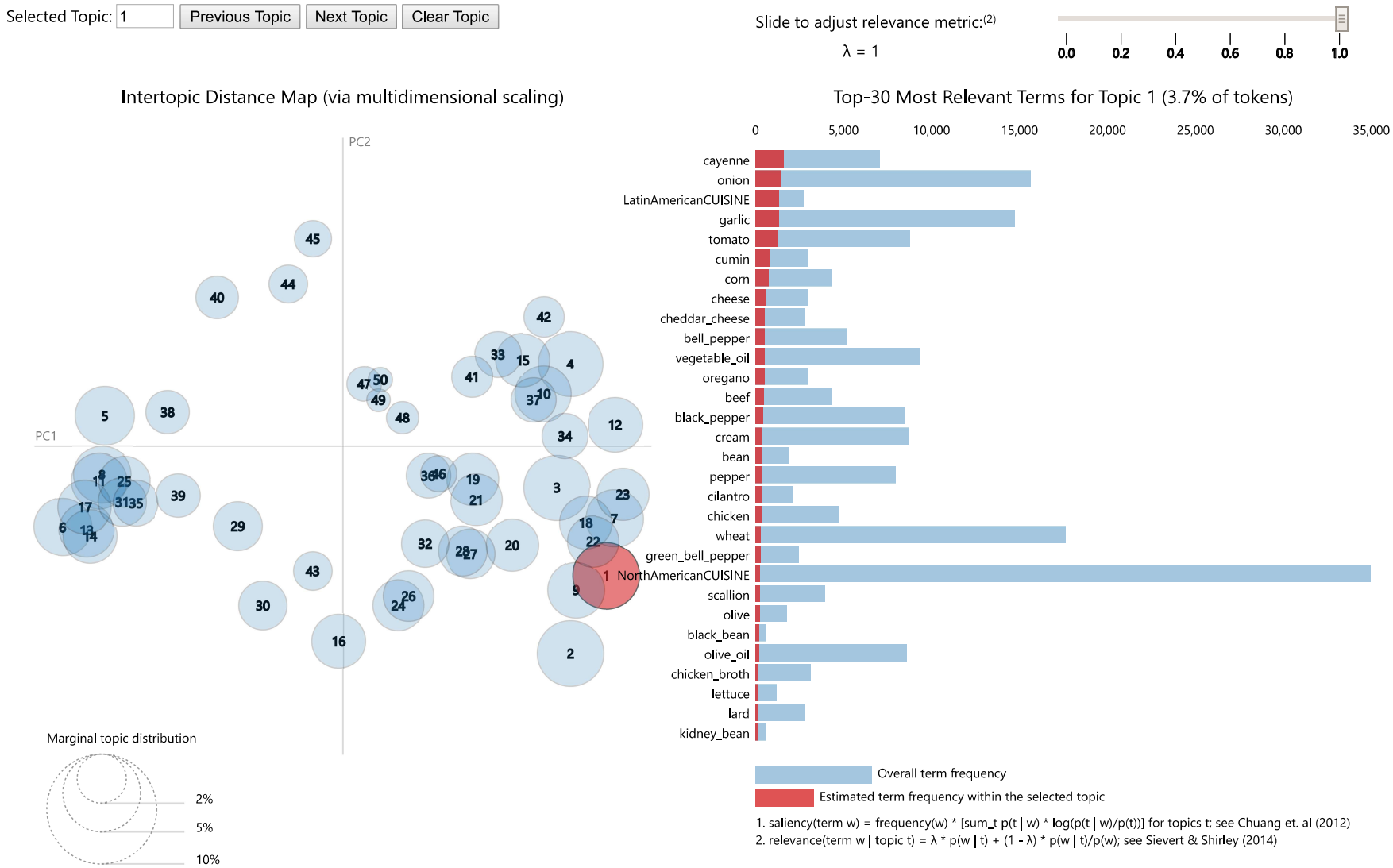
Topic Model №2 (α=1/k, β=1/10k, 50 Topics)

№	Top ten words
1	butter NorthAmericanCUISINE olive_oil garlic chicken_broth parmesan_cheese cream white_wine mushroom SouthernEuropeanCUISINE
2	NorthAmericanCUISINE wheat egg milk butter cocoa vanilla cream coffee vegetable_oil
3	NorthAmericanCUISINE lemon lime orange_juice orange pineapple lime_juice rum lemon_juice honey
4	NorthAmericanCUISINE butter WesternEuropeanCUISINE cream bread onion cream_cheese rice egg olive
5	NorthAmericanCUISINE wheat egg butter milk cream vanilla lemon_juice lemon cream_cheese
6	NorthAmericanCUISINE egg wheat cinnamon butter milk nutmeg cane_molasses vanilla ginger
7	cumin turmeric coriander pepper fenugreek onion garlic SouthAsianCUISINE vegetable_oil cayenne
8	cayenne onion LatinAmericanCUISINE garlic tomato cumin corn cheese cheddar_cheese bell_pepper
9	NorthAmericanCUISINE wheat egg cinnamon butter vanilla walnut cane_molasses vegetable_oil raisin
10	NorthAmericanCUISINE onion thyme celery sage black_pepper butter rosemary chicken_broth bread

Interactive pyLDAvis HTML Visualization Topic Model №1



Interactive pyLDAvis HTML Visualization Topic Model №2

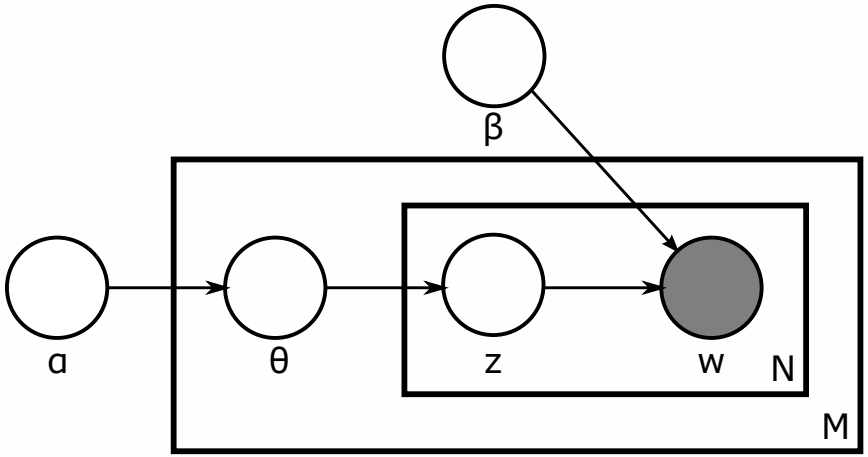


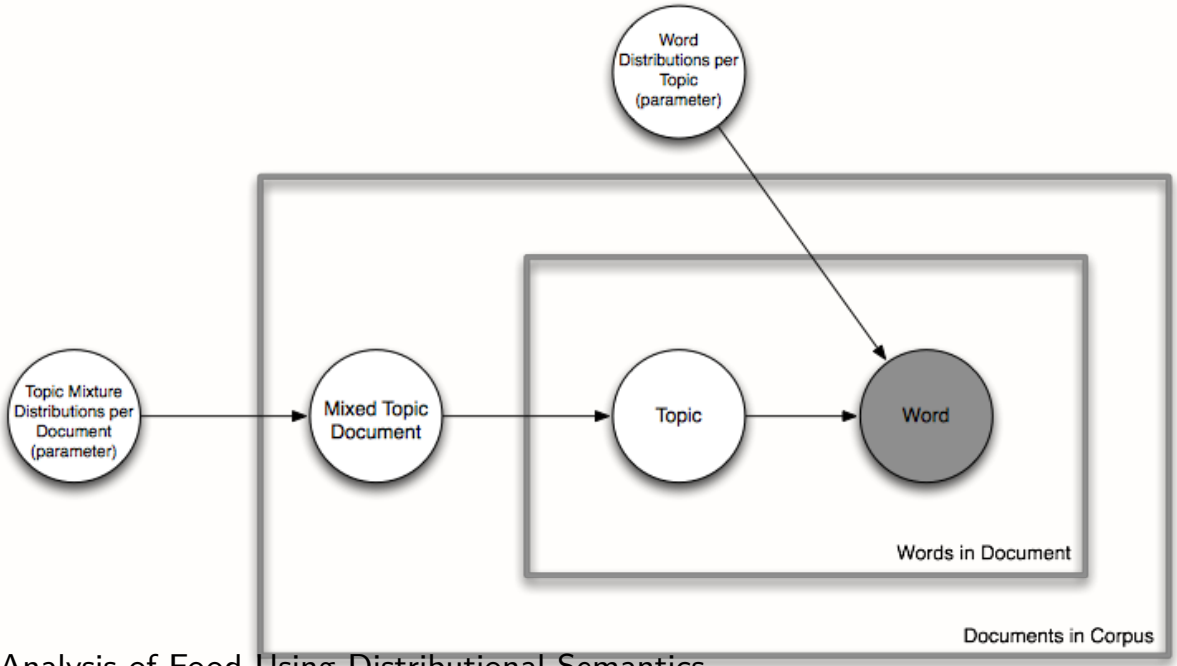
Topic Modeling with Latent Dirichlet Analysis[7]

- A topic is a word probability distribution [26, cf. for detailed explanation]
- Select a number Z as the quantity of topics to use.
- For each document in the corpus do the following steps [26]:
  - select a topic mixture (document-topic) distribution from a Dirichlet distribution influenced by α.
  - For each word in the document do the following steps:
    - select a topic from the document-topic distribution
    - select a word from the topic-word distribution selected above influenced by β.

Topic Model Plate Notation[7, p. 997, 5, figure as .svg, 26, 2nd figure]

square=loop, circles=variables, shaded=observed, white=latent[26, cf.]





Topic Model [7]

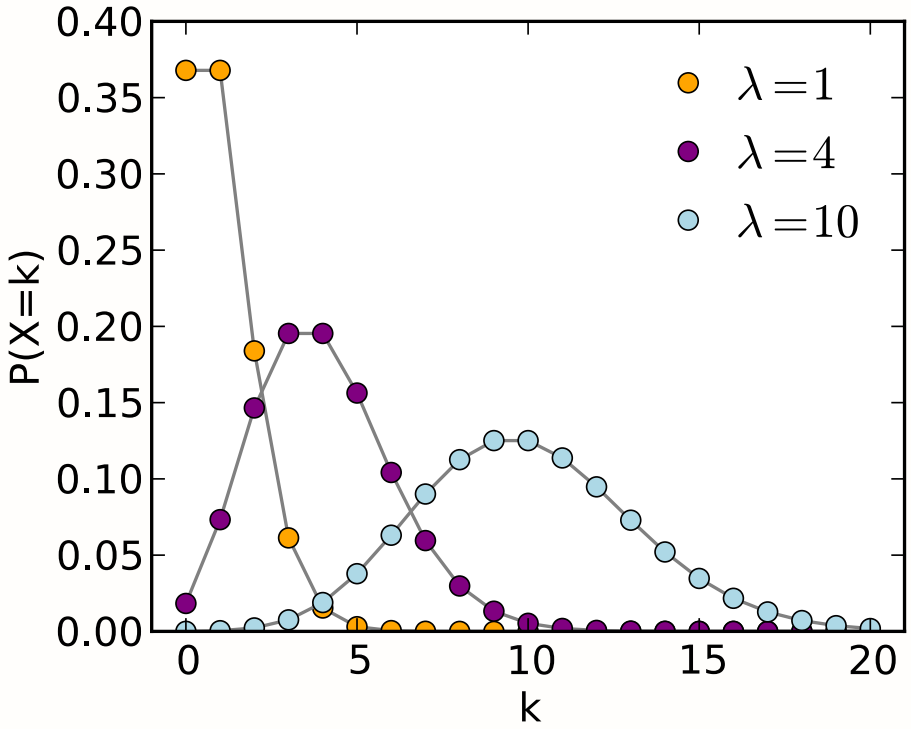
As described by [7, p. 996] Latent Dirichlet Allocation "assumes the following generative process for each document **w** in a corpus D:"

1. Choose a sequence of N words $\sim$ $Poisson(\xi)$.
2. Choose $\theta \sim Dirichlet(\alpha)$.
3. For each of the N words w_n:

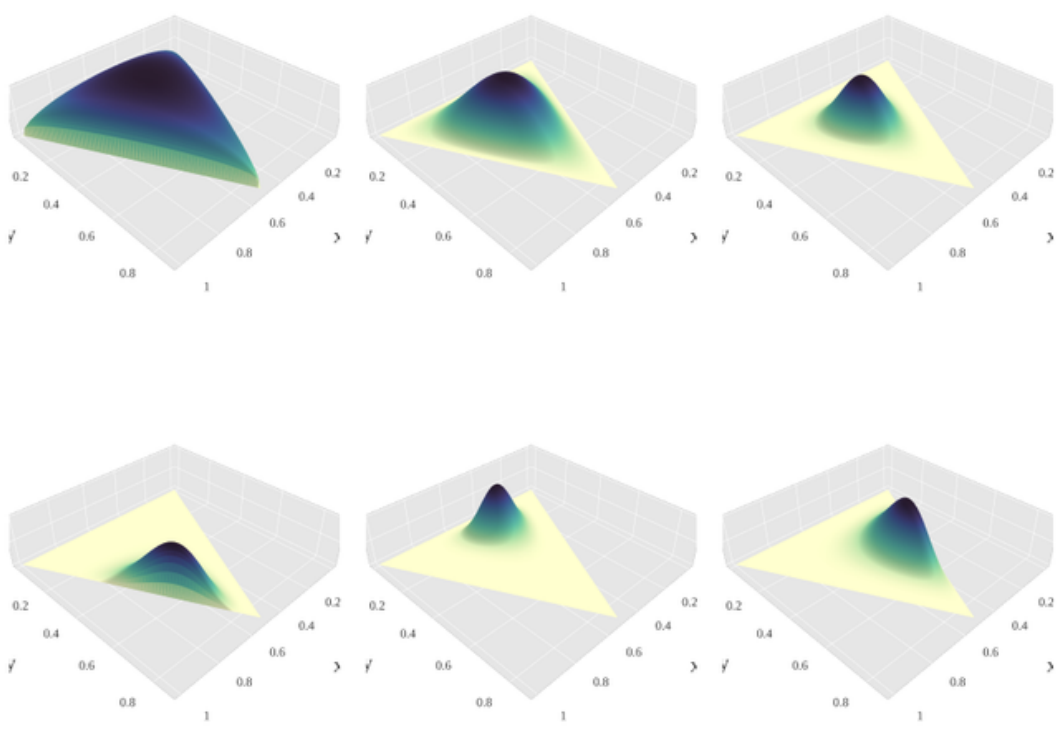
(a) Choose a topic $z_n \sim Multinomial(\theta)$.

(b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Topic Modeling: Draw Number of Topics from Poisson Distribution, figure by [40]



Topic Modeling: Draw Number of Topics from Dirichlet Distribution[36, cf.], figure by [13]



- 3-D Dirichlet probability density function
- Two dirichlet processes to generate words: 1x get a topic 1x get a word given this topic
- Here k $= 3 ==$ number of topics

Kullback-Leibler Divergence (cf. [3])

- difference between true and approximate entropy of a probability density function[14, p. 4]
  - topic specificity: distance between topic-word distribution to uniform distribution[8]
- goal: minimize variance between divergence values [3, p. 400]
- entropy
  - entropy $p_1, p_2 \ldots, p_n$ is the amount of uncertainty about the experiment outcome [34, p. 547, 38]
  - measured in $H(p_1, p_2 \ldots, p_n) = \sum_{k=1}^{n} p_k log_2 \frac{1}{p_k}$
  - approximate entropy or cross-entropy with s traversing all possible states v of $P(v)$ and $P'(v)$[14, p. 4]:
  - $H' = -\sum_s P(s) log(P'(s))$

[1] mallet.cs.umass.edu/diagnostics.php

Topic Model Metrics: Average Correlation Between Topic-Pairs[10]

- goal: minimize average overlap between document-topic vector pairs A and B using cosine simiarity
- similar results as minimizing perplexity (equivalent to inverse of geometric mean per-word likelihood[10, p. 1780])
- formula:

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Topic Model Metrics: Logistic Likelihood (cf. [9])

- goal: maximize mean of observed document logistic likelihoods[9]
- often done on separate test set (cf. [41])
- often negatively correlated to human judgement[11]

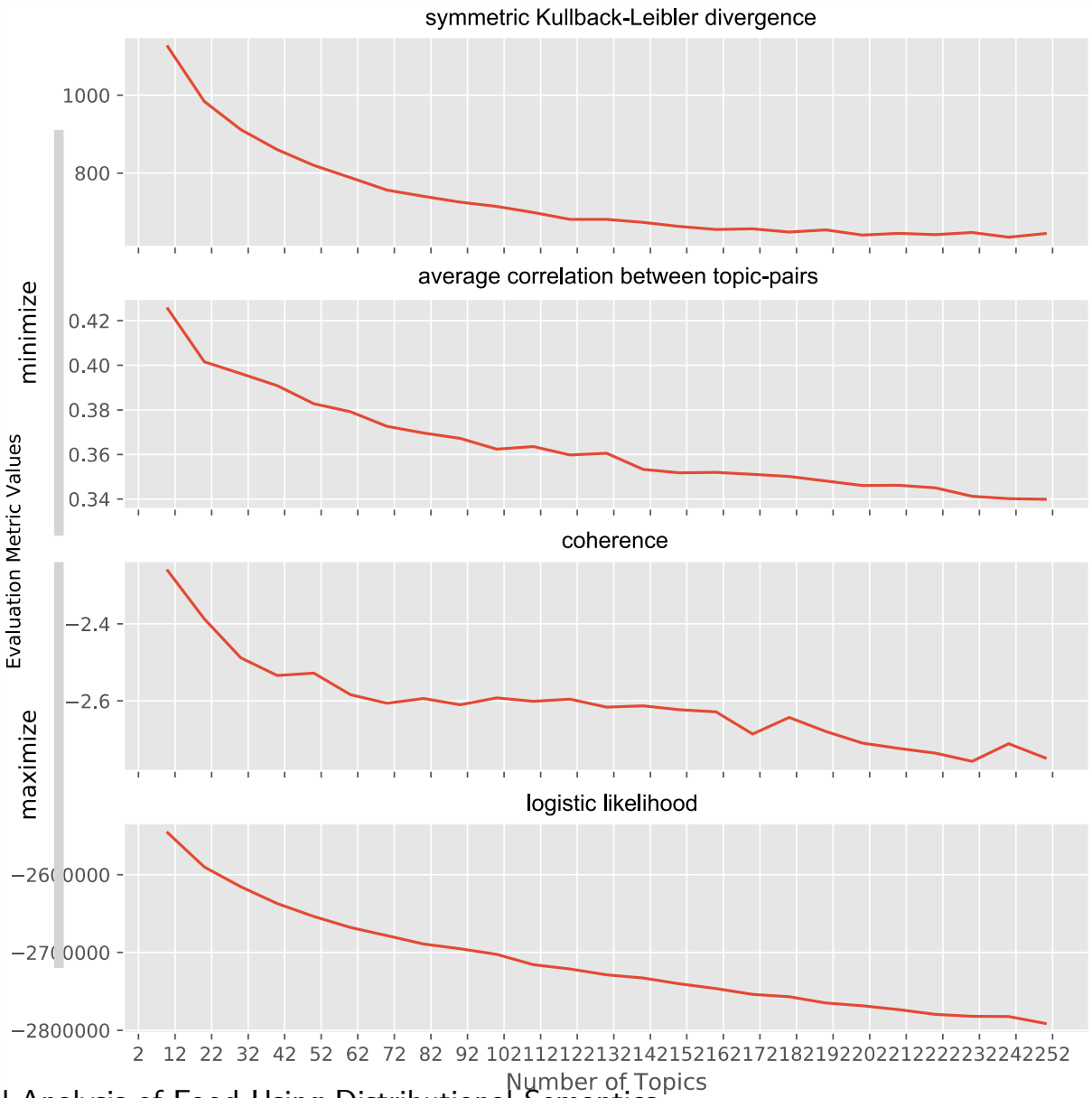
[1] datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/

Topic Model Metrics: Coherence[29]

- goal: maximize coherence per topic[29, p. 265] using:
  - correlates very well to human intuition
  - evaluation using word intrusion and topic intrusion[11]:
    - dog, cat, horse, apple, pig, cow
    - apple is intruding as the other words make sense together
  - use of word co-ocurrence:
    - list of most probable words in each topic $V^{(t)}$
    - logistic document-word frequency $D(v)$ (the number of documents containing at least one word type v)
    - logistic co-document-word frequency (the number of documents containing at least one word type v and at least one word type v$\prime$)
    - $C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)})+1}{D(v_l^{(t)})}$

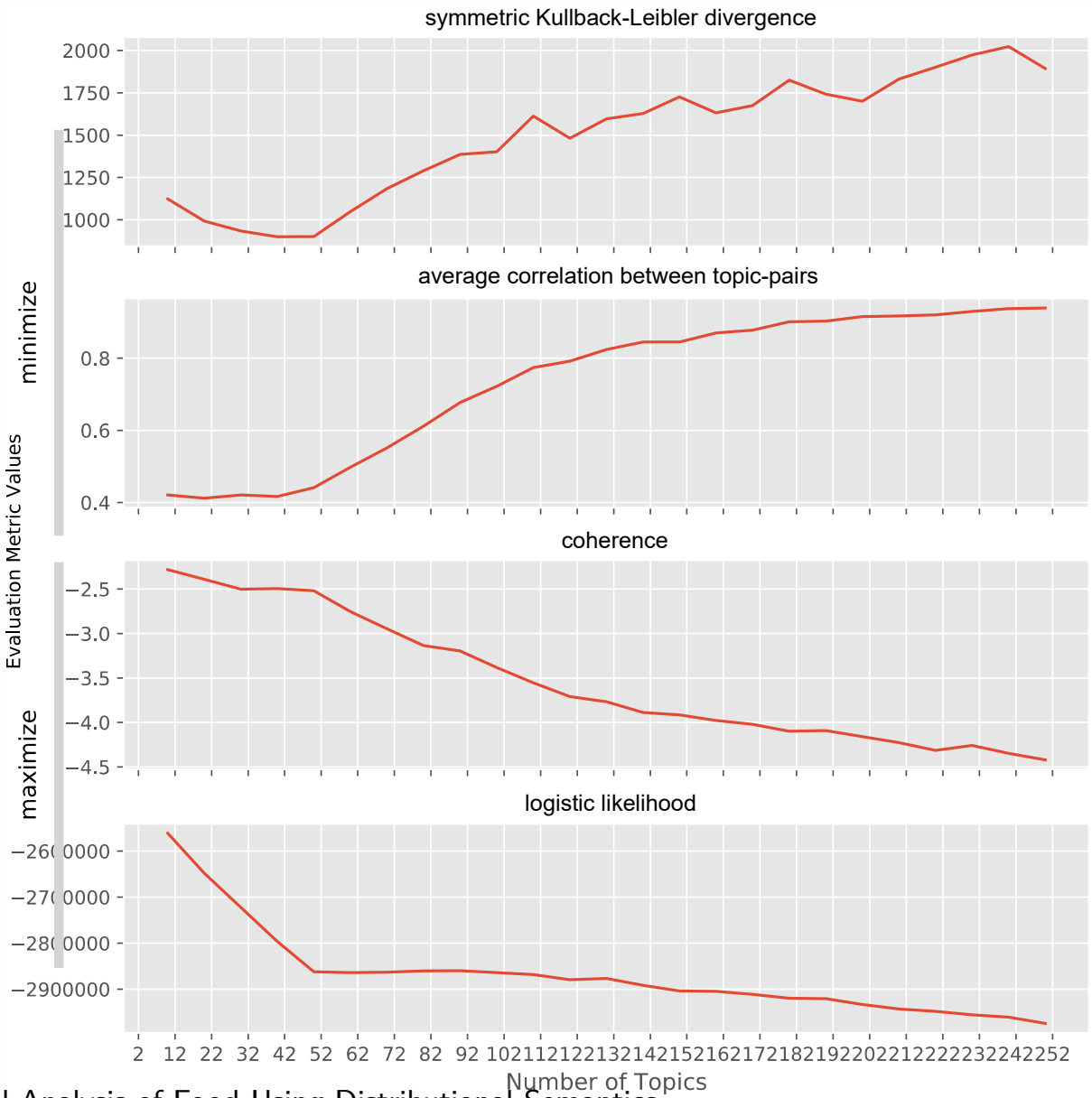
LDA-Metrics Topic Model 1

Parameters: α=1/k, β=0.1 for 2-252 Topics



LDA-Metrics Topic Model 2

Parameters: α=1/k, β=1/10k for 2-252 Topics



Least Salient Words and Most Common Words Topic Model №1

	TD-IDF: Most Common Words	Topic Model №1: Least Salient Words
1	NorthAmericanCUISINE	wheat
2	egg	egg
3	wheat	milk
4	butter	vanilla
5	onion	butter
6	garlic	olive_oil
7	milk	vinegar
8	vegetable_oil	NorthAmericanCUISINE
9	cream	tomato
10	tomato	garlic

TF-IDF = Term Frequency / Inverse Document Frequency

Distinctiveness = How informative is a word for a topic ([11, p. 75])

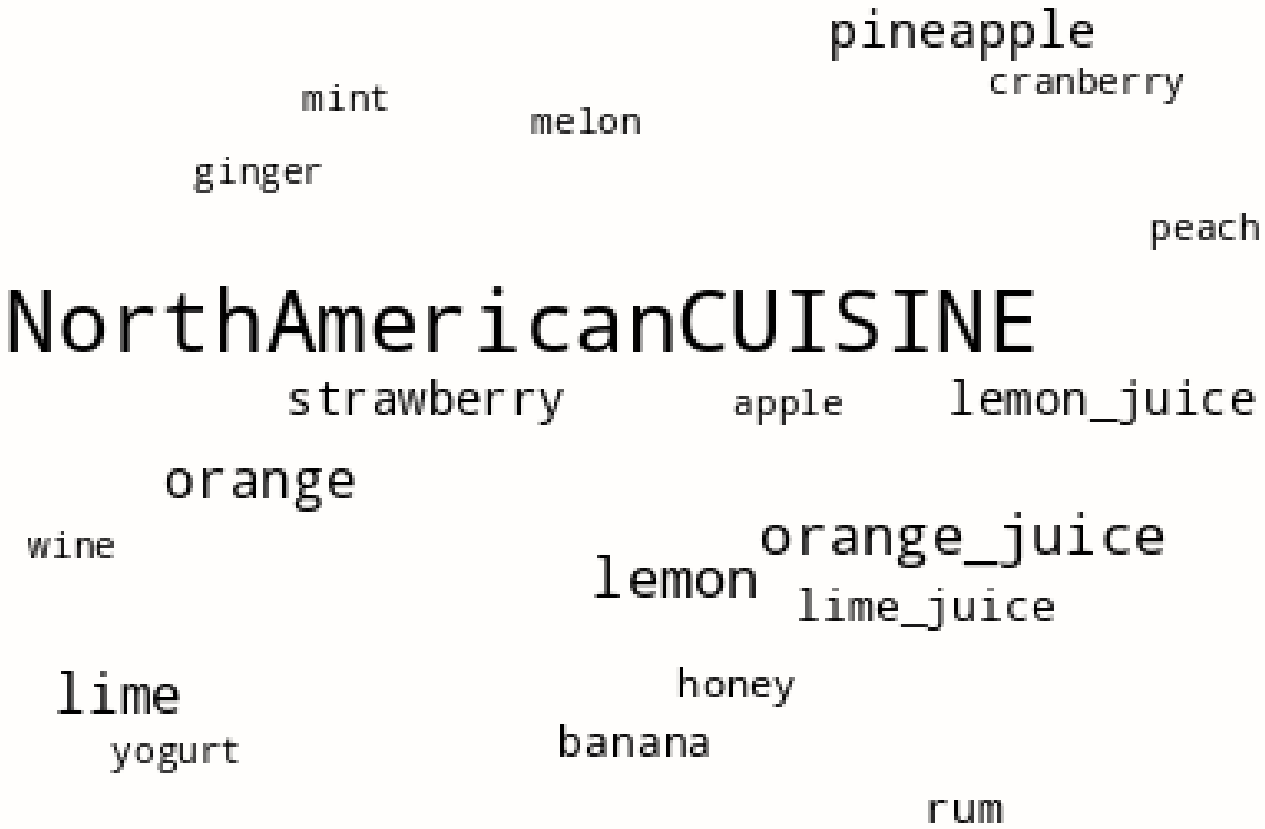
Saliency = Distinctiveness * Word Frequency

Topic №5 of Topic Model №1

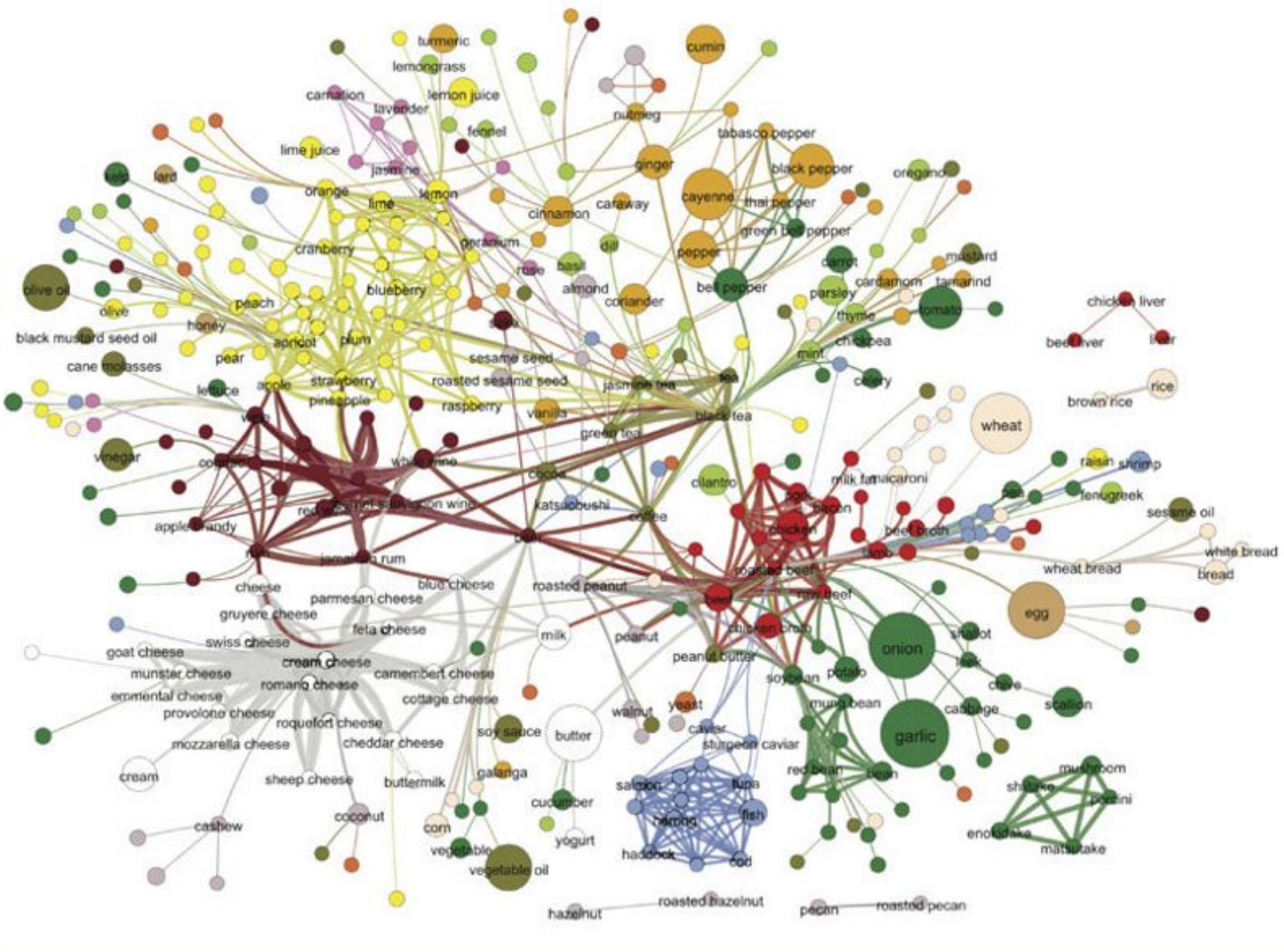
Topic №7 of Topic Model №1

Topic №11 of Topic Model №1

Topic №16 of Topic Model №1, (cf. [1, p. 2])



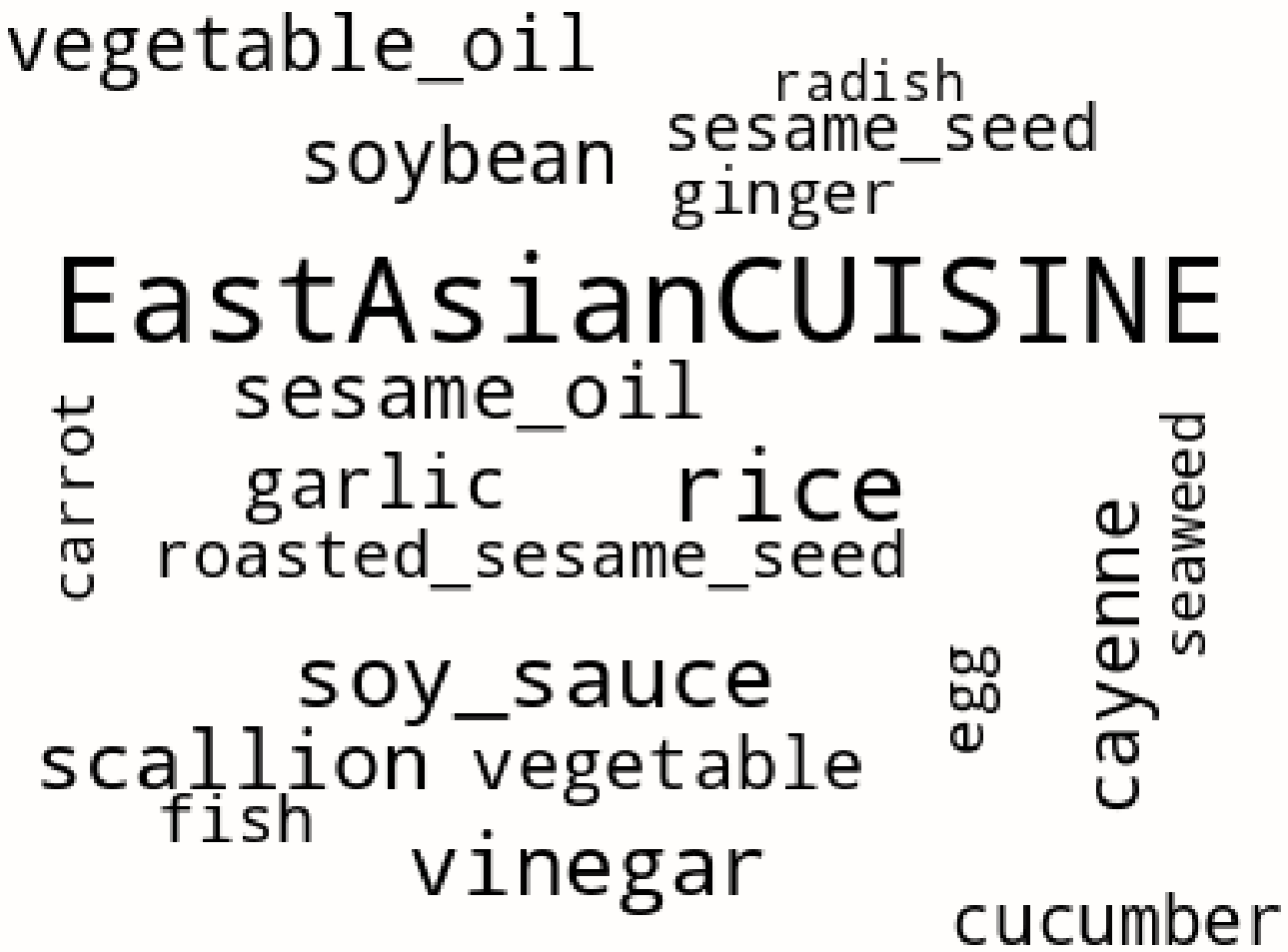
Flavor Network[1, p. 2]



Topic №17 of Topic Model №1

Topic №28 of Topic Model №1

Topic №30 of Topic Model №1 (cf. [31, p. 9])

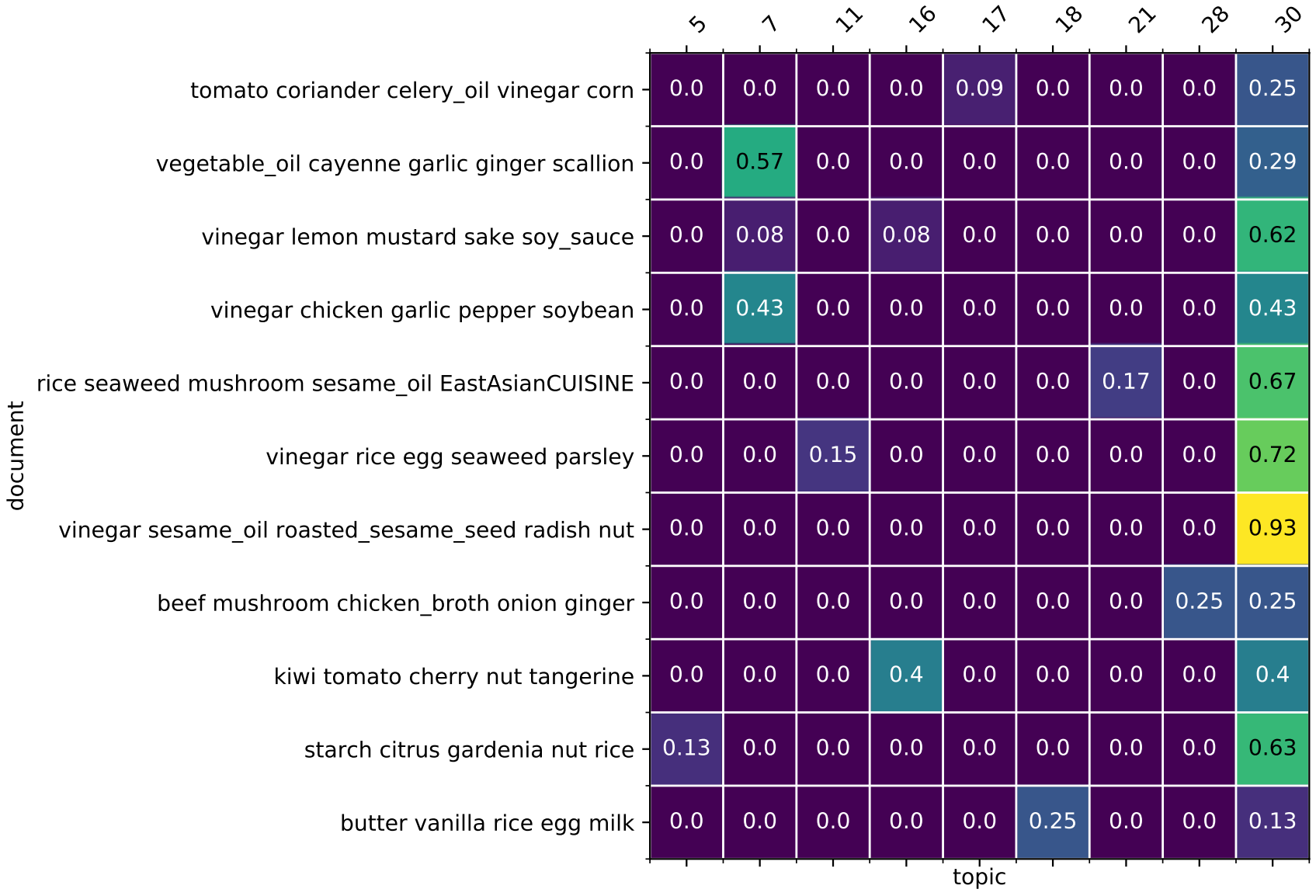


Yummly66K Corpus Topic Modeling[31, p. 9]

TABLE II: Some examples of discovered topics.

Topic ♯2	Topic ♯12	Topic ♯36	Topic ♯68	Topic ♯72
ground-cumin 0.106	butter 0.074	ricotta-cheese 0.066	soy-sauce 0.195	milk 0.204
chili-powder 0.074	flour 0.062	mozzarella-cheese 0.059	sesame-oil 0.079	honey 0.062
fresh-cilantro 0.067	ground-cinnamon 0.056	spaghetti 0.059	green-onion 0.066	cinnamon 0.053
vegetable-oil 0.046	vanilla-extract 0.042	ground-beef 0.046	scallion 0.060	vanilla-extract 0.043
sour-cream 0.043	salted-butter 0.039	italian-sausage 0.043	ginger 0.054	lemon 0.033
dried-oregano 0.041	milk 0.037	lasagna-noodle 0.042	fresh-ginger 0.049	ground-cinnamon 0.031
corn-tortilla 0.041	sugar 0.037	tomato-sauce 0.037	sugar 0.049	whipped-cream 0.026
diced-tomato 0.037	chopped-pecan 0.033	parmesan 0.033	rice-vinegar 0.039	brown-sugar 0.025
bell-pepper 0.030	sour-cream 0.029	pasta-sauce 0.024	vegetable-oil 0.037	orange 0.023
canola-oil 0.023	vegetable-oil 0.029	garlic-clove 0.023	water 0.036	coconut-milk 0.018
Topic ♯73	Topic ♯88	Topic ♯89	Topic ♯91	Topic ♯93
egg 0.107	noodle 0.059	mango 0.093	boneless-skinless-chicken 0.236	spanish-paprika 0.120
milk 0.099	mushroom 0.046	coconut-milk 0.085	chicken 0.102	pimento 0.058
bread 0.096	cabbage 0.035	milk 0.080	chicken-breast 0.089	smoked-paprika 0.048
cinnamon 0.072	sprout 0.034	ice-cube 0.0691	bell-pepper 0.040	green-olive 0.035
butter 0.063	ground-pork 0.030	water 0.064	onion 0.035	spanish-onion 0.035
vanilla-extract 0.050	star-anise 0.030	tea 0.0377	chicken-thigh 0.035	olive 0.034
maple-syrup 0.039	baby-bok-choy 0.029	coconut 0.034	boneless-chicken-breast 0.029	manchego-cheese 0.032
brown-sugar 0.035	bok-choy 0.028	thai-basil 0.023	shrimp 0.028	spanish-smoked-paprika 0.030
ground-cinnamon 0.035	chicken 0.026	sticky-rice 0.022	ground-pepper 0.027	ground-cumin 0.029
heavy-cream 0.022	firm-tofu 0.025	banana 0.021	beef 0.018	canola-oil 0.027

Topic heatmap of twelve "EastAsian" recipes of Topic Model №1

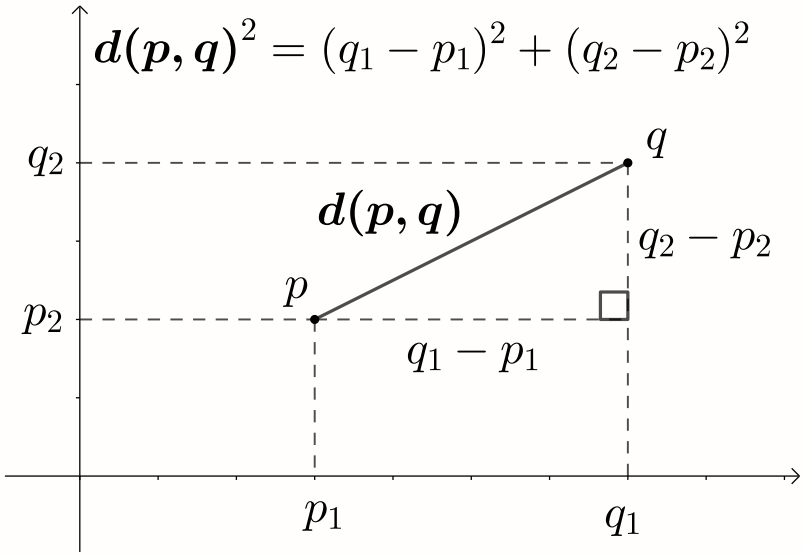


Topic heatmap of twelve "SouthernEuropean" recipes of Topic Model №1

Topic Modeling with Latent Dirichlet Analysis: Disadvantages

- many different parameters need to be tuned
- unsupervised learning: "no gold-standard list of topics to compare against for every corpus"[cf. p. 3][11]
- Dirichlet Distribution cannot model topic correlation:
  - "for example, a document about genetics is more likely to also be about disease than [...] astronomy"[6, p. 1]
- bag-of-words, i.e. no sentence structure preservation[7, cf. p. 995]
  - not as relevant here, as recipes are simple ingredient lists with no importance to ordering

t-Distributed Stochastic Neighbor Embedding[23, p. 2581] Dimensionality Reduction (figure by [19]



- SNE[2, cf.] converts "high-dimensional Euclidean distances" [...] "into conditional probabilities that represent similarities"[23, p. 2581]
- similarity of datapoint $\mathbf{x}_j$ to datapoint $\mathbf{x}_i$ is the conditional probability, $\mathbf{p}_{j|i}$, that $\mathbf{x}_i$ would pick $\mathbf{x}_j$ as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at $\mathbf{x}_i$.

t-Distributed Stochastic Neighbor Embedding[23, p. 2582]: Perplexity[10]

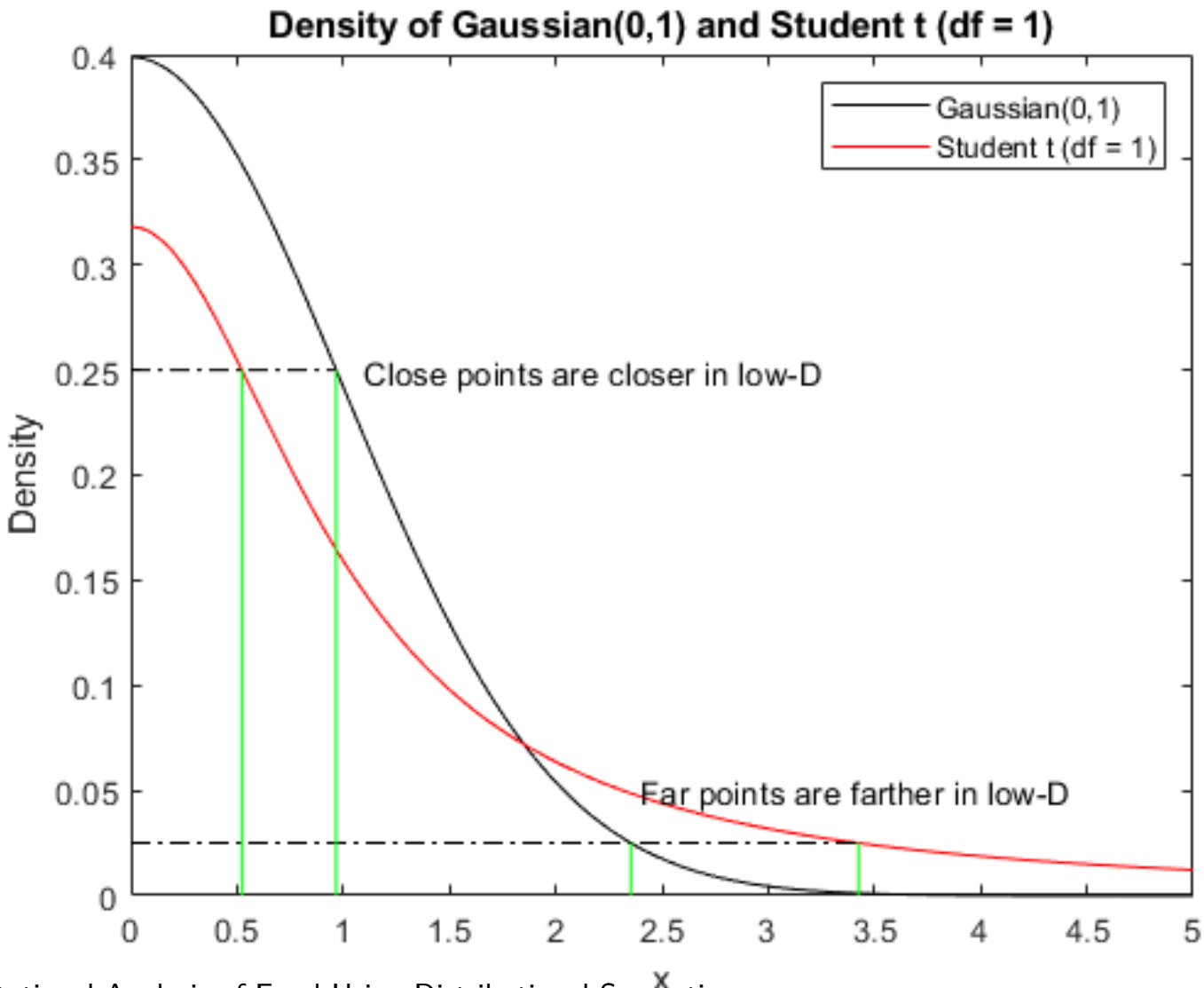
- perplexity of a fair k-sided die = k
- sets the number of effective nearest neighbors
- the denser or larger the data, the higher perplexity to use (rule of thumb)
- typical values between 5 and 50

[1] lvdmaaten.github.io/tsne/

t-Distributed Stochastic Neighbor Embedding (cont.)

- find a "low-dimensional counterparts y_i and y_j of the high-dimensional datapoints x_i and x_j"
- conditional probabilities in the higher and lower dimension p_{ij} and q_{ij} are jointly minimized using Kullback-Leibler divergence
- preserves as much "of the significant structure of the high-dimensional data as possible"[23, p. 2580]
- non-linear, contrary to p.e. Principal Component Analysis

"Student-t distribution" to alleviate crowding problems instead of Gaussian, figure by[25]



t-Distributed Stochastic Neighbor Embedding: Disadvantages

- cannot map unseen data
- parameters need to be tuned
- non-deterministic
- global similarity (inter-cluster) interpretability[23, cf. p. 2582]