

DATA SCIENCE APPLICATIONS

MAY MERKLE-TAN, PHD

Recipe difficulty tagger

Building a Menu Planner Helper App

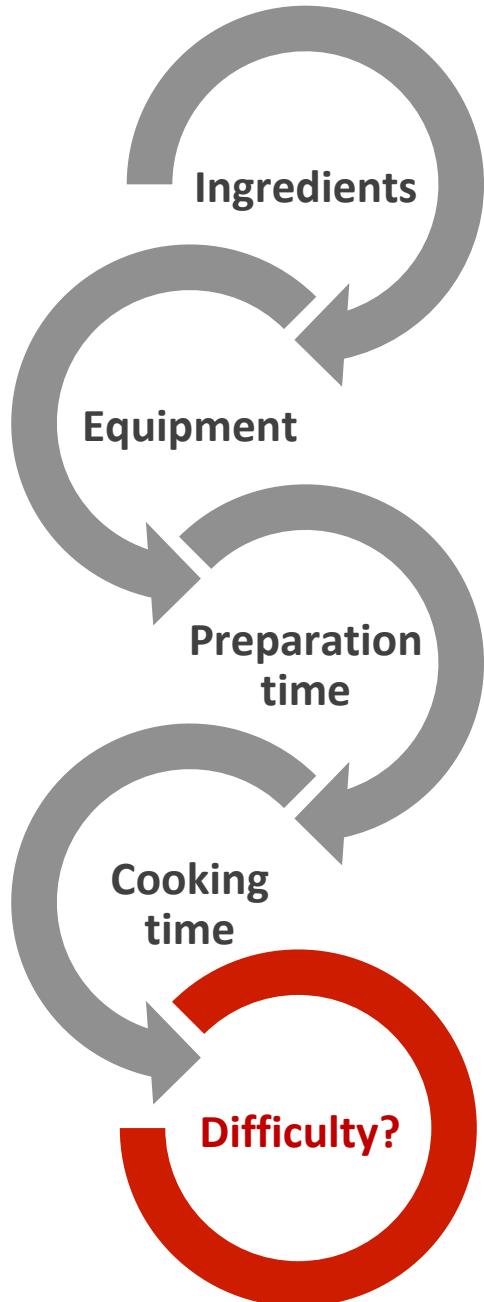
~ RECOMMENDER SYSTEM ~

Menu Planning & Preparation



Multiple Chefs | Multi-tasking



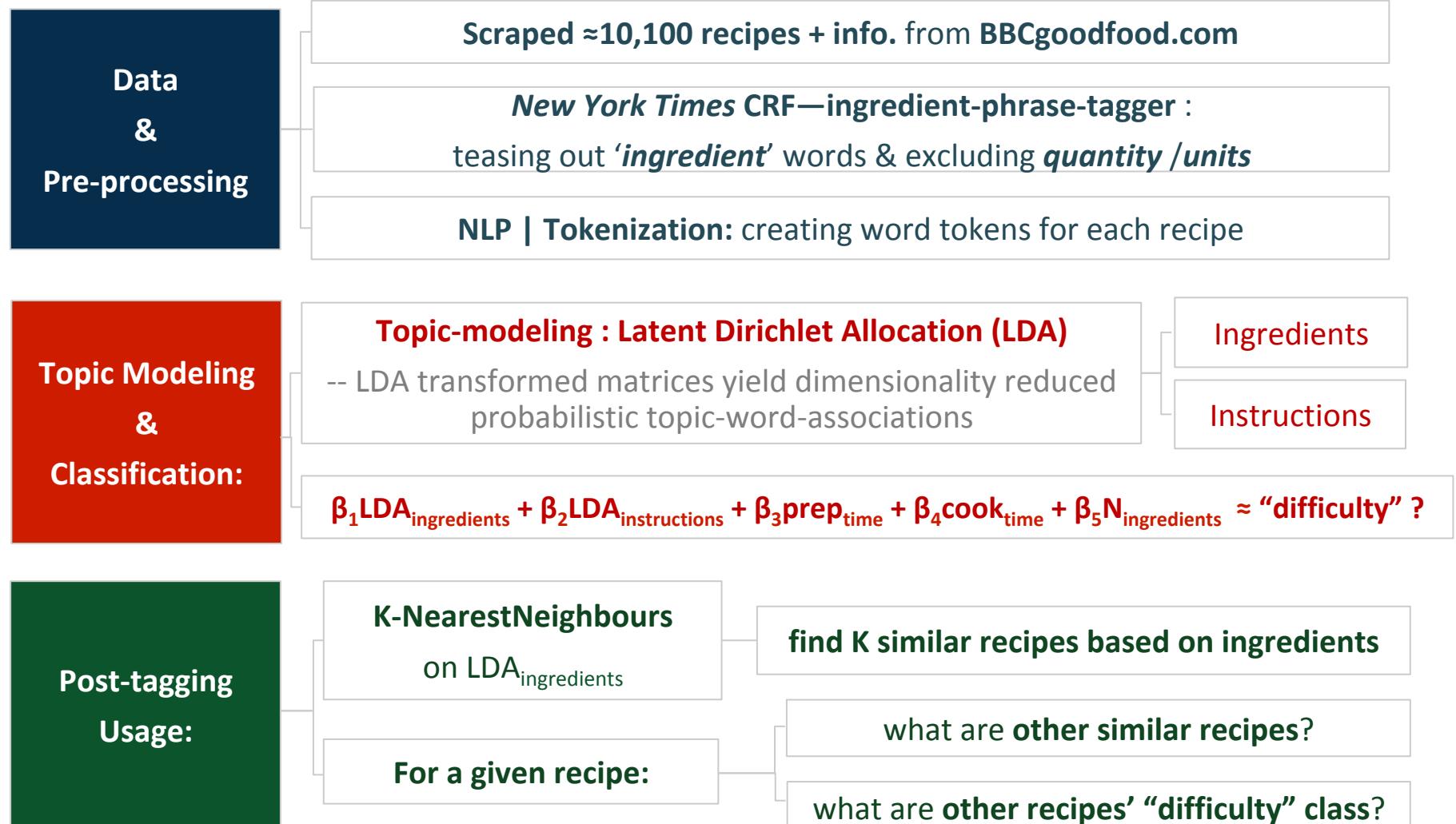


Not all recipes are created equal

- How can we simplify planning & cooking?
- What aspects of a recipe make it challenging?
- **Can we classify recipes by their “difficulty”?**

**very few sites/books include this info.

Creating & Using a recipe-difficulty-tagger



NYT-ingredient phrase tagger

- The model predicts *sequences of labels* for *sequences of words*.
- ~ Parts-Of-Speech-tagging | Named-Entity-tagging
- 89% sentence-level accuracy when trained on 130,000 labeled ingredient phrases.

Example

Ingredient Phrase	1	tablespoon	fresh	lemon	juice
Ingredient Labels	QUANTITY	UNIT	COMMENT	NAME	NAME

predict NAME, UNIT, QUANTITY, COMMENT and OTHER.

Examples

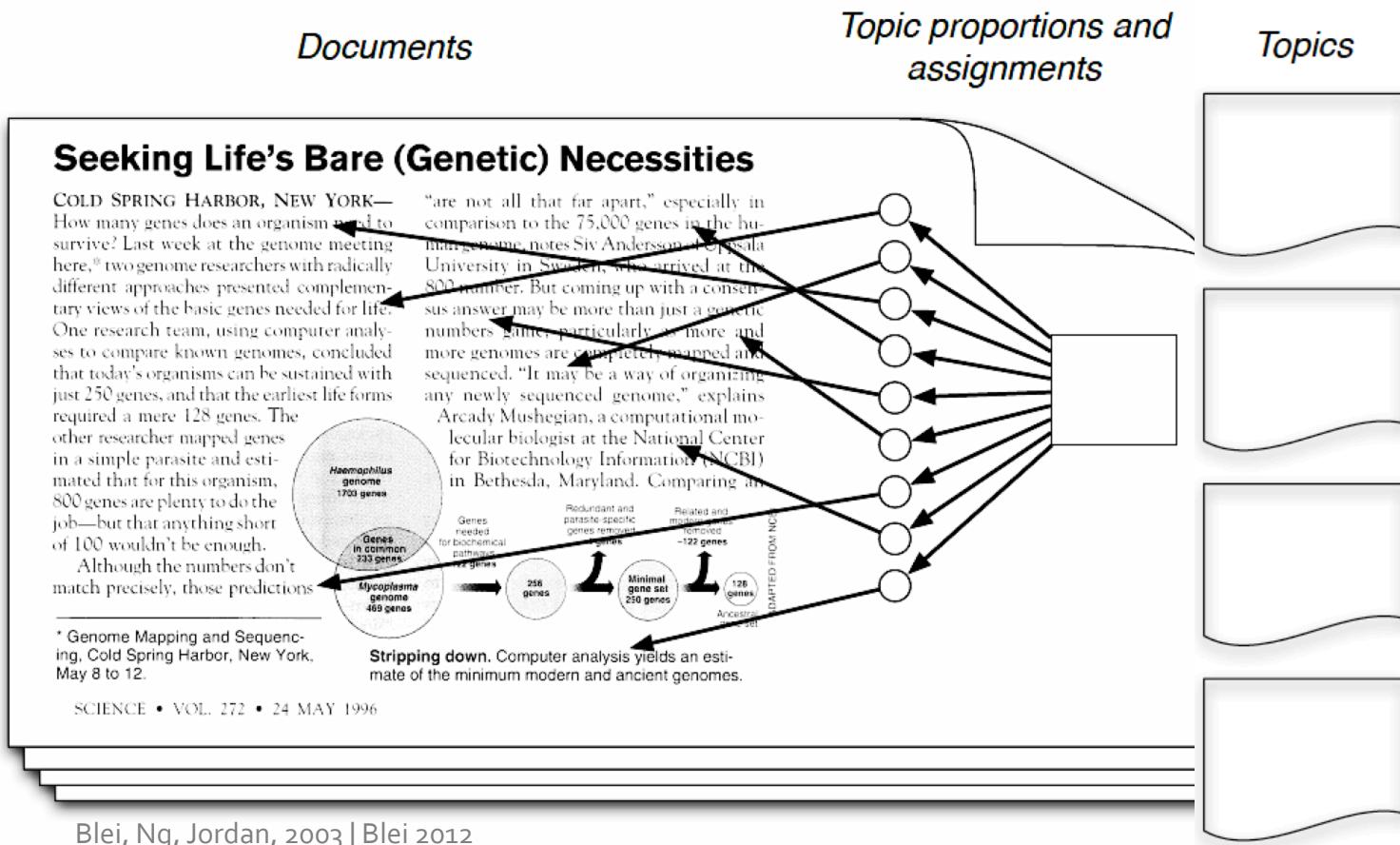
INPUT	NAME	QUANTITY	UNIT	COMMENT
1 6-inch white-corn tortilla	white-corn tortilla	1.0		6-inch
3 cups seedless grapes, equal amounts of red and green grapes	grapes	3.0	cup	seedless, equal amounts of red and green
1/4 cup good quality olive oil	good quality olive oil	0.25	cup	
3 large cloves garlic, smashed	garlic	3.0	clove	smashed
Rind from 1/2 pound salt pork	salt pork	0.5	pound	Rind from

Tagged ingredient phrases

bbcggfd_469o.txt

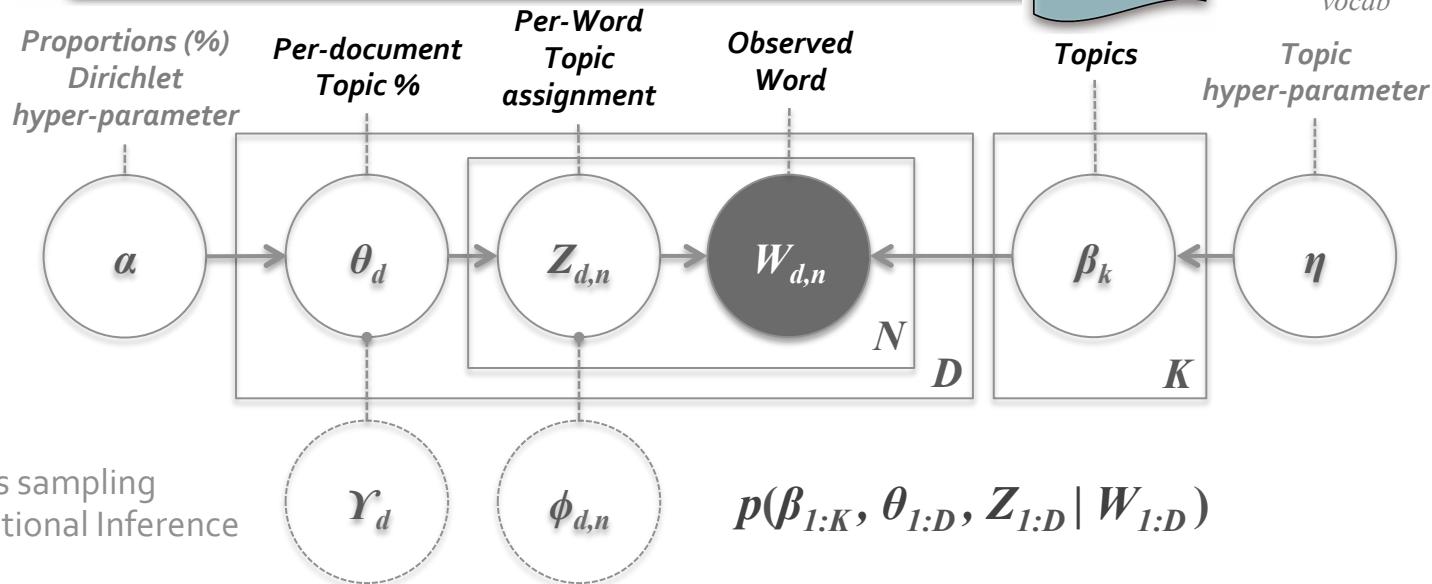
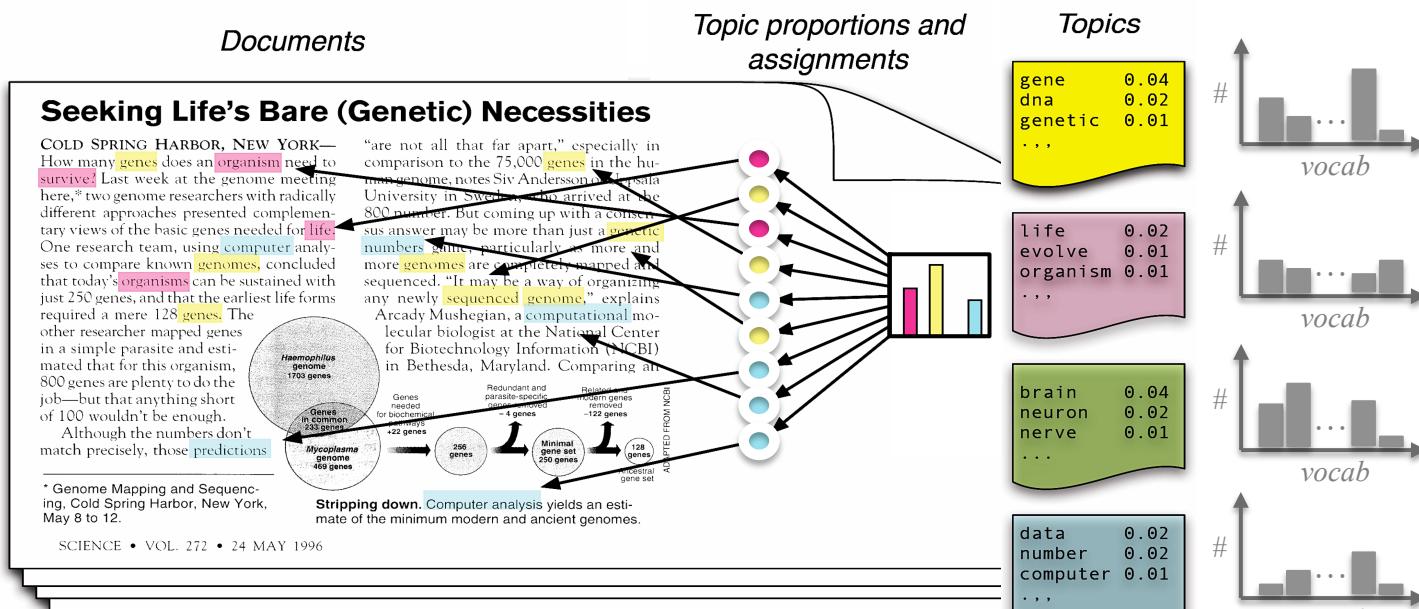
# 0.774212	250gram	I1		L4	NoCAP	NoPAREN	B-QTY/0.826164	
	dark-chocolate	I2		L4	NoCAP	NoPAREN	B-NAME/0.871789	
250gram dark-chocolate								
2 tablespoon golden syrup			# 0.841936					
568milliliter carton double cream	2	I1		L8	NoCAP	NoPAREN	B-QTY/0.989355	
4 teaspoon instant coffee granules	tablespoon	I2		L8	NoCAP	NoPAREN	B-UNIT/0.979152	
1 teaspoon ground cinnamon	golden	I3		L8	NoCAP	NoPAREN	B-NAME/0.867876	
cocoa powder, for dusting	syrup	I4		L8	NoCAP	NoPAREN	I-NAME/0.879950	
			# 0.238208					
	568milliliter	I1		L8	NoCAP	NoPAREN	B-QTY/0.753478	
	carton	I2		L8	NoCAP	NoPAREN	B-UNIT/0.301439	
	double	I3		L8	NoCAP	NoPAREN	B-NAME/0.549777	
	cream	I4		L8	NoCAP	NoPAREN	I-NAME/0.873516	
input	name	other	qty	unit	qty1	unit1	name1	comment
0 250gram dark-chocolate	dark-chocolate	NaN	250gram	NaN	250	gram	dark-chocolate	NaN
1 2 tablespoon golden syrup	golden syrup	NaN	2	tablespoon	2	tablespoon	golden syrup	NaN
2 568milliliter carton double cream	double cream	NaN	568milliliter	carton	568	milliliter	double cream	NaN
3 4 teaspoon instant coffee granules	instant coffee granules	NaN	4	teaspoon	4	teaspoon	instant coffee granules	NaN
4 1 teaspoon ground cinnamon	cinnamon	NaN	1	teaspoon	1	teaspoon	cinnamon	ground
5 cocoa powder, for dusting	cocoa powder	,	NaN	NaN	NaN	NaN	cocoa powder	for dusting
	for	I4		L8	NoCAP	NoPAREN	B-COMMENT/0.857110	
	dusting	I5		L8	NoCAP	NoPAREN	I-COMMENT/0.965766	

LDA topic-modeling



Infer the hidden topic structure from observed documents
→ Reversing the generative process...

An intuition - text generation

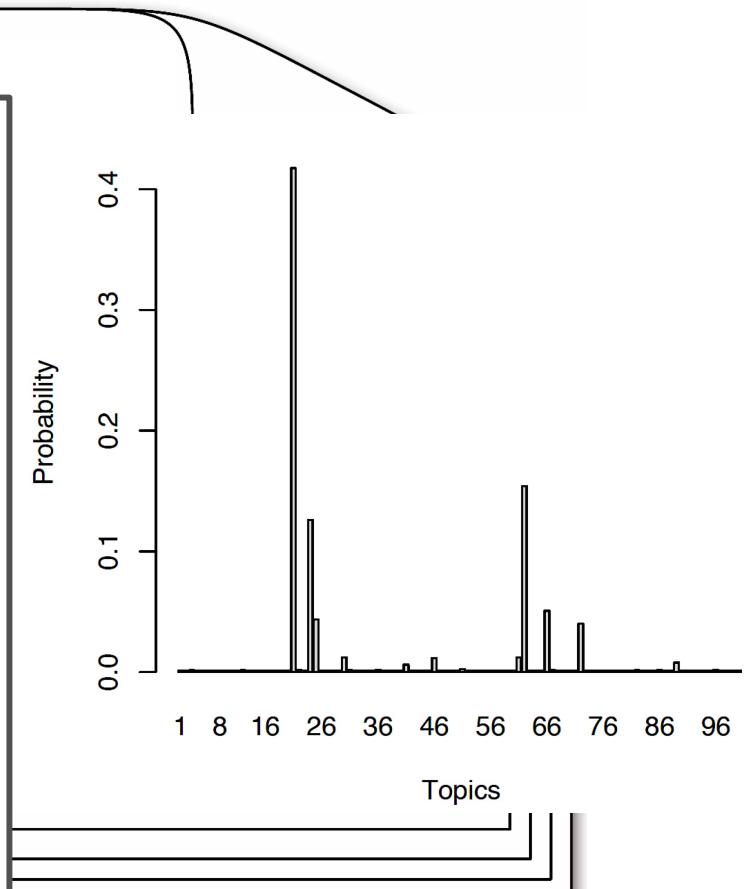


- Gibbs sampling
- Variational Inference

LDA topic-modeling

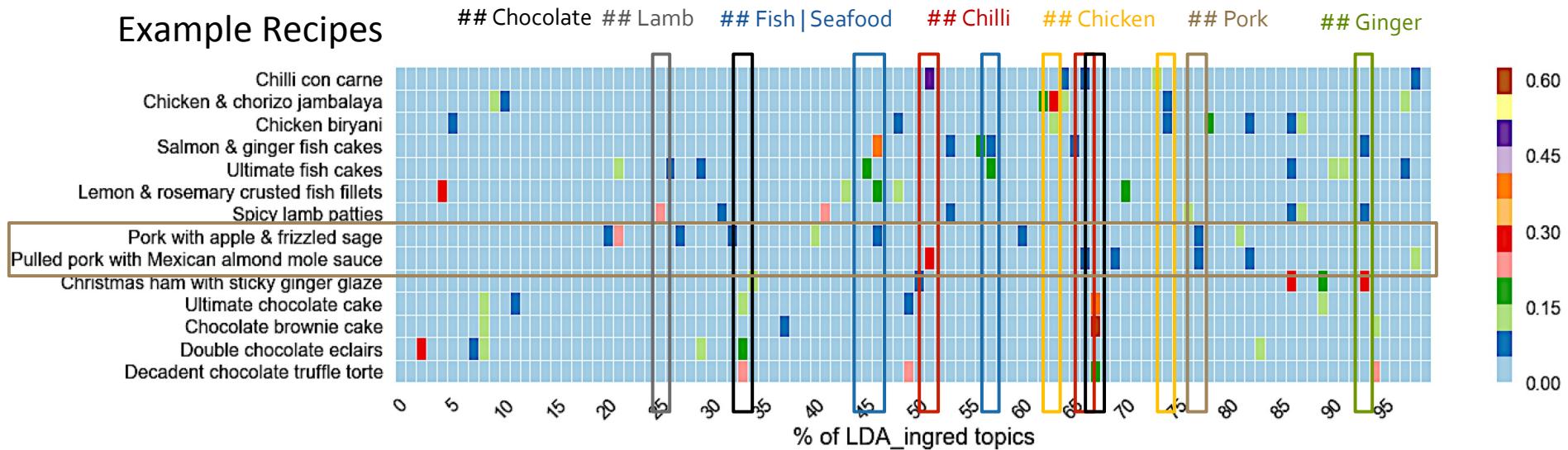
Seeking Life's Bare (Genetic) Necessities

“Genetics”	“Evolution”	“Disease”	“Computers”
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations



Ingredients topic-modeling

Example Recipes



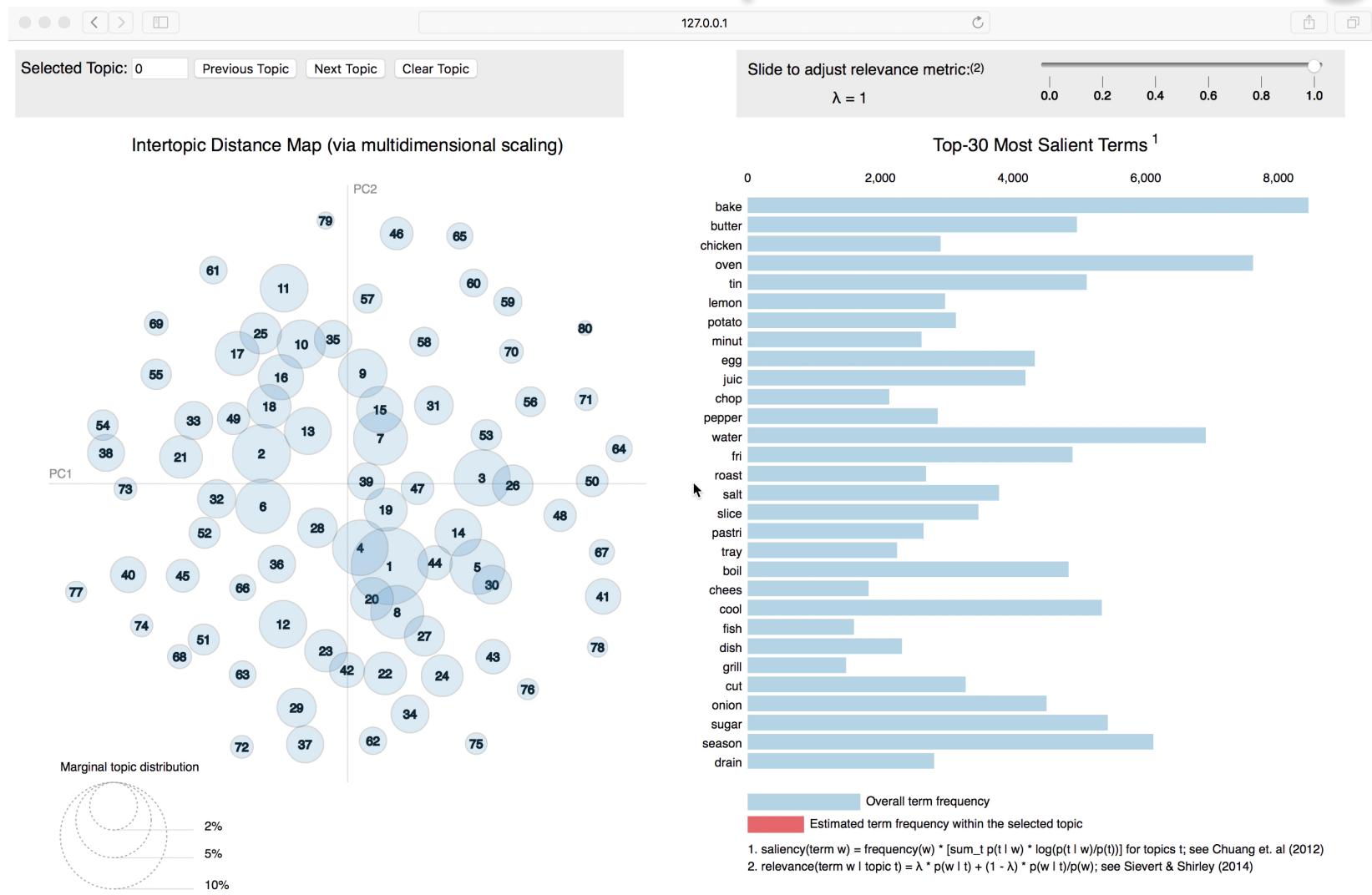
Chocolate-y

Topic 67: sugar flour powder egg butter self-rais bake caster milk cocoa extract plain vanilla dark-chocol ice
Topic 33: cream doubl butter dark-chocol milk sugar whip serv egg coffe liqueur liquid glucos squar disaronno

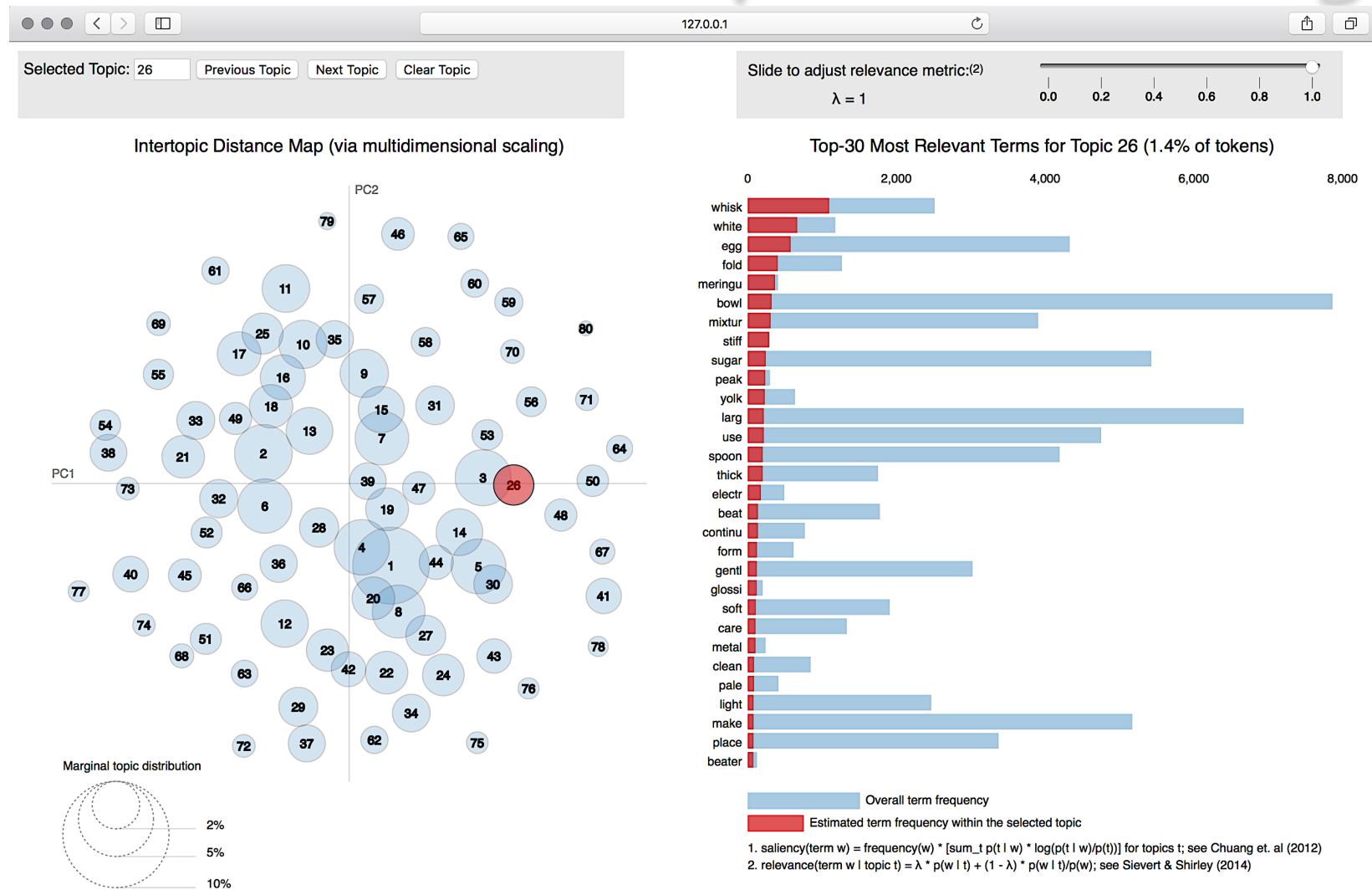
Chicken

Topic 63: chicken stock thigh onion garlic butter duck liver drumstick sunflow chorizo cornichon
Topic 74: breast chicken boneless skinless oliv blue onion sunflow strip zest flour free-rang

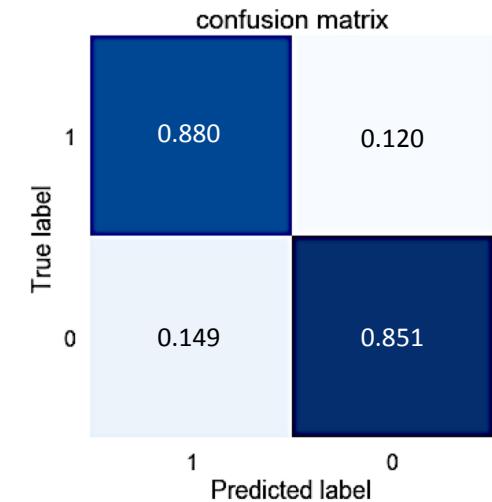
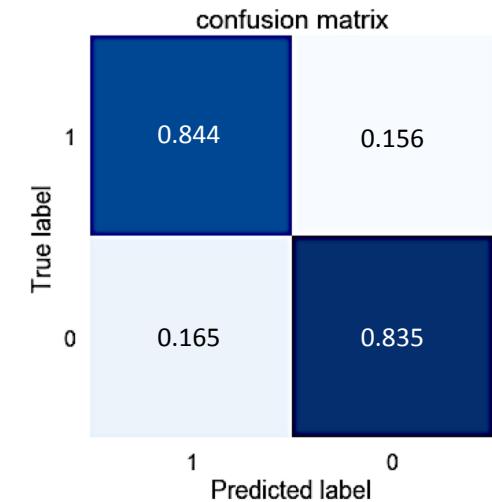
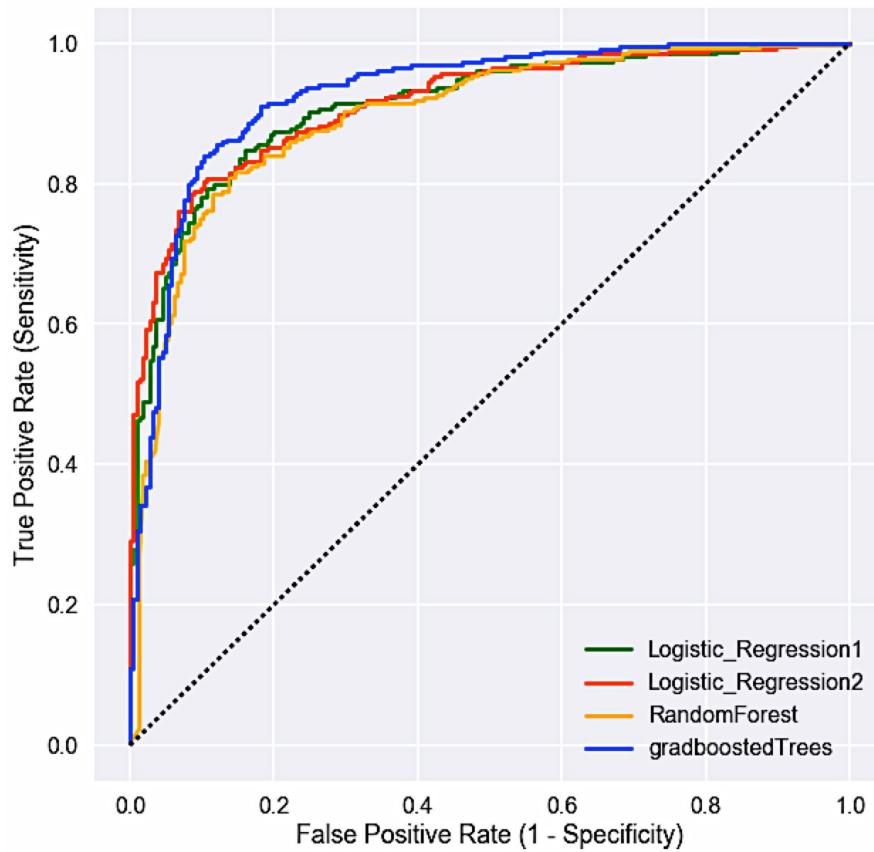
Instructions topic-modeling



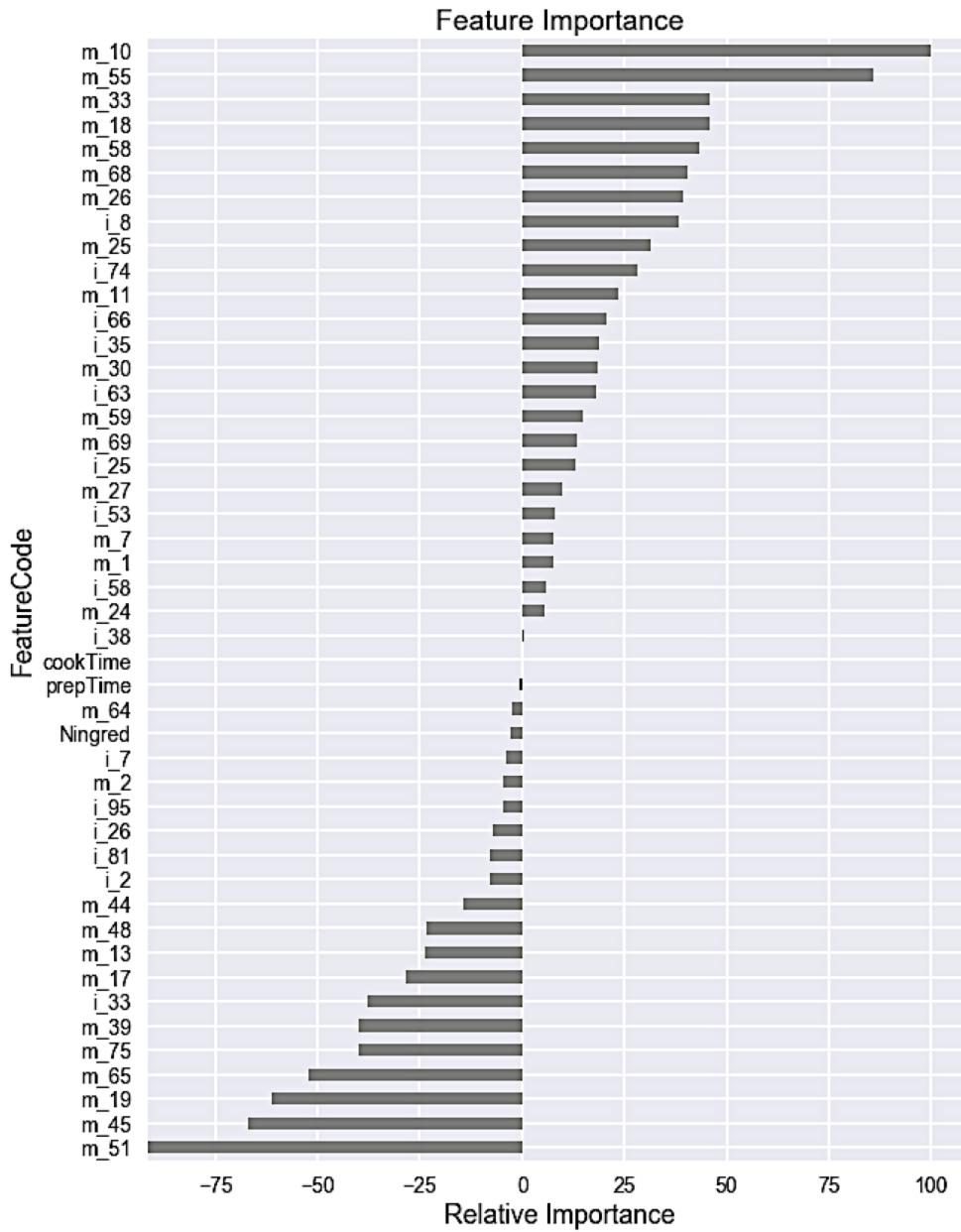
Instructions topic-modeling



Classification outcomes on hold-out data:



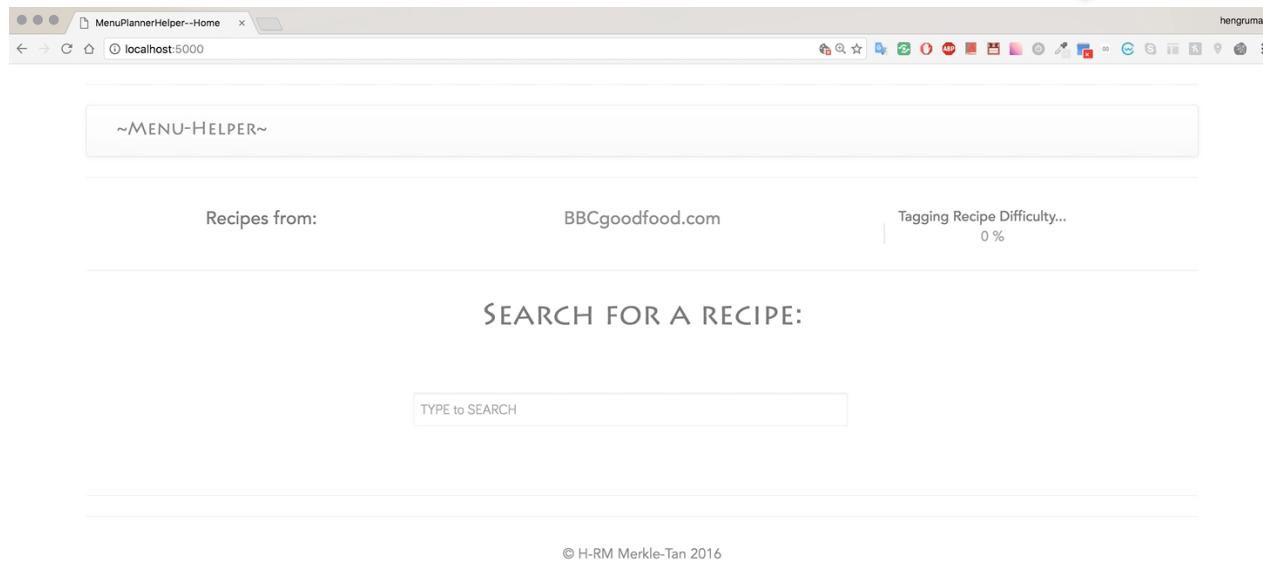
What makes a recipe ~difficult?



- # MTopic 10: soup/stew
- # MTopic 55: baked oven dish
- # MTopic 33: pasta
- # MTopic 18: curry
- # MTopic 26: casserole dishes
- # ITopic 8: general baking
- # ITopic 74: chicken-related
- # ITopic 66: fingerfood / stew
- # ITopic 35: basil/Italian -- salad/finger-food

- # ITopic 33: chocolate/ganache
- # MTopic 39: pastry/baking/case
- # MTopic 75: parcel-type method
- # MTopic 65: deep frying
- # MTopic 19: Roast meat/duck
- # MTopic 45: custard
- # **MTopic 51: meringue**

Menu Planner Helper



AWS: <http://bit.ly/menuplannerhelper>

#Interim Summary

- Recipe Difficulty Tagger: a way to organize/categorize documents
 - Train CRF model for ingredient phrases on BBCgoodfood.com/additional recipes
 - LDA topic modeling disregards word-order: maybe useful for recipe instructions
→ LDA2Vec ~ = LDA + Word2Vec
- Usage:
 - Tag recipes from other sites and test recipe similarity suggestions
 - Collect users' feedback to assess and update model
 - Personalized 'learning' tool: tailoring your cooking adventure – where the model could be adapted with user's experience
- Apply methodology to other types of documents

With thanks to:



Natural Language
Analyses with NLTK



gensim
topic modelling for humans



CRF++: Yet Another CRF toolkit

BBC
GoodFood

RDB
pyLDAvis

B
Bootstrap



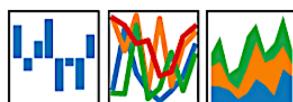
scikits
learn

machine learning in Python



python™

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



 **Flask**
web development,
one drop at a time

THANK YOU

MAY MERKLE-TAN, PHD

HENGRUMAY@GMAIL.COM

GITHUB.COM/HENGRUMAY

WWW.LINKEDIN.COM/IN/HENGRUMAY