David Tsukiyama

CSE 190 Dahta Mining and Predictive Analytics

Professor Julian McAuley

Amazon Fine Food Reviews…wait I don't know what they are reviewing

Dataset

This paper uses Amazon Fine Food reviews from Stanford University's Snap Datasets, https://snap.stanford.edu/data/web-FineFoods.html. The Fine Foods datasets consists of 568,454 reviews between October 1999 and October 2012; 256,059 users and 74,258 products.

The data format is as follows:

```
product/productId: B001E4KFG0

review/userId: A3SGXH7AUHU8GW

review/profileName: delmartian

review/helpfulness: 1/1

review/score: 5.0

review/time: 1303862400

review/summary: Good Quality Dog Food

review/text: I have bought several of the
Vitality canned dog food products and
havefound them all to be of good quality.
The product looks more like a stew than a
processed meat and it smells better. My
Labrador is finicky and she appreciates this
product better than most.
```

The distribution of review/scores is skewed towards scores of 4 and 5:

Figure 1: Distribution of Scores

I counted the frequency of reviews by reviewer id, the histogram has 100 bins to extract some granularity:

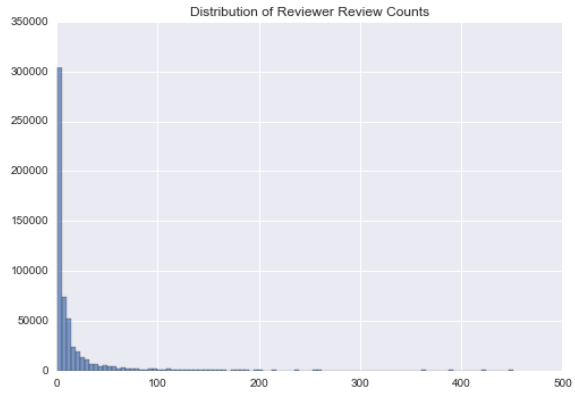

Figure 2: Distribution of Reviewers

The temporal dimensions of the data are important to fully understand user behavior. The review/helpfulness variable was deconstructed into two components, the number of actual helpful ratings received and the number out of. The following shows helpfulness votes over time.

In "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews," McAuley and Leskovec demonstrate

that recommendations engines should take into consumer experiences in addition to their tastes [1]. Therefore I take a look at the temporal dimensions of user scores and helpfulness between different levels of users. I divided users into 5 categories:

1. Light: less the 10 reviews
2. Medium: Greater than 10, but less than or equal to 50 reviews
3. High: Greater than 50, but less than or equal to 75 reviews
4. Very heavy: Greater than 75, but less or equal to 100 reviews
5. Expert: More than 100 reviews

I confess that these breakpoints may be arbitrary. The actual distribution of review frequencies is as follows:

count	568462.000000
mean	18.124276
std	41.320272
min	1.000000
25%	1.000000
50%	5.000000
75%	14.000000
max	451.000000

The breakpoints I selected more or less bin medium, high very high, and expert users into similar sized bins which gives me a sufficient amount of observations per user type to run predictive tasks on.

Number of observations per user type:

Light: 390758

Medium: 128326

High: 16270

Very Heavy: 8763

Expert: 24345

Running moving average (MA) regressions on scores and helpfulness over all users gives us the following two plots. Both demonstrate a downward trend over time.

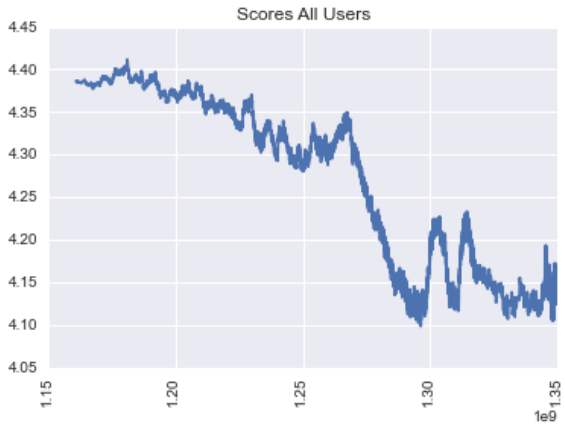

Figure 3: Scores for all Users

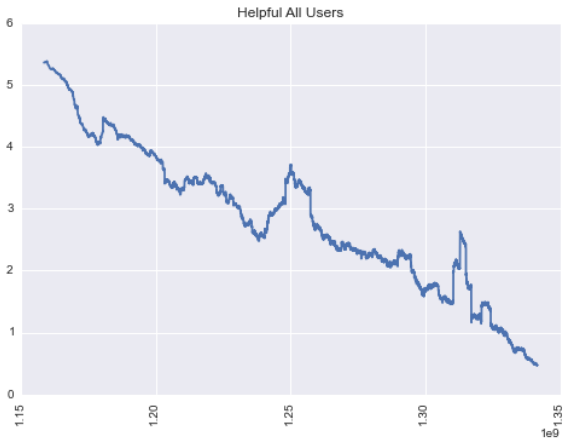

Figure 4: Helpfulness for all Users

Moving average time series regression were run for all user types for scores.

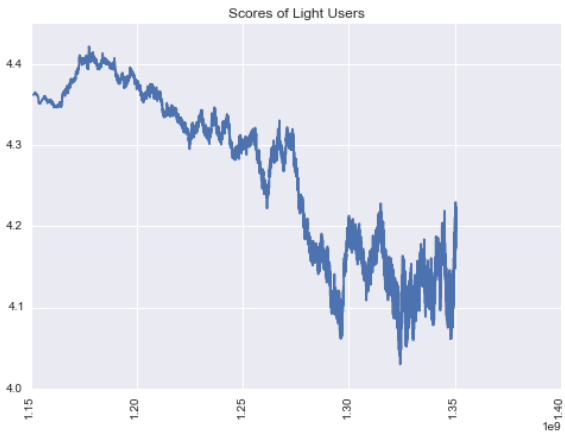

Figure 5: Scores of light Users

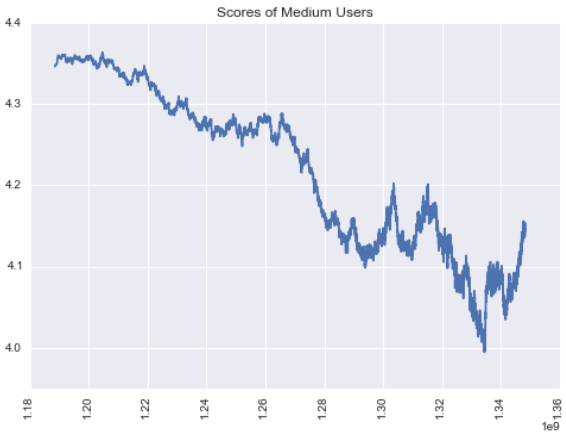

Figure 6: Scores of medium Users

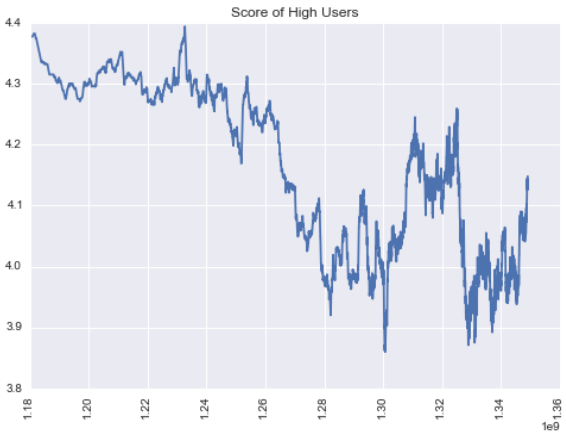

Figure 7: Scores of high Users

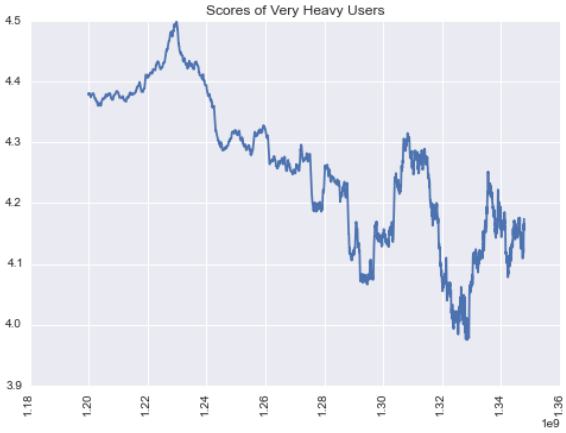

Figure 8: Scores of very heavy Users

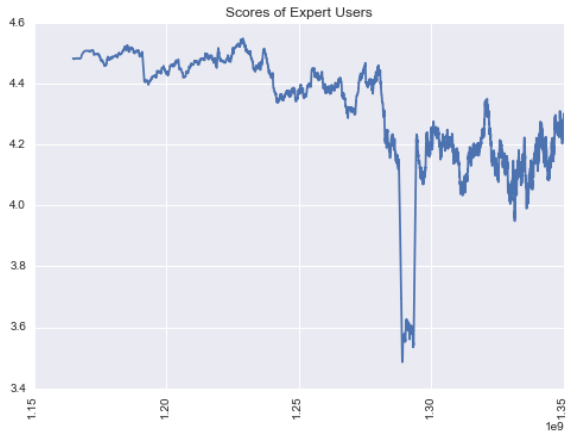

Figure 9: Scores of expert Users

Review scoring behavior varies among user type.

Predictive Task

I had my mind set on finding a compelling predictive task, digging through the data I noticed that there were no product names or descriptions in the dataset that were easily accessible. Review summaries sometimes mentioned the product under review, otherwise there is no category label that provides a way to simply group products and user preferences.

The predictive task at hand is to represent text reviews in the terms of the topics they describe, i.e. topic modeling.

The technique used to extract topics from Amazon fine food reviews is Latent Dirichlet Analysis (LDA). We assume that there is some number of topics (chosen manually), each topic has an associated probability distribution over words and each document has its own probability distribution over topics; which looks like the following [2]:

$$p(d|\theta, \emptyset, Z) = \prod_{j=1} length\ of d\ \ \theta_{z_{d,j}} \emptyset_{z_{d,j}, w_{d,j}}$$

Gibbs Sampling is used to extract the aforementioned distributions. Where we only know some of the conditional distributions, Gibbs Sampling takes some initial values of parameters and iteratively replaces those values conditioned on its neighbors.

Every word in all the text reviews is assigned a topic at random, iterating through *each* word, weights are constructed for each topic that depend on the current distribution of words and topics in the document, and we iterate through this entire process until "we get bored." [3]

Literature Review

The literature on LDA is significantly more sophisticated than this paper's goal of finding whether a reviewer reviewed dog food or not. Perhaps the seminal paper on LDA is "Latent Dirichlet Allocation" by David M. Blei, Andrew Y. Ng, and Michael I. Jordan. The authors use LDA to model topics from Associated Press newswire articles [4].

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Model and Results

The dataset was split into training and test datasets, random 50-50 split.

In order to utilize Latent Dirichlet Allocation for topic modeling feature vectors needed to be created from the text reviews. All the reviews were converted to a bag of words and stop words removed.

Quantitatively evaluating the model is done through a 'perplexity' score, which measures the log-likelihood of the held-out test set [5]

$$perplexity(text\ set\ w) = \exp\{-\frac{L(w)}{word\ count}\}$$

Lower perplexities are better. However model-fit in regards to topic-modeling does not seem to be an intuitive way to measure whether the topics chosen are 'accurate' from a human perspective. Indeed in 'Reading Teat Leaves: How Humans Interpret Topic Models,' Sean Gerrish, Chong Wang, and David Blei find that traditional metrics do not capture whether topics are coherent, human measures of interpretability are negatively correlated with traditional metrics to measure the fit of topic models [6].

This assumption will be tested when labels are assigned to the topics created with model. The perplexity metric is used to choose the final model that will be interpreted.

Topics	Perplexity
10	2439.96
15	2459.95
20	2478
25	2487
30	2434.61
50	2573.42

The ultimate model chosen for this task models 30 topics. 20 words with highest probability are shown (training set topics).

0	1	2	3	4	5	6	7	8	9
tea	food	br	love	br	product	dog	good	br	good
green	cat	chips	butter	taste	amazon	dogs	free	sugar	flavor
drink	cats	eat	find	cheese	pack	treats	flavor	coconut	bag
good	chicken	ingredients	peanut	easy	www	loves	love	oil	don
teas	dry	healthy	made	buy	http	treat	gluten	drink	texture
milk	eat	snack	chocolate	love	bags	teeth	sweet	sweet	bit
water	diet	cereal	delicious	lot	gp	pet	snack	calories	potato
tastes	feed	corn	cream	find	href	chew	tasting	make	hard
black	baby	rice	pretty	2	find	giving	enjoy	honey	tasty
leaves	canned	almonds	wonderful	favorite	don	puppy	bought	powder	favorite
iced	eating	3	perfect	tasty	great	training	fresh	ve	snack
chai	grain	size	eat	noodles	3	toy	natural	hot	buy
strong	feeding	blue	making	doesn	excellent	formula	add	1	package
drinking	meat	fiber	bag	texture	4	pill	granola	don	time
makes	wellness	foods	red	tasted	5	hard	tastes	bottle	healthy
buy	vet	bar	absolutely	people	ounce	year	don	artificial	bar
stash	problems	daily	kind	son	found	chewing	highly	juice	strong
loose	wet	raw	tasted	version	chips	long	licorice	stevia	seeds
powder	issues	feel	amazing	flavor	boxes	ball	favorite	flour	pretty
delicious	happy	doesn	mixed	crackers	24	salmon	prefer	thing	snacks

10	11	12	13	14	15	16	17	18	19
water	high	mix	br	coffee	product	ve	price	br	cookies
taste	store	good	1	flavor	time	br	cup	sugar	bars
sauce	quality	stuff	2	taste	products	make	buying	hair	candy
add	highly	didn	organic	blend	years	store	stores	fat	eat
ve	protein	bread	4	drink	bit	made	back	3	perfect
bottle	2	flavors	ingredients	vanilla	thought	brand	bought	product	eating
sugar	5	oatmeal	sodium	favorite	life	love	shipping	day	good
sweet	ll	work	oz	starbucks	company	people	buy	doesn	soft
added	weight	arrived	5	roast	3	pasta	grocery	didn	family
make	ingredients	brand	fat	bitter	brand	half	years	problem	regular
nice	worth	flavor	8	tastes	price	ll	ve	give	nice
minutes	local	quality	milk	beans	money	put	fine	calories	find
flavor	drinks	white	soy	decaf	tasted	lot	cost	blood	mouth
chicken	recommend	ll	protein	full	recipe	grocery	cheaper	low	hard
heat	times	box	6	espresso	boxes	rice	ordered	bar	fresh
lot	work	brown	vitamin	nice	package	top	mountain	protein	wheat
adding	happy	don	0	found	cookie	light	morning	stuff	pieces
natural	recommended	made	acid	aftertaste	mix	delicious	expensive	clear	products
drinking	hour	worth	12	caffeine	ordered	tasting	medium	thought	tuna
cup	found	awesome	taste	french	waste	strong	long	isn	ginger

20	21	22	23	24	25	26	27	28	29
bought	great	love	salt	good	amazon	coffee	great	br	don
ve	taste	buy	fruit	bad	order	cups	taste	chocolate	hot
small	found	amazon	flavors	box	store	box	cup	taste	price
popcorn	organic	order	kids	smell	2	love	give	dark	bit
make	product	purchased	taste	day	received	keurig	stars	milk	recommend
bag	stuff	product	high	long	oil	flavor	make	cocoa	day
work	foods	recommend	time	cans	day	hot	enjoy	time	big
found	months	tea	br	strong	ago	morning	makes	sweet	gave
buy	reviews	buying	thing	product	didn	machine	real	cinnamon	husband
item	picky	local	soda	give	days	flavored	5	nice	home
size	deal	find	juice	wasn	service	brew	house	recommended	beans
time	pop	ordered	product	big	time	single	nuts	find	bag
makes	fact	soup	cake	ordered	local	bold	ll	body	minutes
low	put	fresh	makes	energy	quickly	rich	thought	magnesium	easy
gum	brands	black	loved	review	olive	wonderful	won	day	spicy
reviews	large	shipping	gift	smells	small	coffees	back	happy	years
perfect	arrived	item	family	expected	great	pod	feel	buy	year
stick	3	make	fresh	recommend	pay	hazelnut	surprised	package	real
read	husband	spice	excellent	natural	put	bitter	things	thing	disappointed
week	free	save	cherry	star	stores	bag	pretty	make	run

Some of these are easy to categorize, topic 6 looks pet related, topic 19, candy. Some are vague, topic 29 which contains words such as: “husband, beans, easy, spicy, and disappointed.” Manually labeling these topics seems fraught with difficulty. However to test the predictions of the model several categories relatively easy topics are labeled.

Topic 6: Dog Treats

Topic: 14 Coffee

Topic 26: Coffee Condiments

The model was used to predict topics for the test dataset. Topic frequency is plotted below.

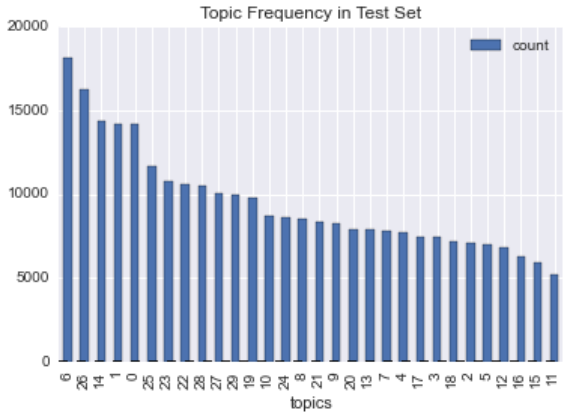

Figure 10: Topic Frequency

Topic 6 is the most frequent, and as mentioned, probably pet related, i.e. dog treats. We can test the accuracy of the topic model on whether these reviews are really about dog treats. Topic 6 (dog treats) has 18,401 entries. ‘Dog’ comes up in 13,128 of those reviews: 71.3%. ‘Dog’ or ‘treat’ comes up in 15,369 of those reviews: 83.5%.

Topics 26 and 14 seem to both deal with coffee, 26 perhaps coffee related goods and 14 actual coffee beans. Topic 26 has 16,284 observations, 11,454 mention ‘coffee’: 70.3%. Topic 14 has 14,326 observations, 10,283 mention ‘coffee’: 71.7%. However differentiating between the two categories is difficult. Manually scouring the reviews under the two topics gives the impression that there is some ephemeral difference, the products in topic 26 perhaps are more likely to be single serve, while topic 14 are beans.

Now that categories are assigned we can track user behavior over time.

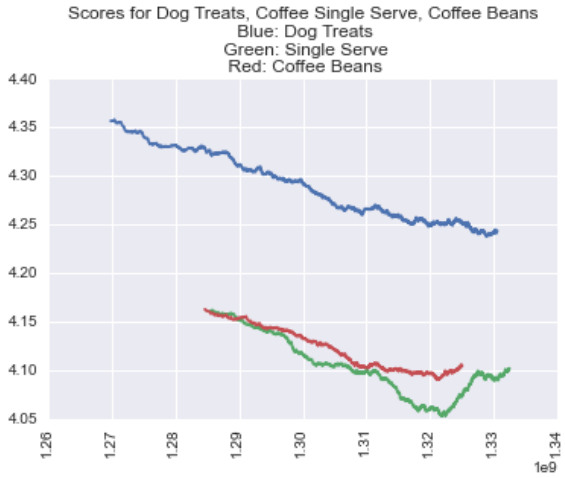

Figure 11: Scores of Test Set Reviewers

Topic frequency between light users, those with 10 or fewer reviews and experts, those with 100 or more reviews.

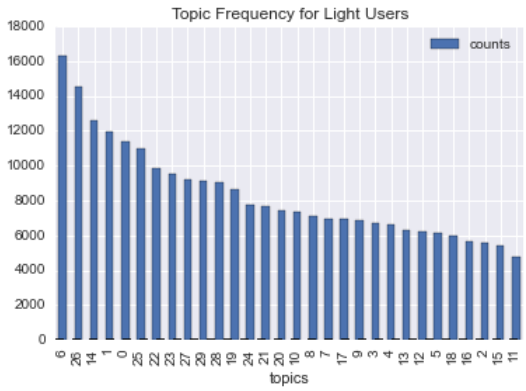

Figure 12: Topic Frequency for Light Users

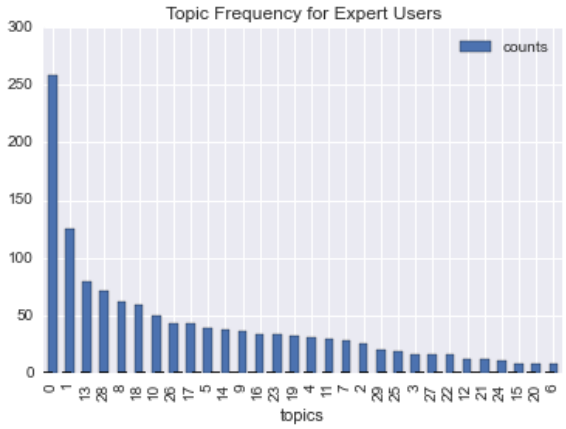

Figure 13: Topic Frequency for Expert Users

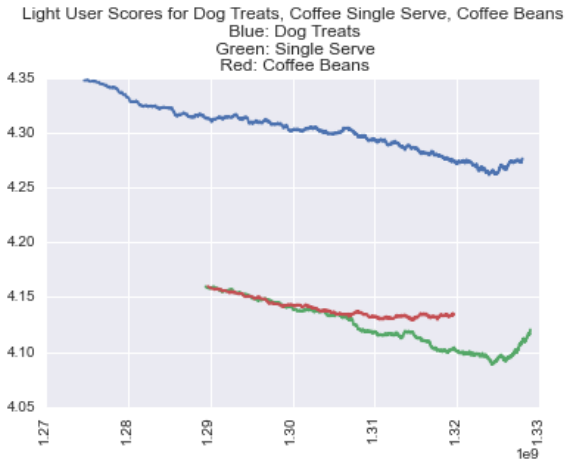

Figure 14: Scores for selected topics for Light Users

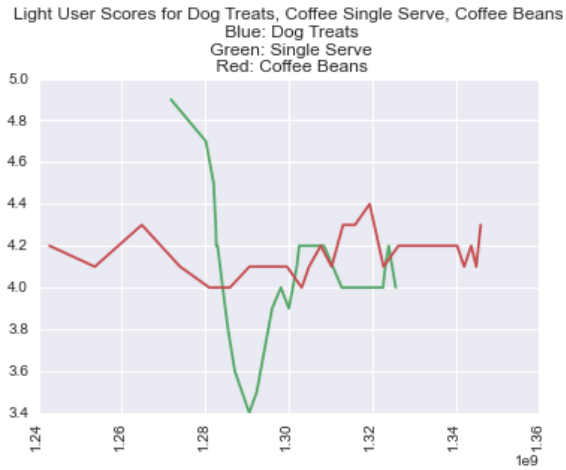

Figure 15: Scores for selected topics for Expert Users

Conclusions

In this paper we sought to create food categories from text reviews with Latent Dirichlet Allocation topic modeling. We observed that LDA is a powerful method to represent documents in terms of the topics they represent and is effective at summarizing a large collection of documents. However, model fit metrics are not easily understood intuitively in regards to human coherence of the resulting representation of documents. Actual real world implementation in model results seems more difficult than other unsupervised machine learning algorithms.

References

[1] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. CoRR, abs/1303.4402, 2013.

[2] Julian McAuley. CSE 190 Dahta Mining and Predictive Analytics. Lecture 13, slide 63, Spring 2015, UCSD.

[3] Julian McAuley. CSE 190 Dahta Mining and Predictive Analytics. Lecture 13, slide 67, Spring 2015, UCSD.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022.

[5] Quintin Pleple. Perplexity To Evaluate Topic Models. http://qpleple.com/perplexity-to-evaluate-topic-models/

[6] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David Blei. Reading Tea Leaves: How Human Interpret Topic Models. Neural Information Processing Systems, 2009.