Improving Text Analysis Using Sentence Conjunctions and Punctuation

Joachim Büschken / Catholic U. of Ingolstadt

Greg Allenby / Ohio State University

Challenges in understanding consumers

- Ambiguous language
- Ratings scales
  - Yeah and nay-saying
  - Haloing
  - Pre-define statements
- Choice-based analysis
  - Conjoint: incentive alignment
  - Revealed demand: short purchase histories

Customer review data

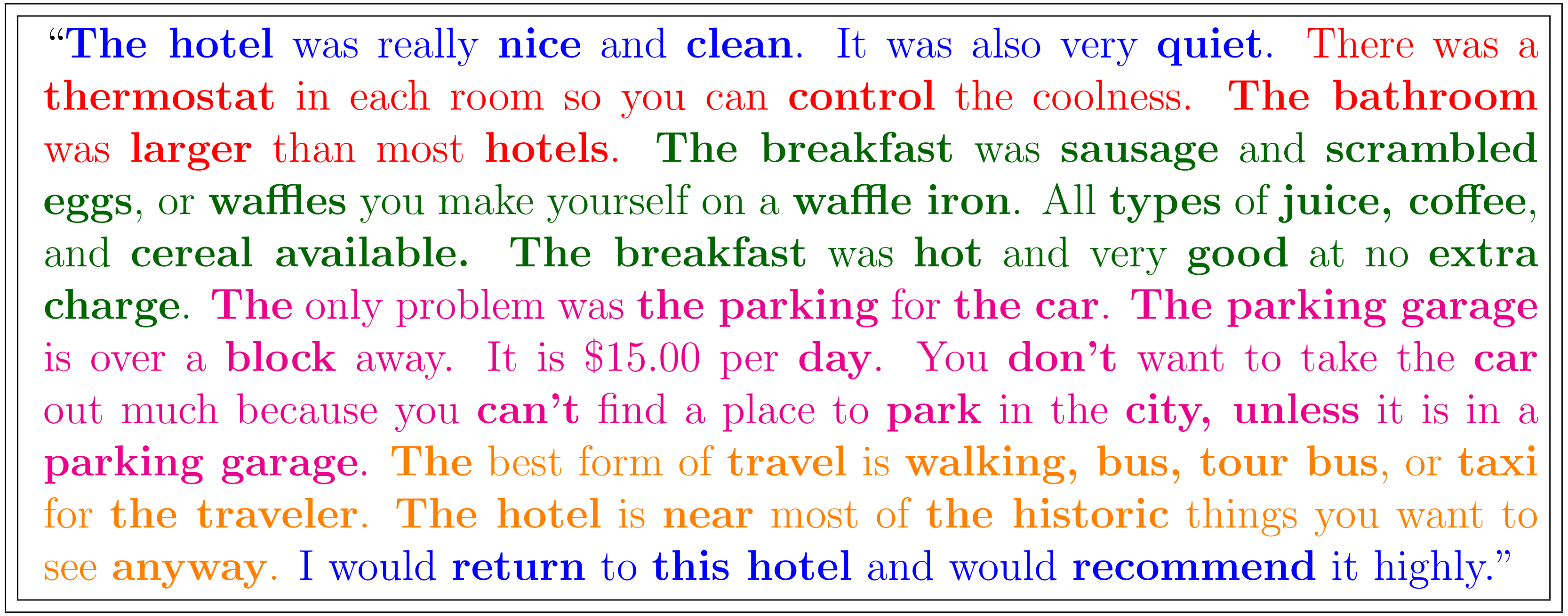
- Unstructured
- Use of dummy-variable regression
  - Valence
  - Appearance of words
  - Applied to intent, experience, conversion, attributes, sales
- Difficulty with high-level interactions
  - 1,000 unique terms are usual
  - Variable selection

Topic modeling

- Topics are comprised of word vocabularies (words and probabilities).
- A review or articles is associated with many potential topics.
- Words are thought to appear after choosing a topic, and then a word from the topic vocabulary.
- Word and topic probabilities are estimated from the observed word counts across documents.
- Topics can be related to other data, such as customer ratings on a fixed-point scale.

Büschken and Allenby (2016)

- Sentence-based analysis
- Predict customer ratings of products



“**The hotel** was really **nice** and **clean**. It was also very **quiet**. There was a **thermostat** in each room so you can **control** the coolness. **The bathroom** was **larger** than most **hotels**. **The breakfast** was **sausage** and **scrambled eggs**, or **waffles** you make yourself on a **waffle iron**. All **types** of **juice, coffee**, and **cereal available. The breakfast** was **hot** and very **good** at no **extra charge**. **The** only problem was **the parking** for **the car**. **The parking garage** is over a **block** away. It is $15.00 per **day**. You **don’t** want to take the **car** out much because you **can’t** find a place to **park** in the **city, unless** it is in a **parking garage**. **The** best form of **travel** is **walking, bus, tour bus**, or **taxi** for **the traveler**. **The hotel** is **near** most of **the historic** things you want to see **anyway**. I would **return** to **this hotel** and would **recommend** it highly.”

LDA model

The n^{th} word appearing in review d, w_{dn}, is thought to be generated by the following process in the LDA model:

1. Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$.

2. Choose a word $w_{dn} \sim$ from $p(w_{dn}|z_{dn}, \Phi)$.

$$p(\theta_d) \sim \text{Dirichlet}(\alpha)$$

$$p(\phi_t) \sim \text{Dirichlet}(\beta)$$

Ratings sub-model

Cut-point model for ratings:

$$r_d = k \quad \text{if} \quad c_{k-1} \leq \tau_d \leq c_k$$

Latent regression on topic probabilities:

$$\tau_d \sim N(\theta'_d \beta, \sigma)$$

Topic probabilities for each review

Latent regression results

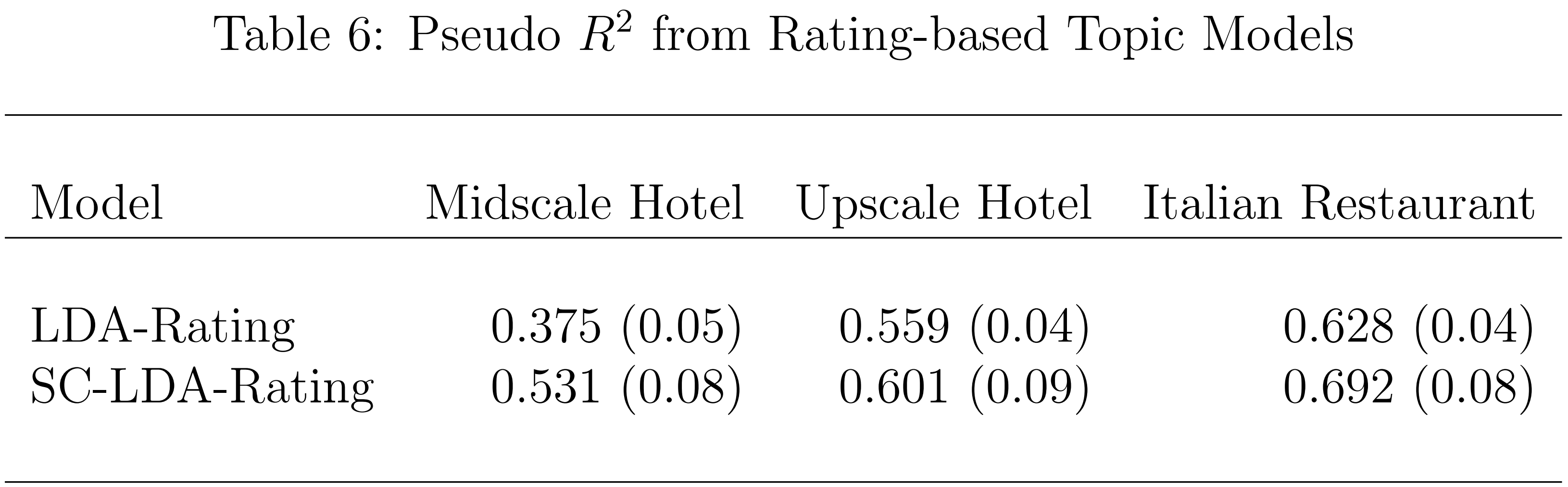


Table 6: Pseudo R^2 from Rating-based Topic Models

Model	Midscale Hotel	Upscale Hotel	Italian Restaurant
LDA-Rating	0.375 (0.05)	0.559 (0.04)	0.628 (0.04)
SC-LDA-Rating	0.531 (0.08)	0.601 (0.09)	0.692 (0.08)

What about conjunctions and other punctuation?

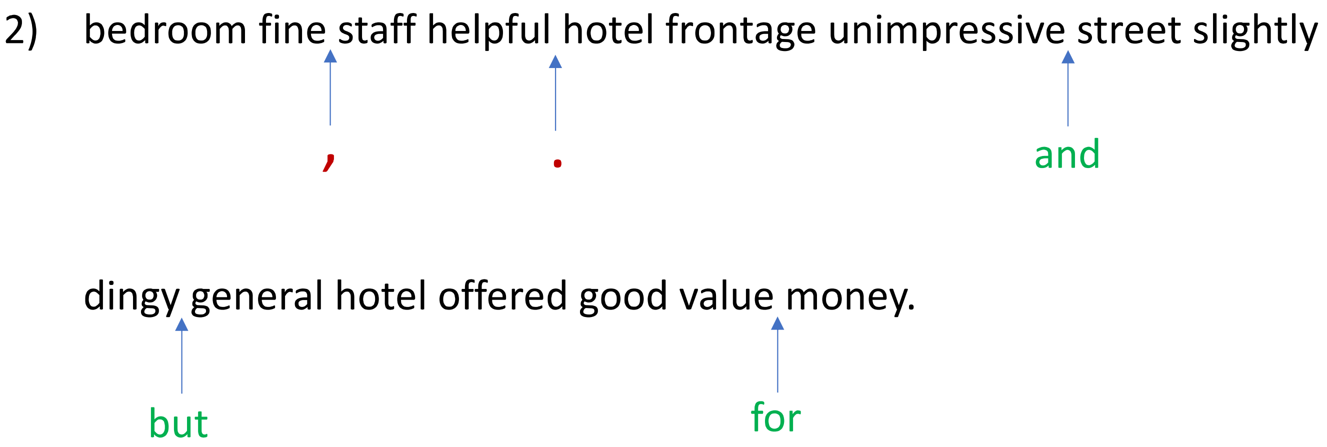
- Introduce an autocorrelated topic model where topic carryover is more generally driven by conjunctions (and, but, because) and punctuation (periods, exclamation marks, commas).
- These syntactic elements are routinely stripped out of text analysis because they are content-neutral.
- Our model assumes that a reviewer stays with a topic for some time before switching to another. The probability of topic change is a function of these elements of speech.

Illustration

Original review

Bedroom was fine, staff were helpful. Hotel frontage very unimpressive and street slightly dingy but in general hotel offered good value for the money.

1) bedroom fine staff helpful hotel frontage unimpressive street slightly dingy general hotel offered good value money.



2) bedroom fine staff helpful hotel frontage unimpressive street slightly dingy general hotel offered good value money.

Arrows indicate covariates for topic carryover

Autocorrelated Topic LDA (AT-LDA)

$$\zeta_n = 1 : z_n = z_{n-1}$$

$$\zeta_n = 0 : z_n \sim \text{Multinomial}(\theta_d)$$

Either stay with the topic or not

$$\zeta_n \sim \text{Binomial}(\psi_n | z_{n-1})$$

$$\psi_n | z_{n-1} = \frac{\exp[\delta_{0,z_{n-1}} + \tilde{x}'_n \delta]}{1 + \exp[\delta_{0,z_{n-1}} + \tilde{x}'_n \delta]}$$

Pr(stay) = f(topic constant, speech)

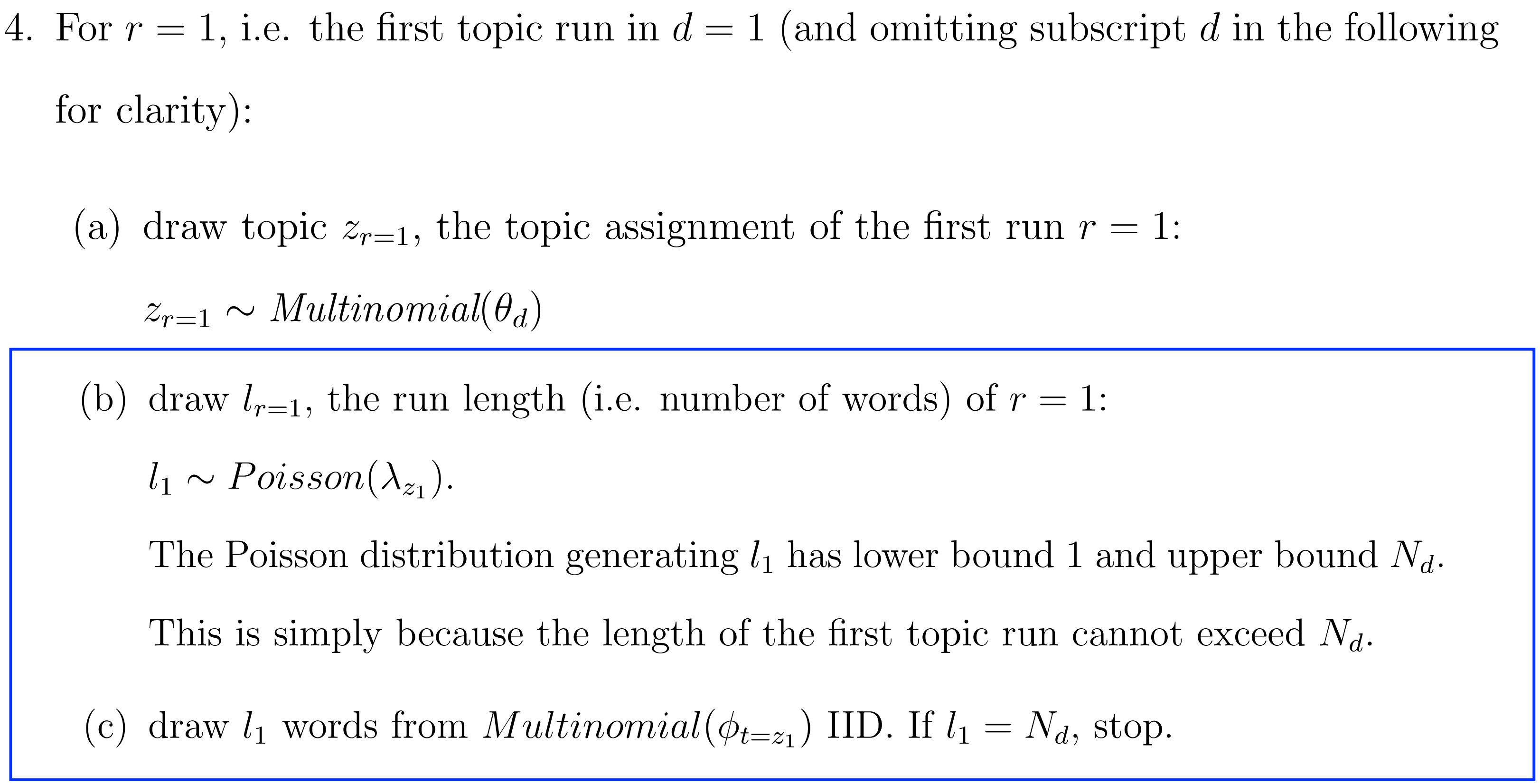
Topic changepoint model (TCPM)

1. draw $\{\theta_d\}$: $\theta_d \sim Dirichlet(\alpha)$, IID $\forall d$

2. draw $\{\phi_t\}$: $\phi_t \sim Dirichlet(\beta)$, IID $\forall t$

3. draw $\{\lambda_t\}$: $\lambda_t \sim Gamma(\delta)$, IID $\forall t$

Topics runs from a Poisson distribution



4. For $r = 1$, i.e. the first topic run in $d = 1$ (and omitting subscript d in the following for clarity):

   (a) draw topic $z_{r=1}$, the topic assignment of the first run $r = 1$:

   $z_{r=1} \sim Multinomial(\theta_d)$

   (b) draw $l_{r=1}$, the run length (i.e. number of words) of $r = 1$:

   $l_1 \sim Poisson(\lambda_{z_1})$.

   The Poisson distribution generating l_1 has lower bound 1 and upper bound N_d.

   This is simply because the length of the first topic run cannot exceed N_d.

   (c) draw l_1 words from $Multinomial(\phi_{t=z_1})$ IID. If $l_1 = N_d$, stop.

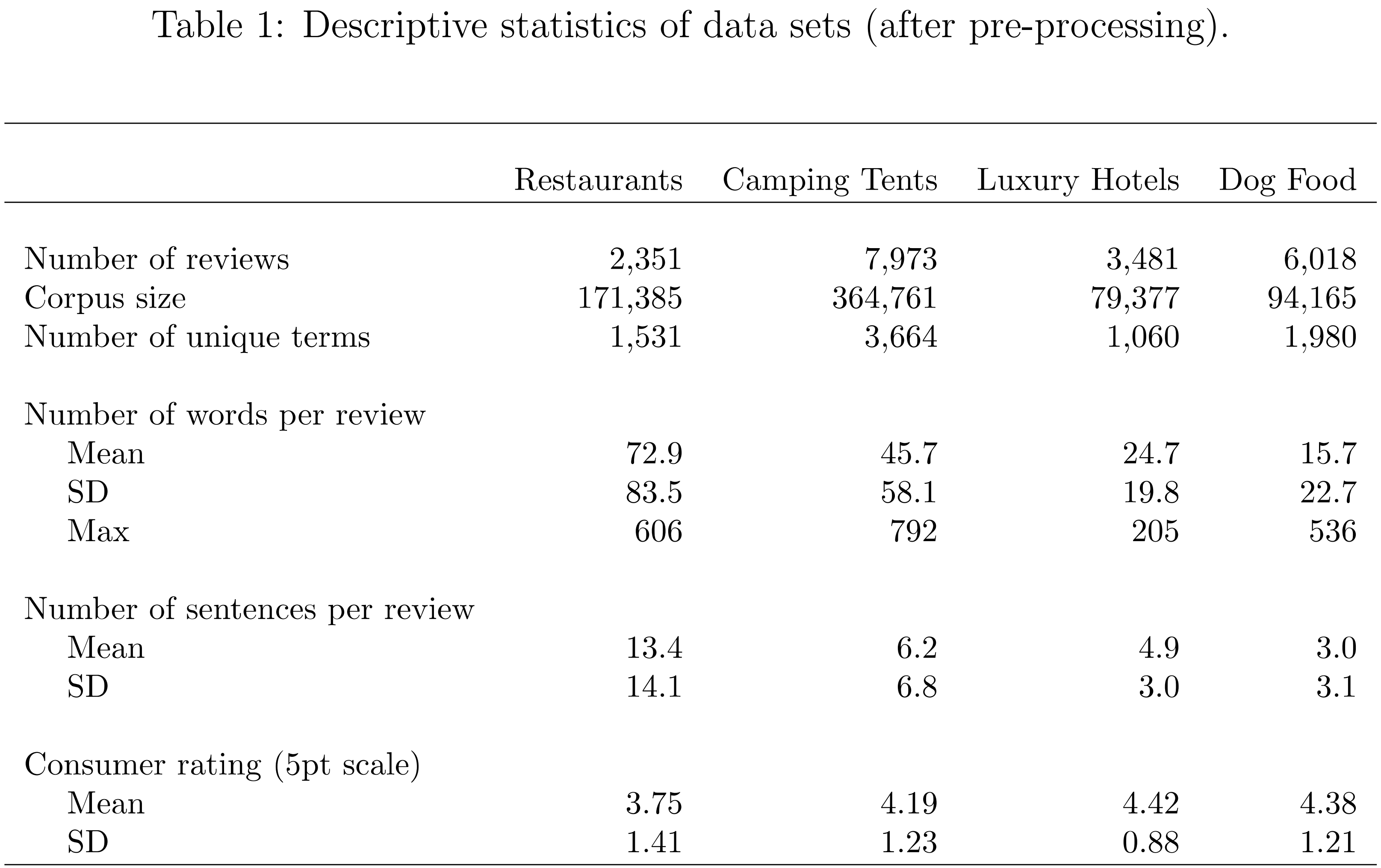


Table 1: Descriptive statistics of data sets (after pre-processing).

	Restaurants	Camping Tents	Luxury Hotels	Dog Food
Number of reviews	2,351	7,973	3,481	6,018
Corpus size	171,385	364,761	79,377	94,165
Number of unique terms	1,531	3,664	1,060	1,980
Number of words per review				
Mean	72.9	45.7	24.7	15.7
SD	83.5	58.1	19.8	22.7
Max	606	792	205	536
Number of sentences per review				
Mean	13.4	6.2	4.9	3.0
SD	14.1	6.8	3.0	3.1
Consumer rating (5pt scale)				
Mean	3.75	4.19	4.42	4.38
SD	1.41	1.23	0.88	1.21

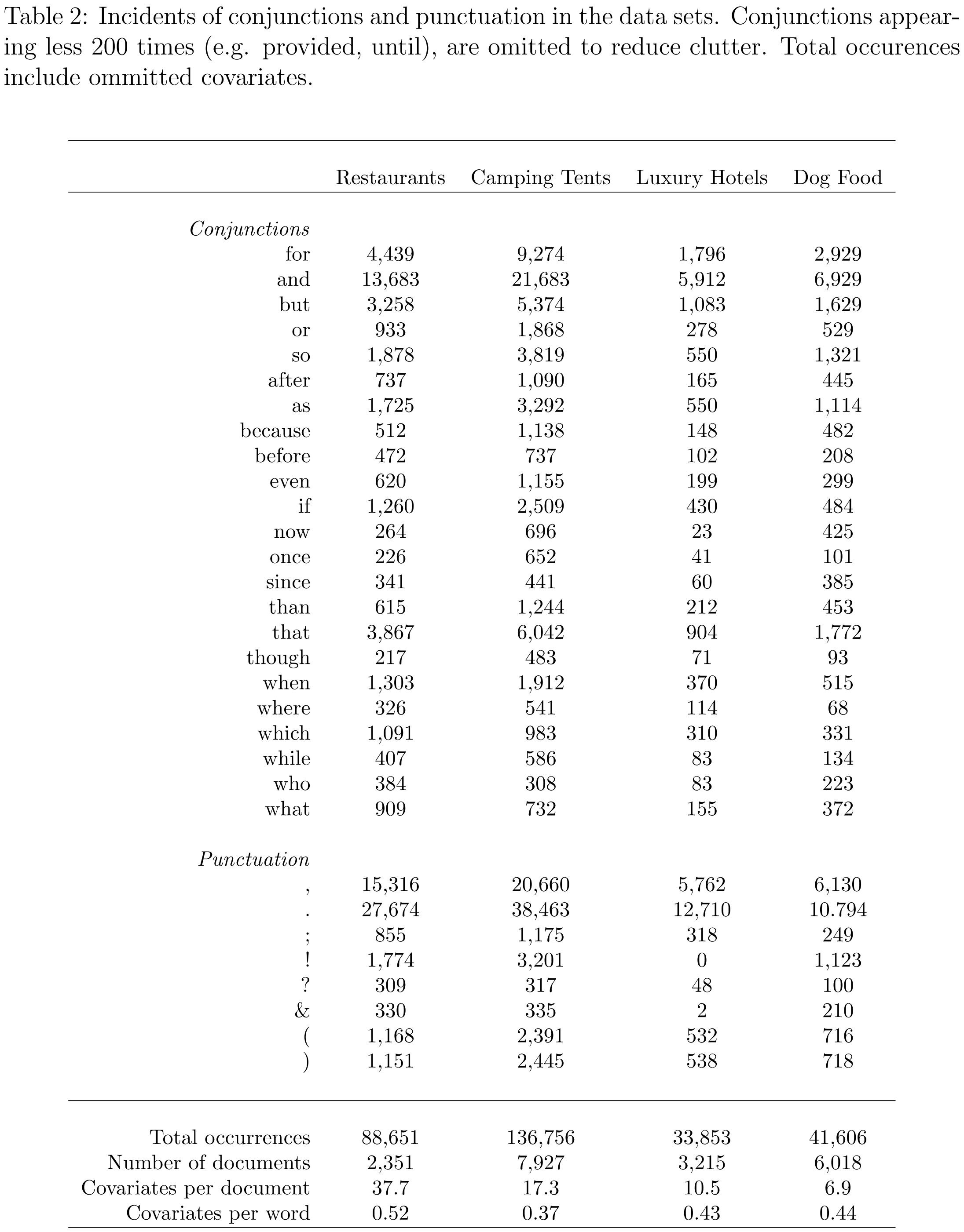


Table 2: Incidents of conjunctions and punctuation in the data sets. Conjunctions appearing less 200 times (e.g. provided, until), are omitted to reduce clutter. Total occurences include ommitted covariates.

	Restaurants	Camping Tents	Luxury Hotels	Dog Food
Conjunctions				
for	4,439	9,274	1,796	2,929
and	13,683	21,683	5,912	6,929
but	3,258	5,374	1,083	1,629
or	933	1,868	278	529
so	1,878	3,819	550	1,321
after	737	1,090	165	445
as	1,725	3,292	550	1,114
because	512	1,138	148	482
before	472	737	102	208
even	620	1,155	199	299
if	1,260	2,509	430	484
now	264	696	23	425
once	226	652	41	101
since	341	441	60	385
than	615	1,244	212	453
that	3,867	6,042	904	1,772
though	217	483	71	93
when	1,303	1,912	370	515
where	326	541	114	68
which	1,091	983	310	331
while	407	586	83	134
who	384	308	83	223
what	909	732	155	372
Punctuation				
,	15,316	20,660	5,762	6,130
.	27,674	38,463	12,710	10.794
;	855	1,175	318	249
!	1,774	3,201	0	1,123
?	309	317	48	100
&	330	335	2	210
(	1,168	2,391	532	716
)	1,151	2,445	538	718
Total occurrences	88,651	136,756	33,853	41,606
Number of documents	2,351	7,927	3,215	6,018
Covariates per document	37.7	17.3	10.5	6.9
Covariates per word	0.52	0.37	0.43	0.44

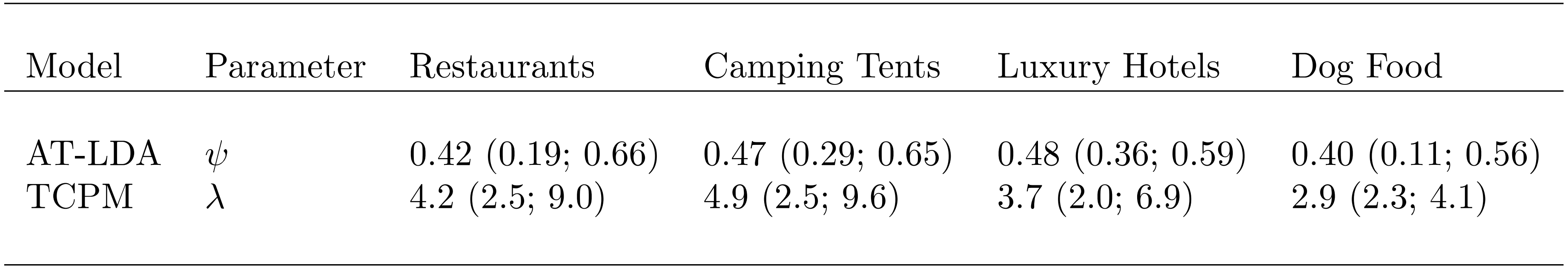
Fit results

Model Category	Model	Restaurant Reviews	Camping Tent Reviews	Luxury Hotel Reviews	Dog Food Reviews
Bag of Words	LDA	-963,344 (10)	-1,878,546 (15)	-421,002 (11)	**-361,274** (8)
Topic Chunking	SC LDA	-1,064,867 (6)	-2,164,162 (14)	-427,964 (10)	-449,155 (8)
	Sticky SC LDA	-1,064,805 (6)	-2,151,230 (14)	-425,981 (10)	-456,477 (8)
	CPC LDA	-1,012,657 (8)	-2,006,406 (14)	-410,286 (10)	-425,301 (8)
	Sticky CPC LDA	-1,052,217 (8)	-2,092,006 (14)	-409,896 (10)	-443,896 (8)
	TCPM	-971,125 (10)	-1,949,083 (14)	-369,832 (10)	-388,957 (9)
Topic Carry-over	AT-LDA w/o covariates	-920,380 (10)	-1,743,012 (16)	**-347,457** (10)	-387,500 (8)
	AT-LDA w covariates	**-916,784** (10)	**- 1,731,181** (16)	-347,824 (10)	-383,425 (8)

Predictive fit results

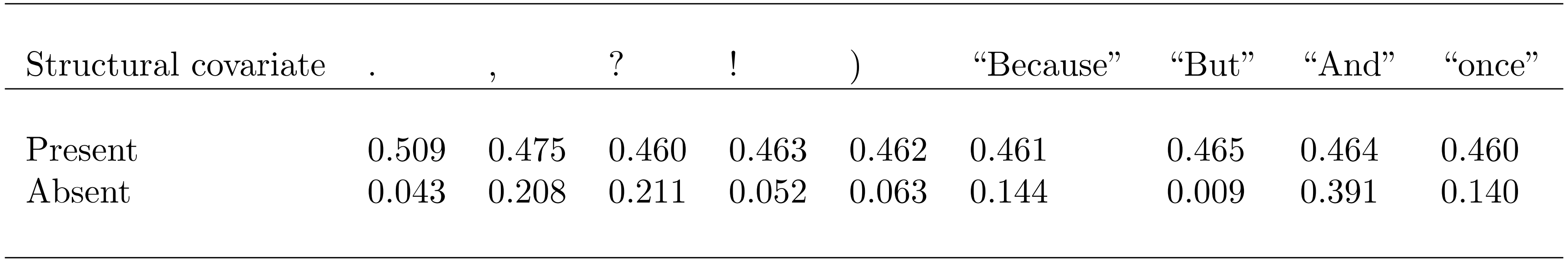
Model Category	Model	Restaurant Reviews	Camping Tent Reviews	Luxury Hotel Reviews	Dog Food Reviews
Bag of Words	LDA	-221,900	-481,697	-93,614	-129,378
Topic Chunking	SC LDA	-221,709	-479,977	-92,957	-129,550
	Sticky SC LDA	-221,640	-478,655	-92,986	-129,365
	CPC LDA	-221,713	-479,509	-92,952	-129,423
	Sticky CPC LDA	-221,594	-478,683	-92,962	-129,289
	TCPM	-221,671	-478,622	-92,942	-129,347
Topic Carry-over	AT-LDA w/o covariates	-221,600	-478,753	-92,976	-129,307
	AT-LDA w covariates	**-220,661**	**-475,974**	**-92,785**	**-128,802**

Topic chunking



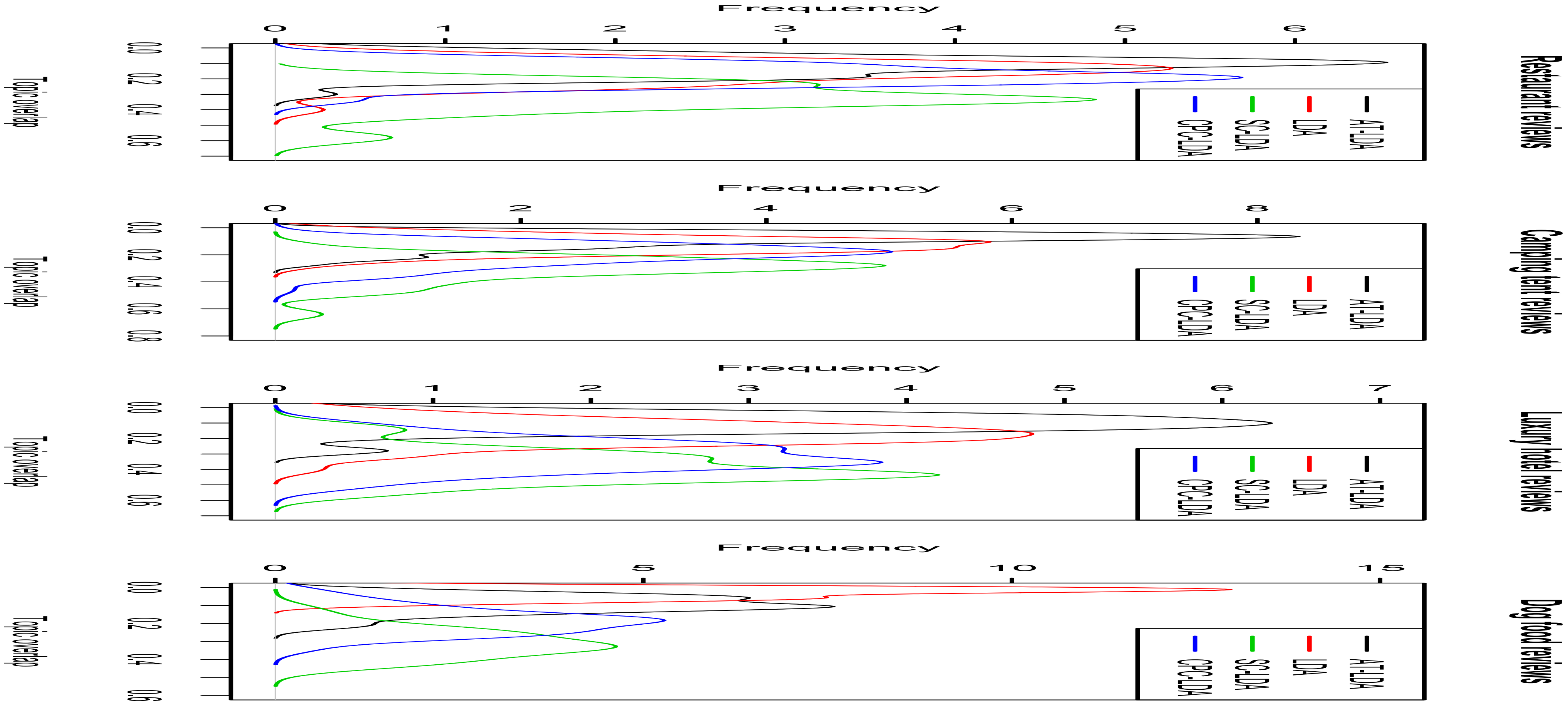
Model	Parameter	Restaurants	Camping Tents	Luxury Hotels	Dog Food
AT-LDA	ψ	0.42 (0.19; 0.66)	0.47 (0.29; 0.65)	0.48 (0.36; 0.59)	0.40 (0.11; 0.56)
TCPM	λ	4.2 (2.5; 9.0)	4.9 (2.5; 9.6)	3.7 (2.0; 6.9)	2.9 (2.3; 4.1)

Change in topic change probability

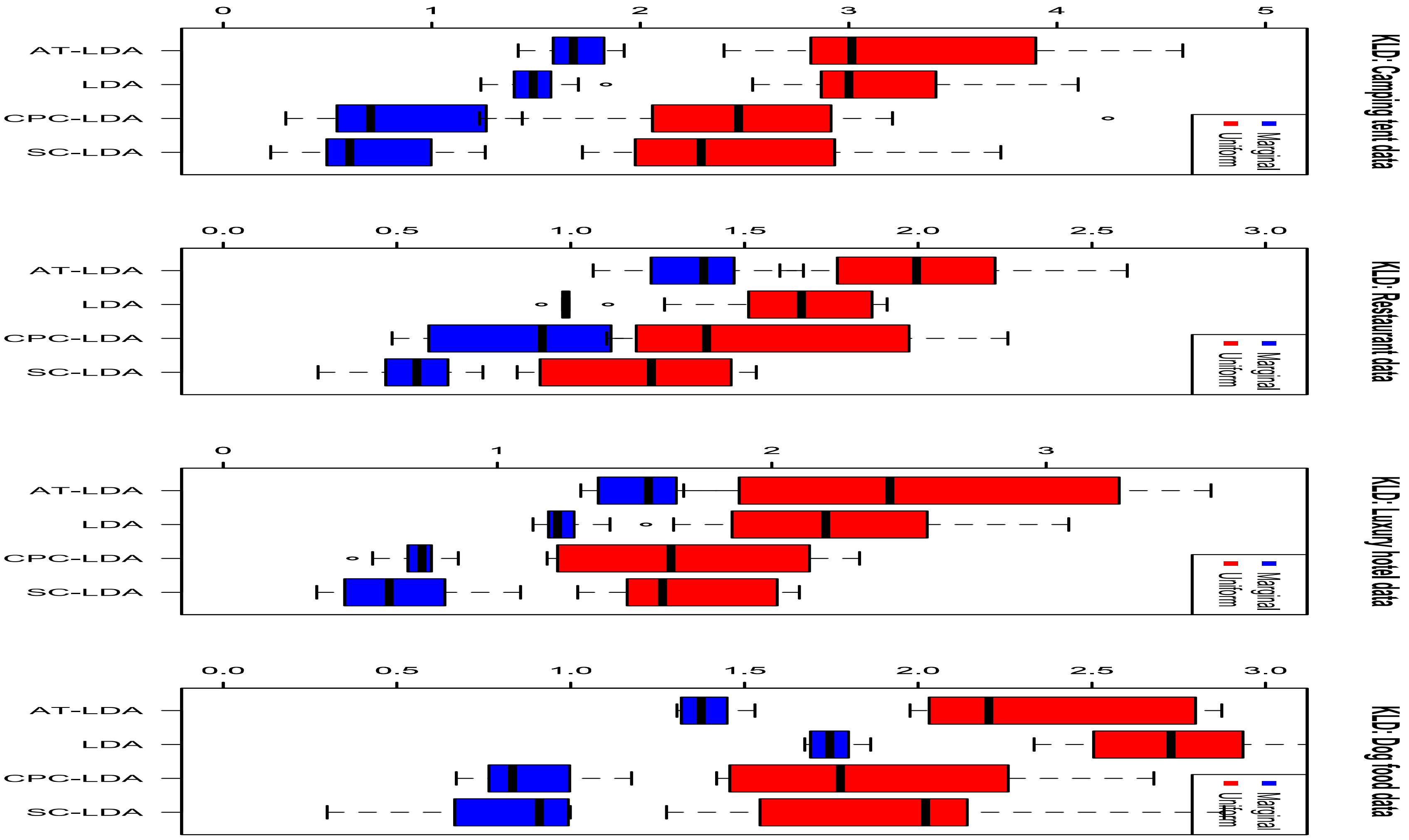


Structural covariate	.	,	?	!	)	“Because”	“But”	“And”	“once”
Present	0.509	0.475	0.460	0.463	0.462	0.461	0.465	0.464	0.460
Absent	0.043	0.208	0.211	0.052	0.063	0.144	0.009	0.391	0.140

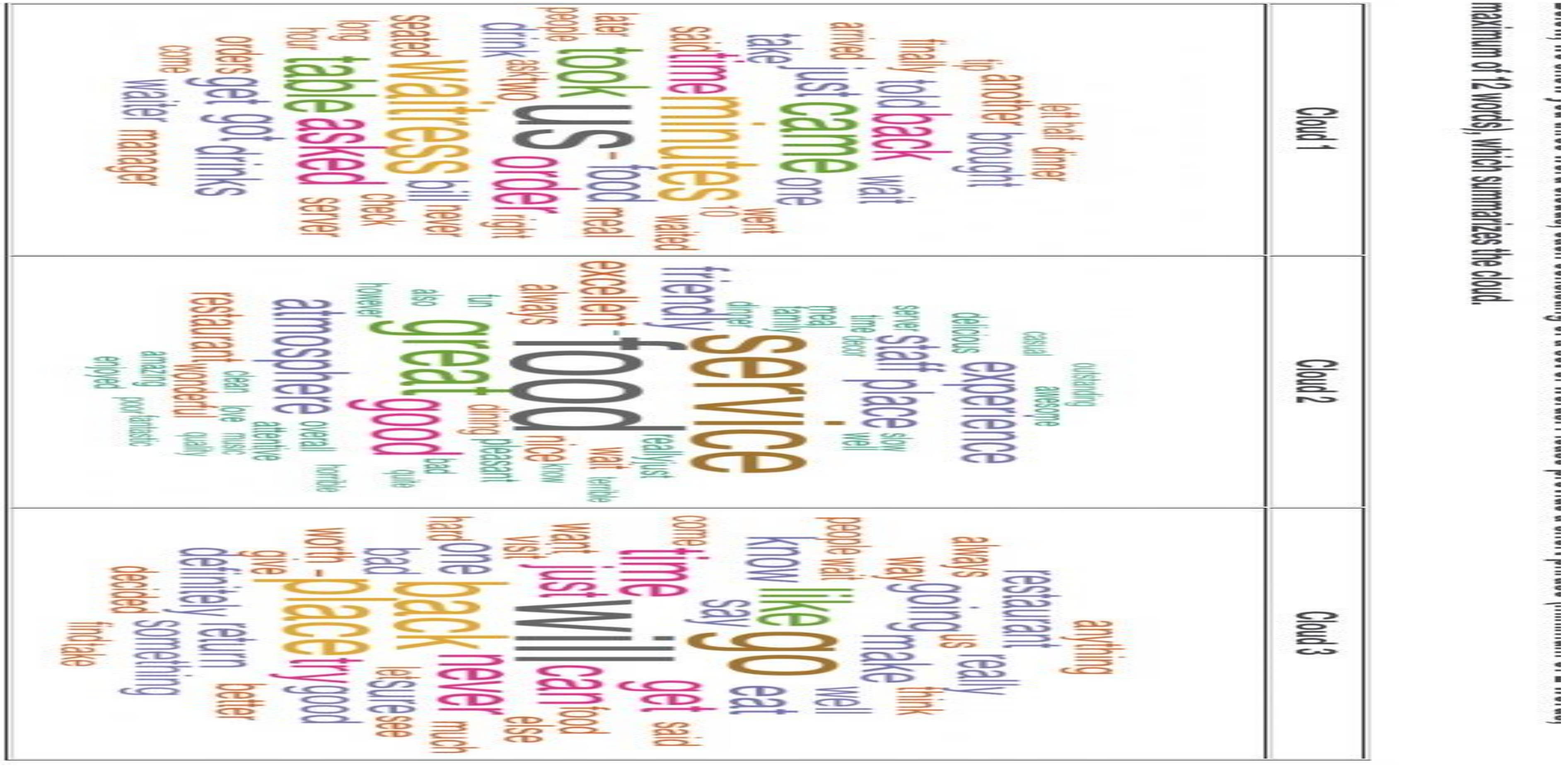
Topic uniqueness – term overlap frequency



Topic coherence – KL divergence



Topic response latency - word cloud stimuli



maximum of 12 words), which summarizes the cloud.

Cloud 1

us
minutes
waitress
came
took
table
asked
order
back
time
food
get
got
drinks
waiter
manager
another
brought
told
just
said
seated
bill
never
server
meal
waited
went
dinner
left
half
check
long
hour
drink
asked
two
people
later
arrived
finally
tip
take
one
wait
come
orders
right
10

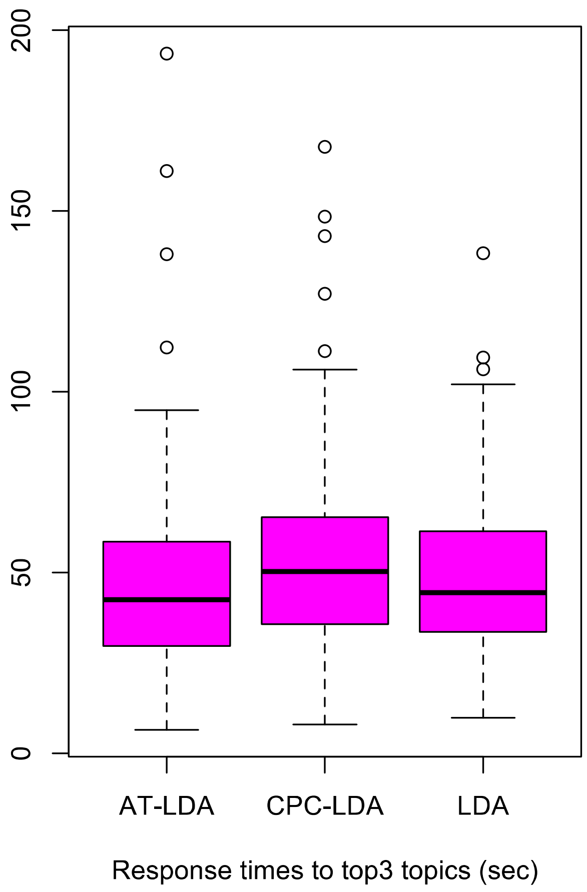
Cloud 2

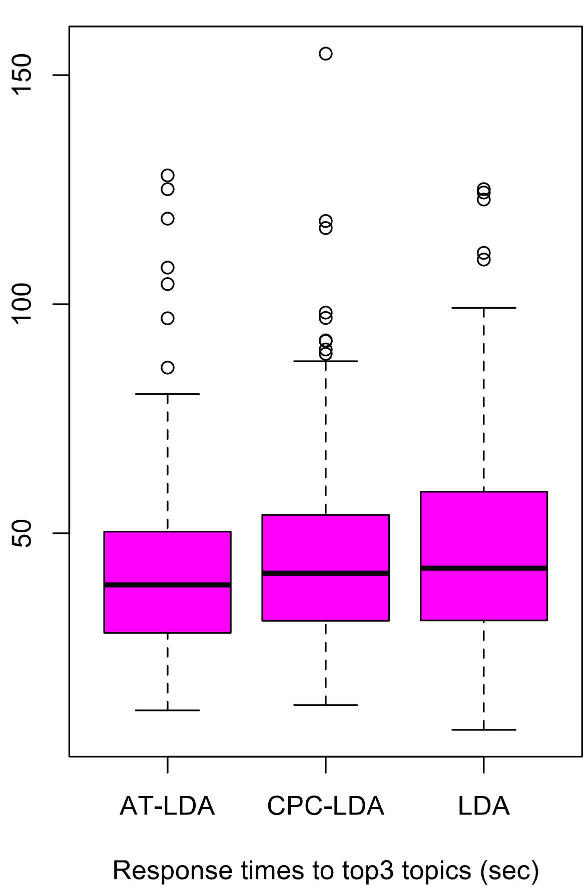
food
service
great
good
atmosphere
experience
staff
place
friendly
excellent
always
restaurant
wonderful
delicious
server
time
decor
meal
family
dinner
fun
also
however
nice
pleasant
dining
really
just
wait
terrible
know
bad
quite
attentive
overall
horrible
clean
love
music
quality
amazing
enjoyed
poor
fantastic
awesome
outstanding
casual
slow
well

Cloud 3

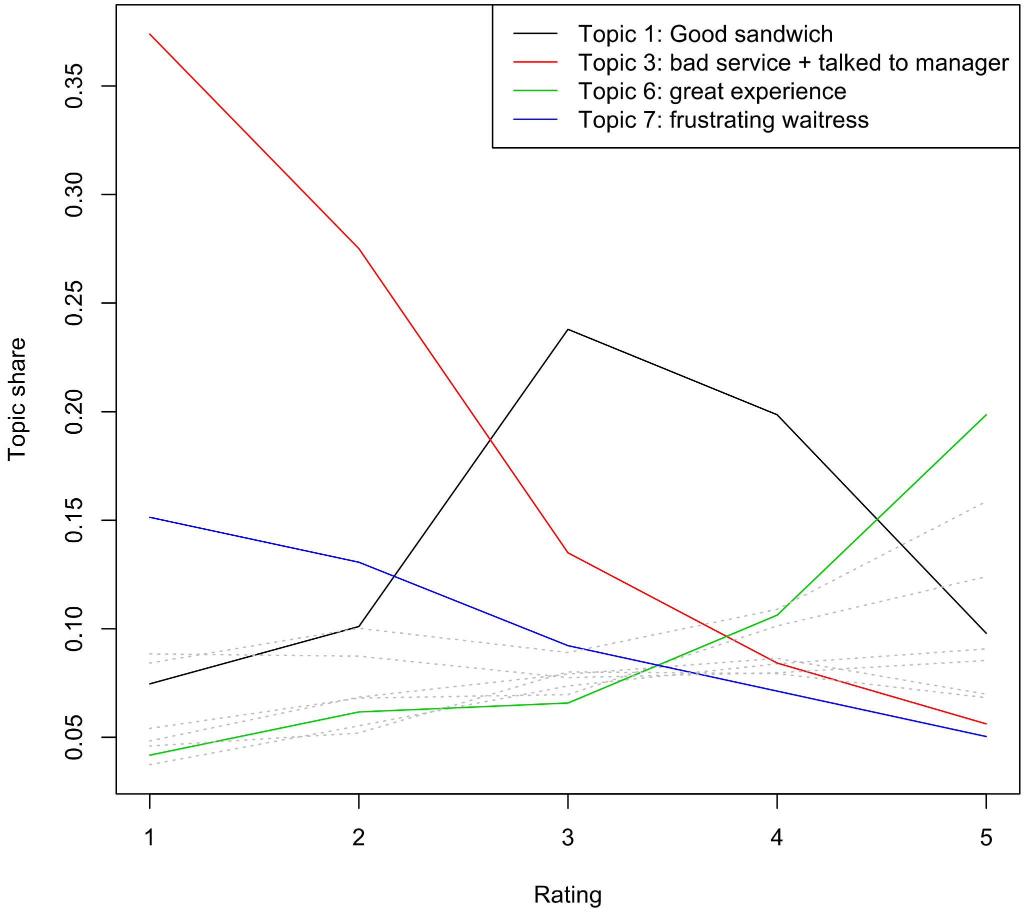
will
go
back
place
like
never
can
time
get
just
want
visit
one
hard
bad
worth
give
definitely
return
decided
something
find
take
try
good
sure
see
much
else
food
said
eat
know
say
people
wait
come
going
make
way
always
us
really
restaurant
anything
think
well
better

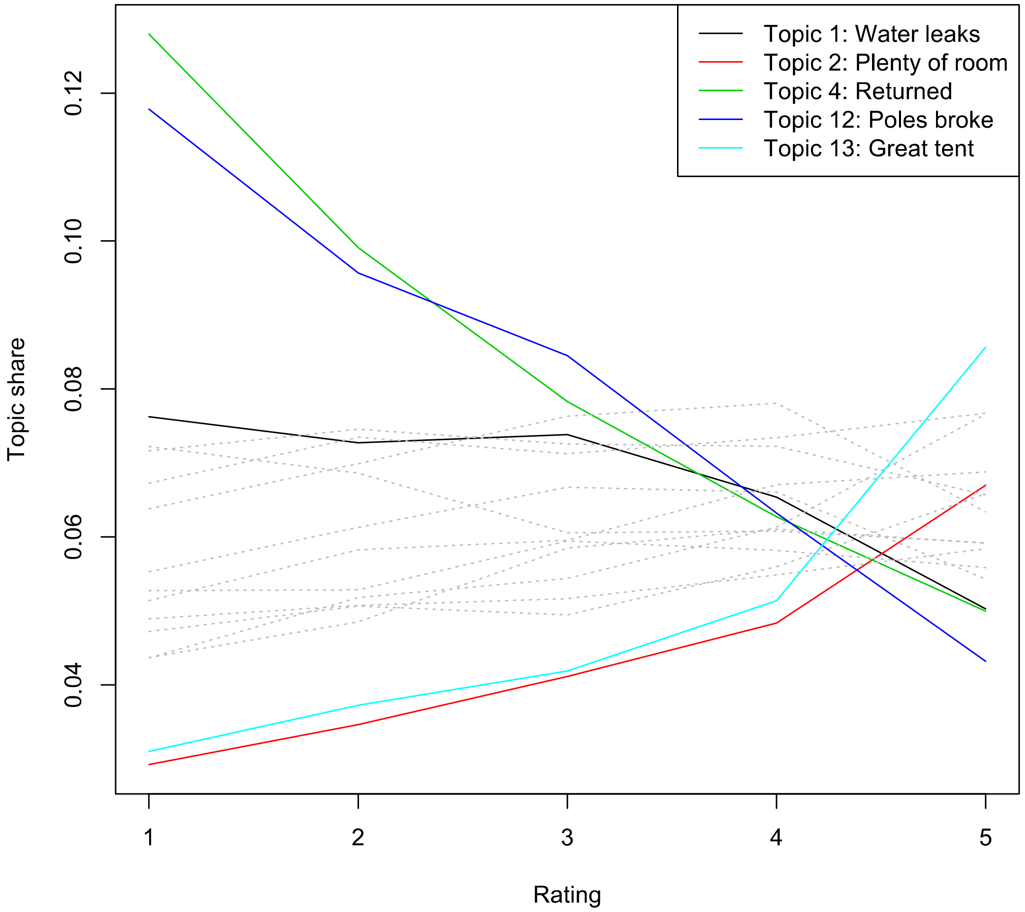
Topic response latency - results


Topics restaurant data


Topics tent data

Distribution of topic probabilities





Example customer reviews

[1] "unprofessional . I was charged on both of my credit cards by the 11 am manager and yet she said my cards did not go through. my bank card companies showed transactions approved on their system. the manager did not take the time to cancel the transaction and try again. my experience was very unpleasant. the cashier did not know what was wrong with the transaction so she asked the manager for help. the manager is apparently very unqualified being that she did not know how to handle the transaction. she just kept sliding the cards over and over. there are many places to eat in Roswell and with poor customer service like this I will not return to this place. I was charged $17 and cents on my cards and was told that my cards had a problem. I did ask for a customer service card and that too was refused. I wonder why?"

[2] "the food at this place is okay, but I found the service to be terrible. the managers are impossible to contact for complaints, and their associates actually tried to bully me! I was called names and hung up on several times when I tried to call the manager. they told me I was 'simple' and 'low' and 'stupid'. nobody has ever said such things about me. it ruined my meal, and my day. I don't care if their food is even excellent (which it's not). I will never eat there again. as of today, I still cannot get hold of the manager to voice my complaints, after leaving multiple messages. I never expected this experience with a pizza place. don't eat there if you don't want insults with your pizza!"

Restaurant Topic 3

[2] "These are my observations in camping in this tent for a week. PROS: The tent did set up very easily and was very roomy with enough room for two queen size air mattresses and gear. We liked the number of windows and the size, it let you get a good breeze through the tent. CONS: This tent is heavy and quite bulky,as you would expect with any instant tent but I found it very cumbersome. I also found the door design a tad bit awkward and kept tripping over the bottom zipper when getting in and out, also with no rain fly when you open the door to get in and out in rain all of the water pours in the tent. That leads me to the biggest issue that I had with the tent, it LEAKED... and definitely not just condensation, humidity or moisture in the tent. I have had plenty of other tents and I know what normal condensation and humidity is and this was definitely rain leaking through at the roof seams. This left us in a damp tent for the better part of a week and made it quite uncomfortable and miserable. Ultimately the cons outweighed the pros for me and I will be searching for a different tent for our next trip."

[3] "Spacious tent, good ventilation. It is NOT WATER prof, had it in the rain, if you touch the inside of the tent with a finger, there will be a drip in that spot. The one would expect waterproof fabric for the money spent. However, if you don't touch it form the inside, it will not drip, so it can be managed. The big pro is that it is EASY to set up. It sets up in seconds."

Camping Tent Topic 1

Restaurant topics

Rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	really	pizza	food	menu	cheese	food	us	bar	go	night
2	good	best	service	salad	sauce	great	came	room	will	times
3	got	italian	restaurant	ordered	chicken	service	waitress	dining	back	dinner
4	just	new	one	wine	fries	good	table	restaurant	can	time
5	like	chicago	just	good	sandwich	place	minutes	tables	place	years
6	little	style	said	also	meat	atmosphere	order	area	time	lunch
7	pretty	place	bad	pasta	flavor	staff	took	back	get	day
8	one	restaurant	like	dessert	bacon	friendly	back	street	try	rib
9	burger	barbecue	never	best	barbecue	restaurant	get	right	eat	many
10	ordered	crust	manager	dinner	bread	experience	asked	parking	make	last
11	much	home	ordered	large	fresh	excellent	drinks	building	one	breakfast
12	sandwich	ever	experience	salads	pork	nice	got	lot	going	year
13	menu	pizzas	husband	dishes	came	always	food	near	never	restaurant
14	went	great	will	chicken	burger	recommend	time	small	say	went
15	fries	york	good	delicious	served	prices	brought	table	like	2
16	also	cheese	told	fresh	salad	wait	drink	front	sure	several
17	meal	places	owner	one	hot	wonderful	bill	inside	definitely	people
18	said	one	however	items	sweet	well	take	around	know	party
19	still	good	steak	dish	taste	pleasant	wait	side	want	sunday
20	lot	family	nothing	house	onion	reasonable	seated	located	visit	week

Ratings sub-model

Cut-point model for ratings:

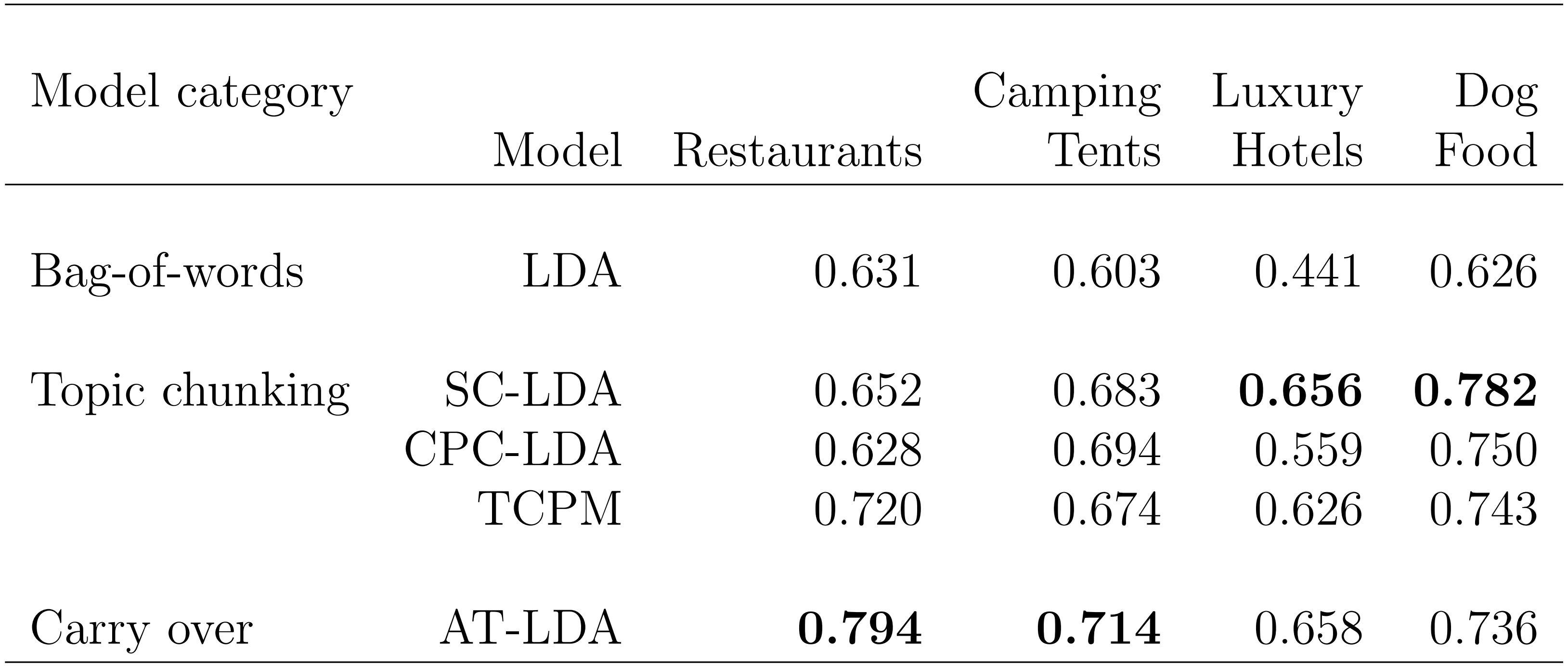
$$r_d = k \quad \text{if} \quad c_{k-1} \leq \tau_d \leq c_k$$

Latent regression on topic probabilities:

$$\tau_d \sim N(\theta'_d \beta, \sigma)$$

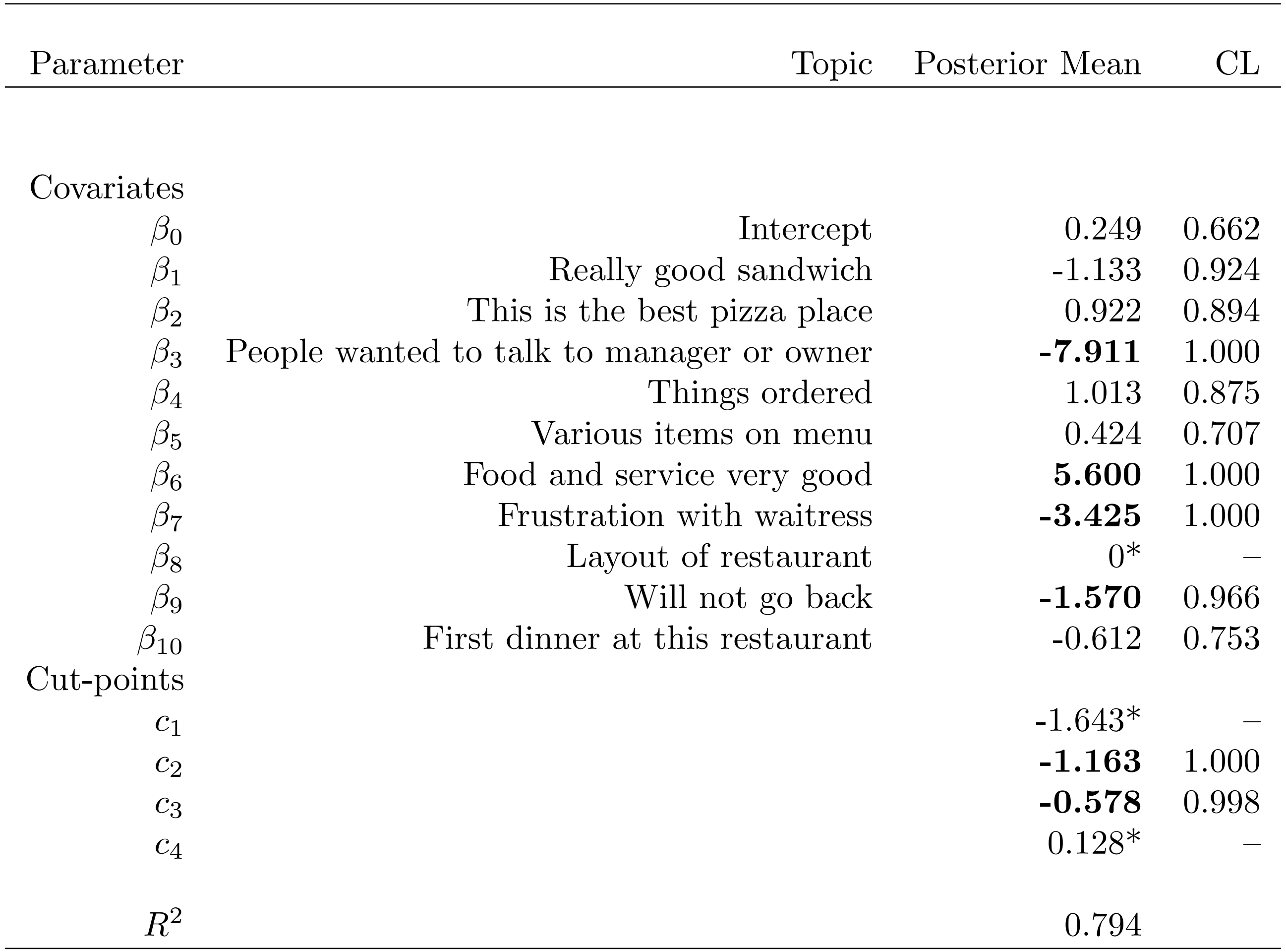
Topic probabilities for each review

Predicting ratings



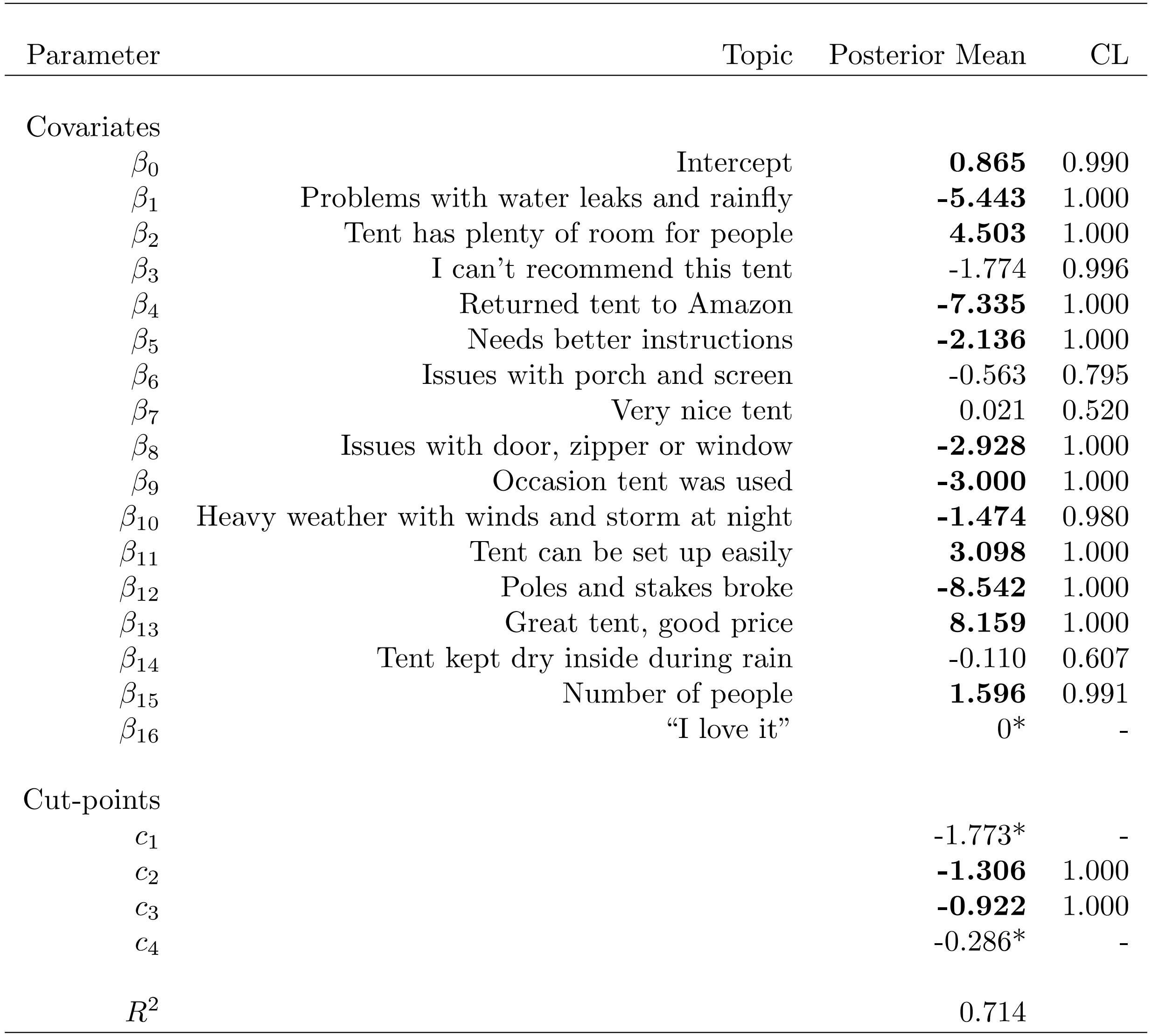
Model category			Camping	Luxury	Dog
	Model	Restaurants	Tents	Hotels	Food
Bag-of-words	LDA	0.631	0.603	0.441	0.626
Topic chunking	SC-LDA	0.652	0.683	**0.656**	**0.782**
	CPC-LDA	0.628	0.694	0.559	0.750
	TCPM	0.720	0.674	0.626	0.743
Carry over	AT-LDA	**0.794**	**0.714**	0.658	0.736

Restaurant regression



Parameter	Topic	Posterior Mean	CL
Covariates			
β_0	Intercept	0.249	0.662
β_1	Really good sandwich	-1.133	0.924
β_2	This is the best pizza place	0.922	0.894
β_3	People wanted to talk to manager or owner	**-7.911**	1.000
β_4	Things ordered	1.013	0.875
β_5	Various items on menu	0.424	0.707
β_6	Food and service very good	**5.600**	1.000
β_7	Frustration with waitress	**-3.425**	1.000
β_8	Layout of restaurant	0*	–
β_9	Will not go back	**-1.570**	0.966
β_{10}	First dinner at this restaurant	-0.612	0.753
Cut-points			
c_1		-1.643*	–
c_2		**-1.163**	1.000
c_3		**-0.578**	0.998
c_4		0.128*	–
R^2		0.794	

Camping Tent regression



Parameter	Topic	Posterior Mean	CL
Covariates			
β_0	Intercept	**0.865**	0.990
β_1	Problems with water leaks and rainfly	**-5.443**	1.000
β_2	Tent has plenty of room for people	**4.503**	1.000
β_3	I can't recommend this tent	-1.774	0.996
β_4	Returned tent to Amazon	**-7.335**	1.000
β_5	Needs better instructions	**-2.136**	1.000
β_6	Issues with porch and screen	-0.563	0.795
β_7	Very nice tent	0.021	0.520
β_8	Issues with door, zipper or window	**-2.928**	1.000
β_9	Occasion tent was used	**-3.000**	1.000
β_{10}	Heavy weather with winds and storm at night	**-1.474**	0.980
β_{11}	Tent can be set up easily	**3.098**	1.000
β_{12}	Poles and stakes broke	**-8.542**	1.000
β_{13}	Great tent, good price	**8.159**	1.000
β_{14}	Tent kept dry inside during rain	-0.110	0.607
β_{15}	Number of people	**1.596**	0.991
β_{16}	"I love it"	0*	-
Cut-points			
c_1		-1.773*	-
c_2		**-1.306**	1.000
c_3		**-0.922**	1.000
c_4		-0.286*	-
R^2		0.714	

Next steps

- Other research
  - Incorporation into choice models
  - GoM models and extremes