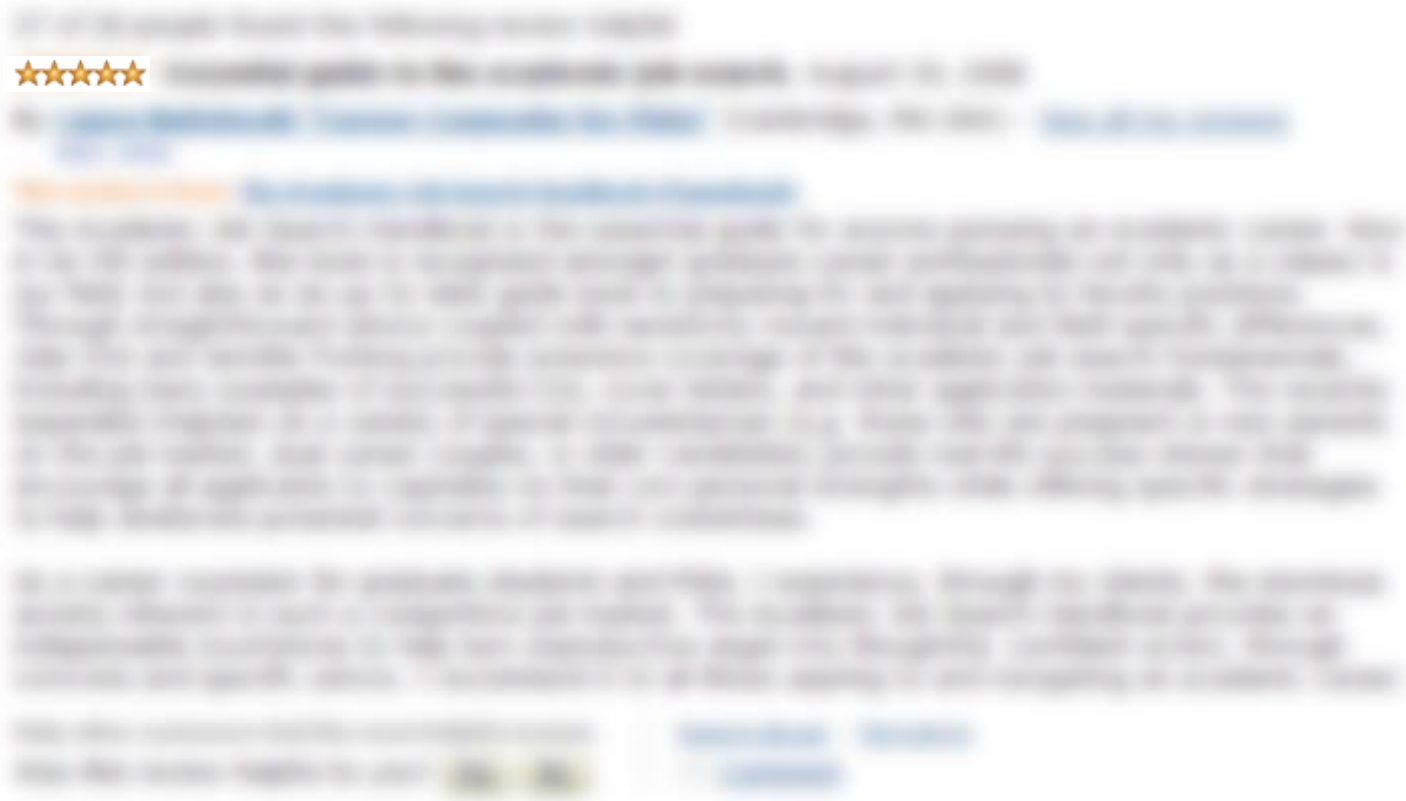


Harnessing reviews to build richer models of opinions

Julian McAuley, UC San Diego

Opinions



“Do I like the product?”

Opinions – a richer view

By [Laura Malisheski "Career Counselor for PhDs"](#) (Cambridge, MA USA) - [See all my reviews](#)

REAL NAME

This review is from: [The Academic Job Search Handbook \(Paperback\)](#)

The Academic Job Search Handbook is the essential guide for anyone pursuing an academic career. Now in its 4th edition, this book is recognized amongst graduate career professionals not only as a classic in our field, but also as an up-to-date guide book to preparing for and applying to faculty positions. Through straightforward advice coupled with sensitivity toward individual and field-specific differences, Julie Vick and Jennifer Furlong provide extensive coverage of the academic job search fundamentals, including many examples of successful CVs, cover letters, and other application materials. The recently expanded chapters on a variety of special circumstances (e.g. those who are pregnant or new parents on the job market, dual career couples, or older candidates) provide real-life success stories that encourage all applicants to capitalize on their own personal strengths while offering specific strategies to help ameliorate potential concerns of search committees.

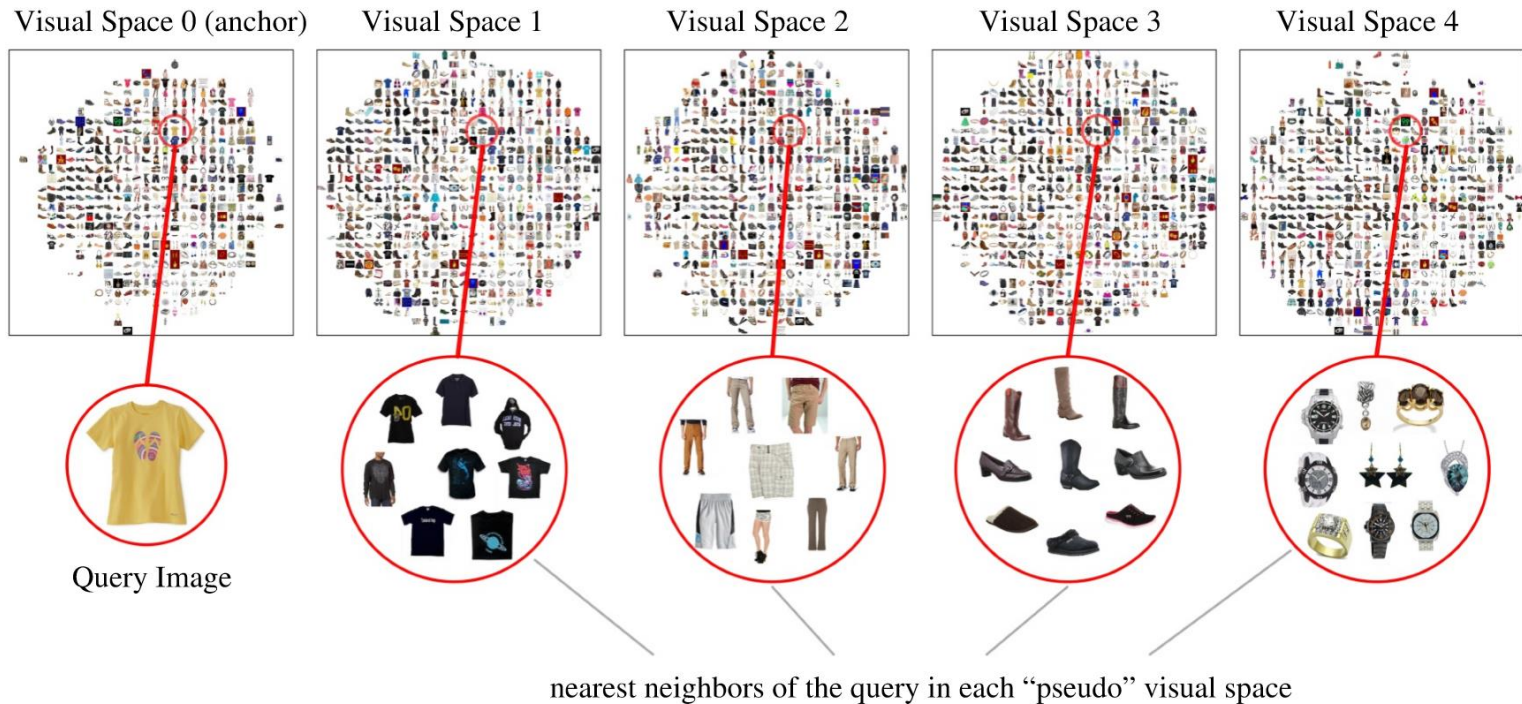
As a career counselor for graduate students and PhDs, I experience, through my clients, the enormous anxiety inherent in such a competitive job market. The Academic Job Search Handbook provides an indispensable touchstone to help turn unproductive angst into thoughtful, confident action, through concrete and specific advice. I recommend it to all those aspiring to and navigating an academic career.

"Why do I like the product?"

How can we build review text into models of people's opinions?

1. How can **latent factor models** be extended to incorporate review content? (RecSys'13)
2. How can reviews be used to **answer questions** about products? (WWW'16)
3. How can personalized reviews be **generated?** (arXiv)

Also: fashion recommendation



- Social recommendation
- Temporal recommendation, sequential recommendation
- etc.

The Amazon logo, featuring the word "amazon" in a bold, black, lowercase sans-serif font. Below the text is a curved orange arrow that starts under the letter 'a' and ends under the letter 'n', pointing to the right.

~**100M** reviews, ~**10M** items, ~**20M** users
1.4M questions and answers

The Beeradvocate logo, featuring the word "Beer" in a bold, orange, lowercase sans-serif font, followed by the word "advocate" in a bold, black, lowercase sans-serif font.

~**3M** reviews, ~**60k** items, ~**30k** users

1. Ratings & text

TE HOTELS
inia Hotels
reservation
es all of the
inia Hotels
(212) 320-8000
11-free: 1-855-246-1234
www.affinia.com

Recommending things to people

We'd like to estimate users' preferences toward items

Could be a (star) rating

$$f(u, i) : U \times I \rightarrow \{1, 2, 3, 4, 5\}$$

rating(julian, Harry Potter) = ?

$$f(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

global offset

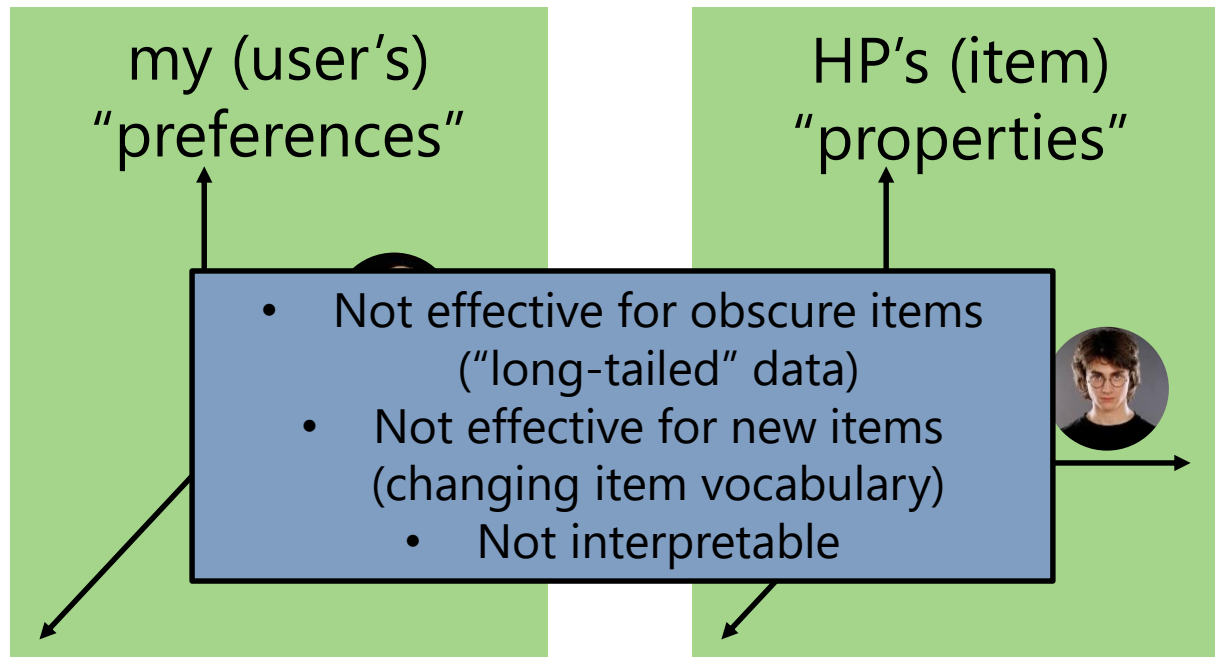
user/item biases

user/item interaction

Recommending things to people

learn my **preferences**, and the product's **properties**

e.g. rating(julian, Harry Potter) =



Latent Dirichlet Allocation

Observation: **Can't** model low-dimensional structure from a single rating, but **can** model it from a single review

87 of 102 people found the following review helpful

★★★★★ **You keep what you kill**, December 27, 2004

By [Schtinky "Schtinky"](#) (Washington State) - [See all my reviews](#)

VINE™ VOICE

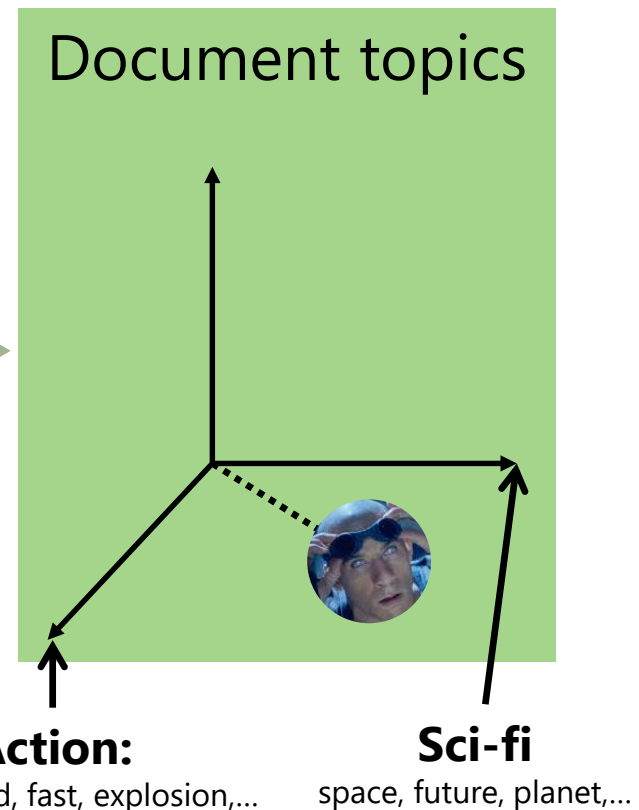
This review is from: [The Chronicles of Riddick \(Widescreen Unrated Director's Cut\) \(DVD\)](#)

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from 'Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to 'Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

LDA →



Combining ratings and reviews

The parameters of a “standard” recommender system

$$rec(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

user/item offset user/item bias latent factors

are fit so as to minimize the mean-squared error

$$\arg \min_{\alpha, \beta, \gamma} \frac{1}{|\mathcal{T}|} \sum_{r_{u,i} \in \mathcal{T}} \underbrace{(rec(u, i) - r_{u,i})^2}_{\text{rating error}} + \underbrace{\lambda \|\gamma\|_2^2}_{\text{regularizer}}$$

where $r_{u,i} \in \mathcal{T}$ is a training corpus of ratings

Note: “compatibility” is ignored when there is too little training data

Combining ratings and reviews

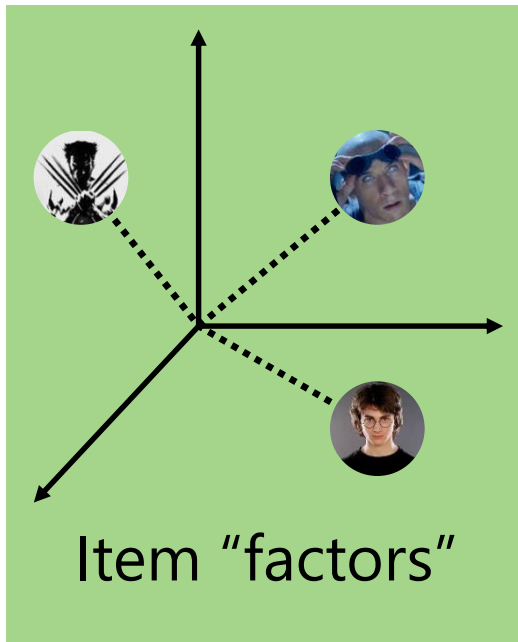
We replace this objective with one that uses the **review text** as a regularizer:

$$\frac{1}{|\mathcal{T}|} \sum_{r_{u,i} \in \mathcal{T}} \underbrace{(rec(u,i) - r_{u,i})^2}_{\text{rating error}} - \mu \underbrace{l(\mathcal{T} | \Theta, \phi, z)}_{\text{corpus likelihood}}$$

rating parameters
 $\alpha, \beta_u, \beta_i, \gamma_u, \gamma_i$

LDA parameters
 Θ, ϕ, z

Combining ratings and reviews



transform

$$\theta_{i,k} = \frac{\exp(\kappa\gamma_{i,k})}{\sum_{k'} \exp(\kappa\gamma_{i,k})}$$



By linking rating and opinion models, we can find topics in reviews that **inform us** about opinions

Model fitting

Repeat steps (1) and (2) until convergence:

$$\arg \min_{\Theta} \frac{1}{|\mathcal{T}|} \sum_{r_{u,i} \in \mathcal{T}} \underbrace{(rec(u, i) - r_{u,i})^2}_{\text{rating error}} - \mu \underbrace{l(\mathcal{T} | \Theta, \phi, z)}_{\text{corpus likelihood}}$$

solved via gradient ascent using L-BFGS
(see e.g. Koren & Bell, 2011)

Step 1:
fit a rating
model
regularized by
the topics

sample $z_{d,j}$ with probability $p(z_{d,j} = k) = \phi_{k,w_{d,j}}$

solved via Gibbs sampling
(see e.g. Blei & McAuliffe, 2007)

Step 2:
identify topics
that "explain"
the ratings

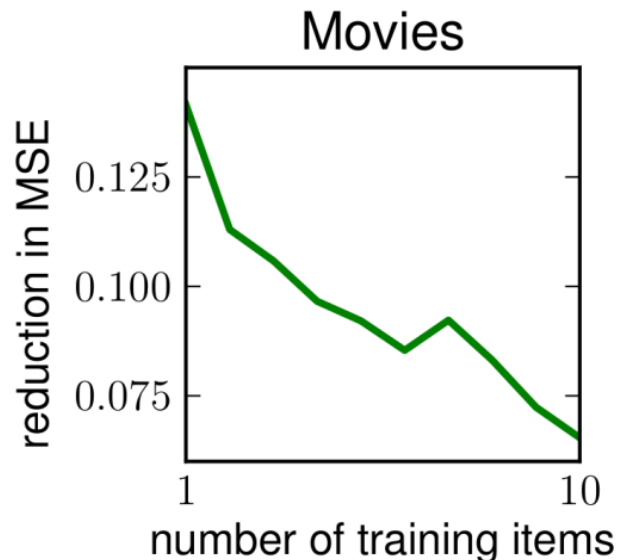
Experiments

Rating prediction:

- Amazon (35M reviews): 6% better than state-of-the-art
- Yelp (230K reviews): 4% better than state-of-the-art

New users:

- Improvements are largest for users with few reviews:



Experiments

Interpretability:

Topics are highly interpretable across all datasets

Beers

pale ales	lambics	dark beers	spices	wheat beers
ipa	funk	chocolate	pumpkin	wheat
pine	brett	coffee	nutmeg	yellow
grapefruit	saison	black	corn	straw
citrus	vinegar	dark	cinnamon	pilsner
ipas	raspberry	roasted	pie	summer
piney	lambic	stout	cheap	pale
citrusy	barnyard	bourbon	bud	lager
floral	funky	tan	water	banana
hoppy	tart	porter	macro	coriander
dipa	raspberries	vanilla	adjunct	pils

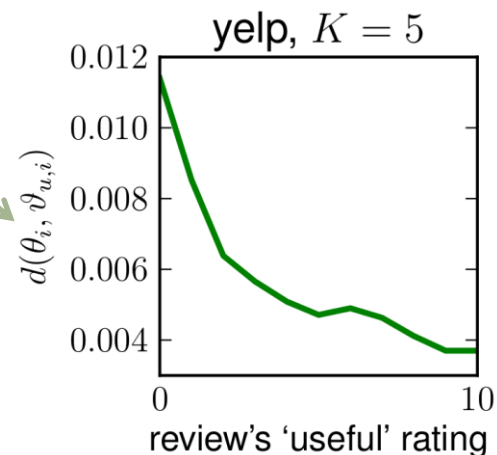
Musical Instruments

drums	strings	wind	mics	software
cartridge	guitar	reeds	mic	software
sticks	violin	harmonica	microphone	interface
strings	strap	cream	stand	midi
snare	neck	reed	mics	windows
stylus	capo	harp	wireless	drivers
cymbals	tune	fog	microphones	inputs
mute	guitars	mouthpiece	condenser	usb
heads	picks	bruce	battery	computer
these	bridge	harmonicas	filter	mp3
daddario	tuner	harps	stands	program

Other ideas...

- Although we used *reviews*, the idea can be adapted to any corpus where we have user or item text and product ratings
 - The notion of what topics are relevant to variance in opinions is a useful measure of the “helpfulness” of a review

Do the topics in **my** review match those that **the community** find important?



- Can also apply the same ideas with more “modern” document embedding techniques



2. Using reviews to answer questions

Answering product-related queries



Q: "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

Suppose we want to answer the question above.
Should we:

- 1) Wade through (hundreds of!) existing reviews looking for an answer → time consuming
- 2) Ask the community via a Q/A system? → have to wait
- 3) Can we answer the question **automatically?**

Answering product-related queries



Q: "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

Challenging!

- The question itself is complex (not a simple query)
- Answer (probably?) won't be in a knowledge base
- Answer is subjective (how loud is "loud enough"?)

Answering product-related queries



Q: "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

So, let's use **reviews** to find possible answers:

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

Answering product-related queries



Q: "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

Still challenging!

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

- Text is only tangentially related to the question
- Text is linguistically quite different from the question
- Combination of positive, negative, and lukewarm answers to resolve

Answering product-related queries



Q: "I want to use this with my iPad air while taking a jacuzzi bath. Will the volume be loud enough over the bath jets?"

So, let's aggregate the results of many reviews

"The sound quality is great, especially for the size, and if you place the speaker on a hard surface it acts as a sound board, and the bass really kicks up."

Yes

"If you are looking for a water resistant blue tooth speaker you will be very pleased with this product."

Yes

"However if you are looking for something to throw a small party this just doesn't have the sound output."

No

=Yes



Challenges

- 1.** Question, answers, and reviews are linguistically heterogeneous
- 2.** Questions may not be answerable from the knowledge base, or may be subjective
- 3.** Many questions are non-binary

Linguistic heterogeneity

Question, answers, and reviews are linguistically heterogeneous

How might we estimate whether a review is “relevant” to a particular question?

1. Cosine similarity?  (won't pick out important words)
2. Tf-idf (e.g. BM25 or similar)?  (won't handle synonyms)
- 3. Bilinear models**

$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} W \mathbf{x}_{\text{review}}^T$$

Linguistic heterogeneity

$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} W \mathbf{x}_{\text{review}}^T$$



$$\text{relevance}(\mathbf{x}_{\text{question}}, \mathbf{x}_{\text{review}}) = \mathbf{x}_{\text{question}} (AB^T + \Delta) \mathbf{x}_{\text{review}}^T$$

(note: also allows questions and reviews to have different features)

- A and B embed the text to account for synonym use, Δ accounts for (weighted) word-to-word similarity
 - But how do we learn the parameters?

Parameter fitting

- We have a high-dimensional model whose parameters describe how relevant each review is to a given question
 - But, we have no **training data** that tells us what is relevant and what isn't
- But we *do* have training data in the form of **answered questions!**

Idea: A **relevant** review is one that helps us to **predict** the correct answer to a question

Parameter fitting

$$p(\text{answer is yes} \mid \text{question } q) = \sum_{r \in \text{reviews}} p(r \text{ is relevant} \mid q) p(\text{yes} \mid r, q)$$

"prediction"
 $\propto \mathbf{x}_{\text{question}} (A' B'^T + \Delta') \mathbf{x}_{\text{review}}^T$

$\propto \mathbf{x}_{\text{question}} (AB^T + \Delta) \mathbf{x}_{\text{review}}^T$
"relevance"

"mixture of experts"

Fit by maximum-likelihood:

Extracting yes/no questions:
"Summarization of yes/no questions using a feature function model" (He & Dai, '11)

$$\ell(\text{corpus}) = \sum_{q \in q_{\text{yes}}} \log p(\text{yes} \mid q) + \sum_{q \in q_{\text{no}}} \log(1 - p(\text{yes} \mid q))$$

Non-binary questions



What about **open-ended** questions?

Mommy's Helper Kid Keeper (amazon.com/dp/B00081L2SU)

Q: "I have a big two year old (30 lbs) who is very active and pretty strong. Will this harness fit him? Will there be any room to grow?"

A: "One of my two year olds is 36lbs and 36in tall. It fits him. I would like for there to be more room to grow, but it should fit for a while. "

- It's no longer practical to predict the answer directly
- But we can still predict whether a review is **relevant**

Non-binary questions

- The model should rank the “true” answer higher than “non”-answers

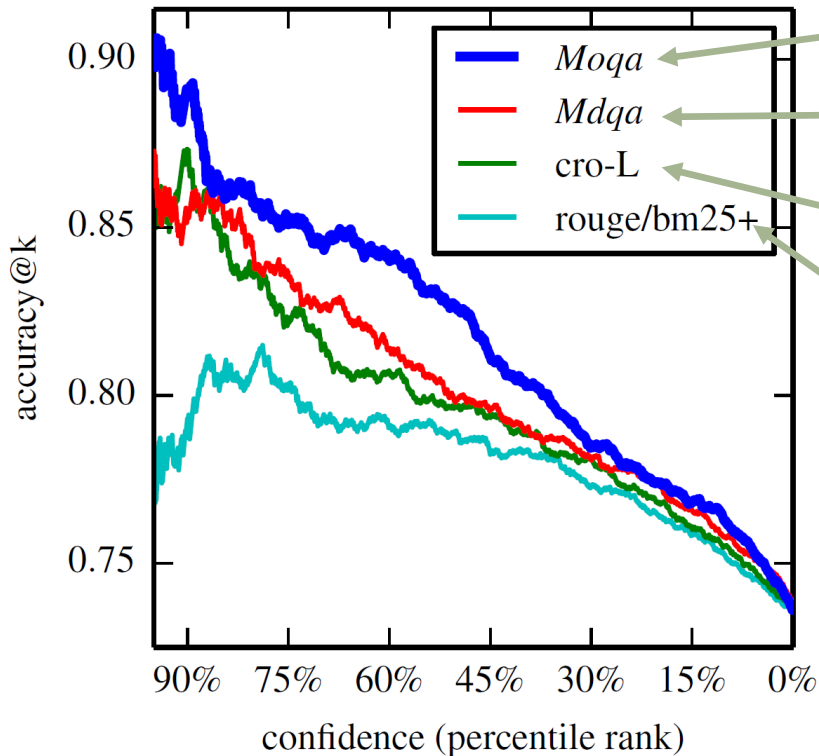
$$p(\text{answer } a > \bar{a} \mid \text{question } q) = \sum_{r \in \text{reviews}} p(r \text{ is relevant} \mid q) p(a > \bar{a} \mid r)$$


$$\propto (\mathbf{x}_a - \mathbf{x}_{\bar{a}})(A'B'^T + \Delta')\mathbf{x}_{\text{review}}^T$$

- We still train by maximum likelihood, sampling many non answers at training time
- Note that at test time (in practice) we'd only use the relevance function, since candidate answers wouldn't be available

Evaluation – binary questions

Accuracy versus confidence (electronics)



Mixtures-of-Opinions for QA

Mixtures-of-Descriptions

Various off-the-shelf similarity measures w/ learned weights

No learning

(~300k questions and answers)

Open-ended questions (AUC)

Dataset	Moqa	Mdqa	cro-L	Rouge
Electronics	0.912	0.865	0.855	0.626
Average	0.883	0.841	0.828	0.631

$$| p(\text{yes}) - 0.5 |$$


Evaluation – user study

mturk interface:

Instructions

Consider a customer's query about the following product:

Think King Mighty Buggy Hook for Stroller, Wheelchair, Rollator, Walker, 2 Pack



" Since the hooks attach with velcro, do they slide or do they stay in place? "

Which of the following sentences is **most relevant** to the above question?

"I originally purchased the Mommy Hooks for our stroller and loved the durability of the metal and being able to put large amounts of stuff on them, but i ended up hating how big and clunky they are especially when folding the stroller and they are not stationary, always sliding around."

"With the hooks attached at the highest part of the main handle, bags that are hung from the hooks press against both the bassinet and the footrest of the backwards-facing seat, but not in such a way that the hooks are unusable."

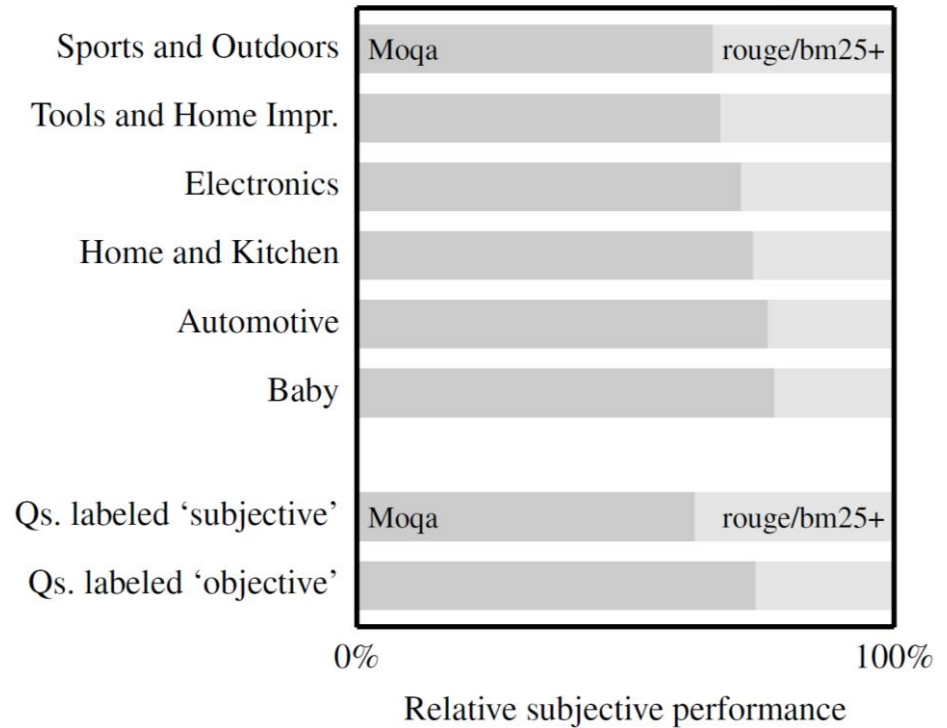
"The hooks stay in place even with multiple bags hanging on it."

Would you say that this question is **subjective**?

Yes

No

Mechanical turk study



Evaluation – binary examples

Product: Schwinn Searcher Bike (amazon.com/dp/B007CKH61C)

Question: "Is this bike a medium? My daughter is 5'8"."

Ranked opinions: "The seat was just a tad tall for my girl so we actually sawed a bit off of the seat pole so that it would sit a little lower." (yes, .698); "The seat height and handlebars are easily adjustable." (yes, .771); "This is a great bike for a tall person." (yes, .711)

Response: Yes (.722)

Actual answer: My wife is 5'5" and the seat is set pretty low, I think a female 5'8" would fit well with the seat raised



Product: Davis & Sanford EXPLORERV (amazon.com/dp/B000V7AF8E)

Question: "Is this tripod better than the AmazonBasics 60-Inch Lightweight Tripod with Bag one?"

Ranked opinions: "However, if you are looking for a steady tripod, this product is not the product that you are looking for" (no, .295); "If you need a tripod for a camera or camcorder and are on a tight budget, this is the one for you." (yes, .901); "This would probably work as a door stop at a gas station, but for any camera or spotting scope work I'd rather just lean over the hood of my pickup." (no, .463)

Response: Yes (.863)

Actual answer: The 10 year warranty makes it much better and yes they do honor the warranty. I was sent a replacement when my failed.

Evaluation – open-ended examples

Product: Mommy's Helper Kid Keeper ([amazon.com/dp/B00081L2SU](https://www.amazon.com/dp/B00081L2SU))

Question: "I have a big two year old (30 lbs) who is very active and pretty strong. Will this harness fit him? Will there be any room to grow?"

Ranked opinions: "So if you have big babies, this may not fit very long."; "They fit my boys okay for now, but I was really hoping they would fit around their torso for longer."; "I have a very active almost three year old who is huge."

Actual answer: One of my two year olds is 36lbs and 36in tall. It fits him. I would like for there to be more room to grow, but it should fit for a while.



Product: : Thermos 16 Oz Stainless Steel ([amazon.com/dp/B00FKPGEB0](https://www.amazon.com/dp/B00FKPGEB0))

Question: "how many hours does it keep hot and cold ?"

Ranked opinions: "Does keep the coffee very hot for several hours."; "Keeps hot Beverages hot for a long time."; "I bought this to replace an aging one which was nearly identical to it on the outside, but which kept hot liquids hot for over 6 hours."; "Simple, sleek design, keeps the coffee hot for hours, and that's all I need."; "I tested it by placing boiling hot water in it and it did not keep it hot for 10 hrs."; "Overall, I found that it kept the water hot for about 3-4 hrs.";

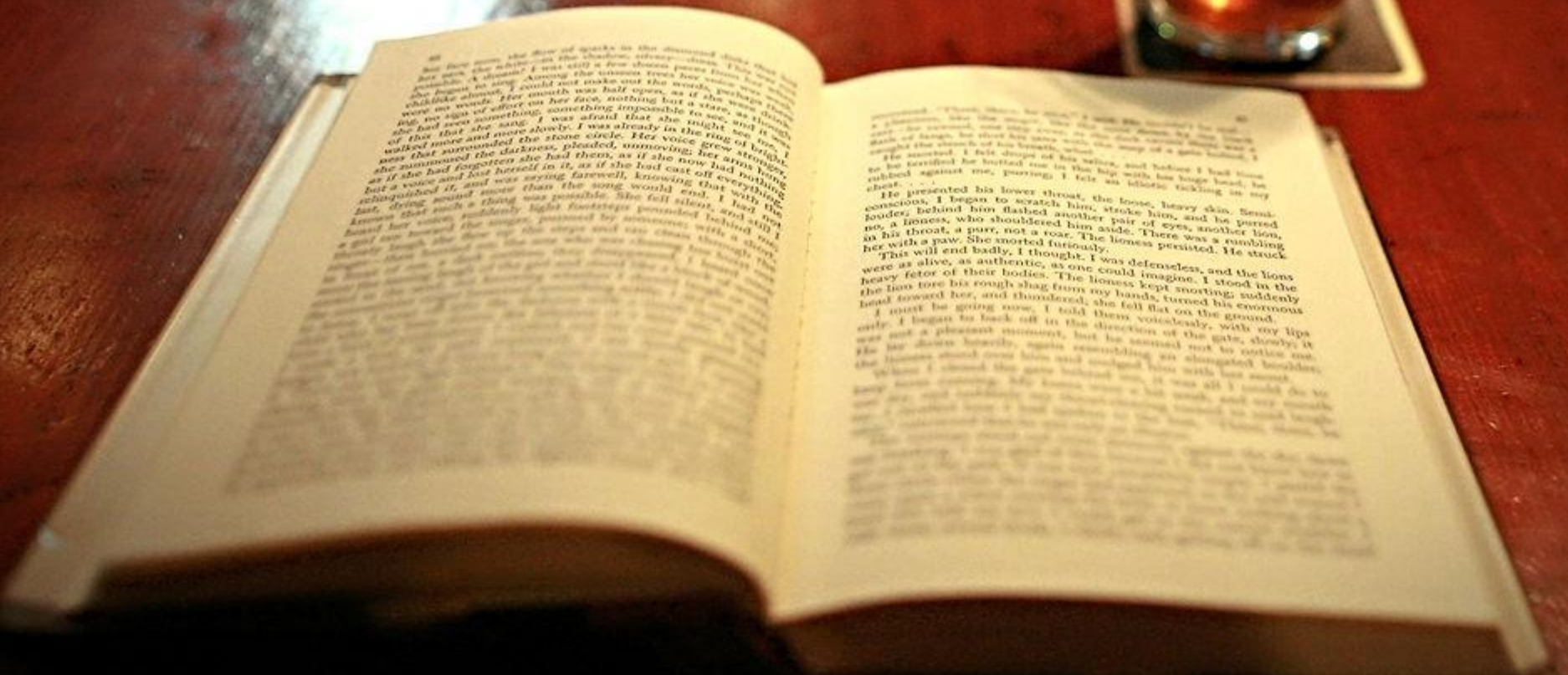
Actual answer: It doesn't, I returned the one I purchased.

Other ideas...

- In this work we just considered a single answer, but many questions have multiple (contradictory!) answers
 - Can also use other features like price, user expertise, product-specific language models, etc.
 - Can also consider the “match” between the questioner and the answerer/reviewer

(see our ICDM 2016 paper w/ Mengting Wan)

3. Generative models of reviews



Richer recommenders

Most “recommender systems” aim to build systems that are useful in a predictive capacity

- Can we predict a user’s rating of a product?
- Can we find reviews that a user would agree with?
- Can we predict the answer to a question, or find reviews that might help to answer it?

“Recommendation” then consists of intelligently making use of this black box

Can we make recommender systems more powerful by predicting not just how a user would **respond** to some stimulus (e.g. a rating), but by predicting what a user would **say**?

Richer recommenders

have:

$$f(u, i) : U \times I \rightarrow \{1, 2, 3, 4, 5\}$$

want:

$$f(u, i) : U \times I \rightarrow$$

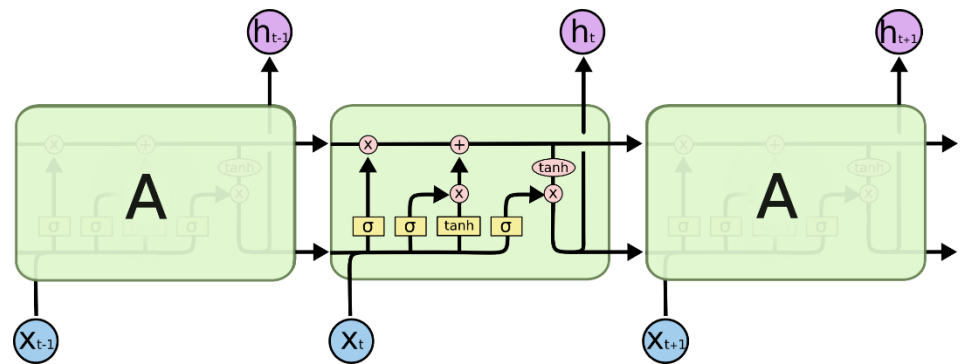
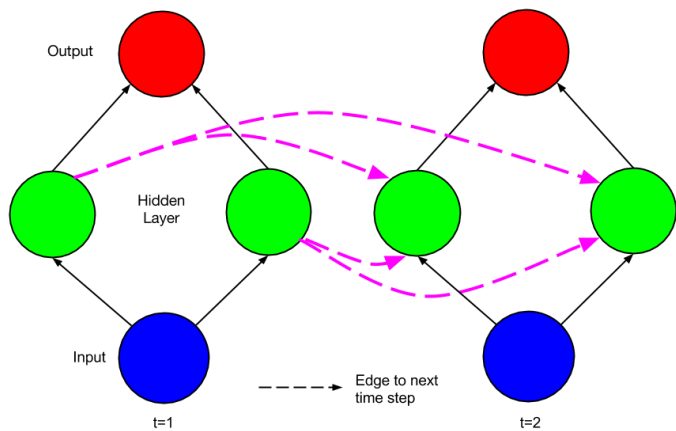
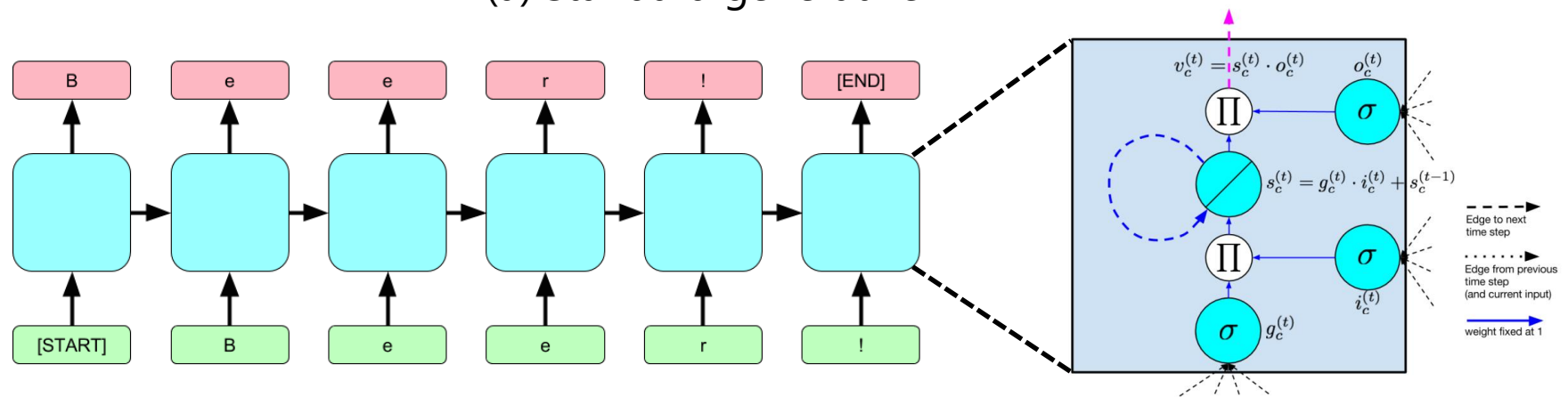
Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from 'Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to 'Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

- "Richer" recommendations, but can also be "reversed", and used for search

Generative models of text

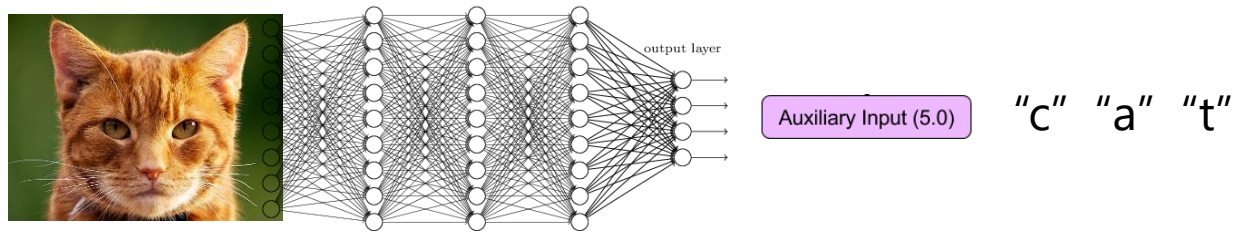
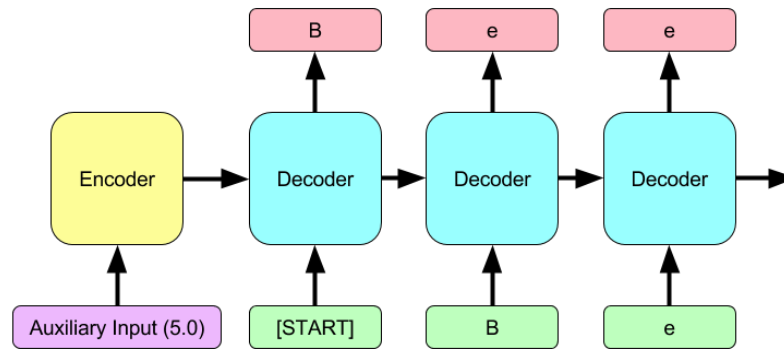
(a) Standard generative RNN



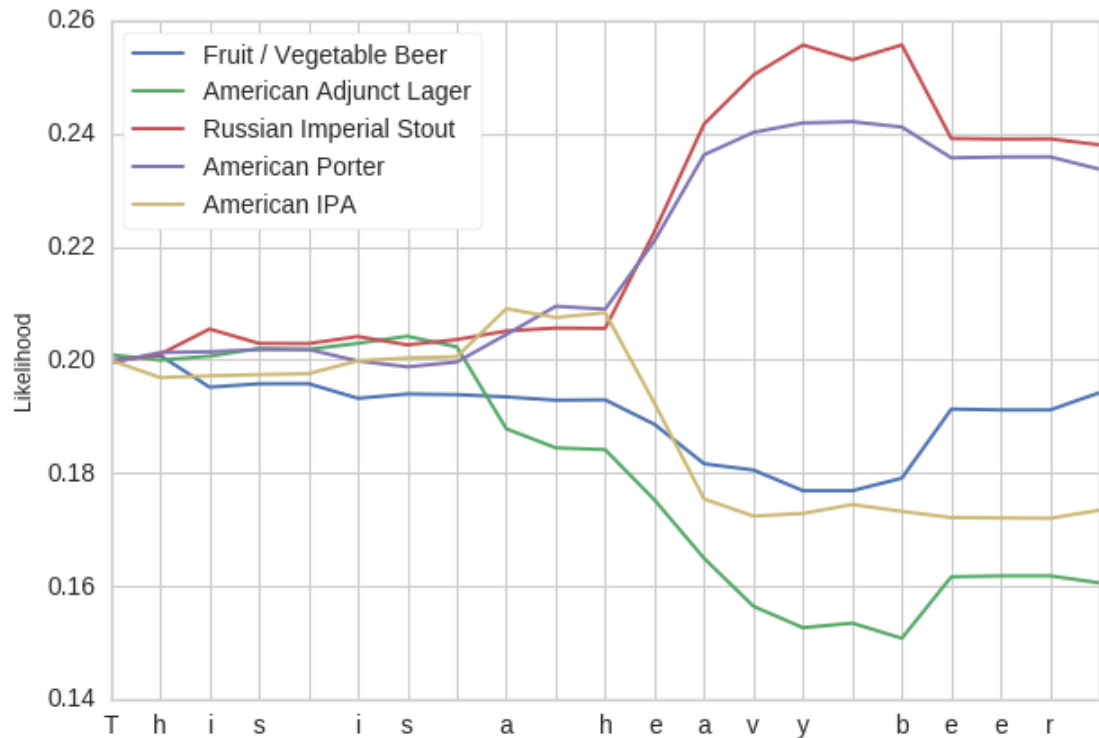
(from Christopher Olah)

Generative models of text

(b) Encoder-decoder RNN



Evaluation



Test set perplexity (median)

Unsupervised:	2.22
Rating:	2.07
Item:	2.17
User:	2.03
User-Item:	1.98

Generating reviews

Poured from 12oz bottle into half-liter Pilsner Urquell branded pilsner glass. **Appearance:** Pours a cloudy golden-orange color with a small, quickly dissipating white head that leaves a bit of lace behind. **Smell:** Smells HEAVILY of citrus. By heavily, I mean that this smells like kitchen cleaner with added wheat. **Taste:** Tastes heavily of citrus- lemon, lime, and orange with a hint of wheat at the end. Mouthfeel: Thin, with a bit too much carbonation. Refreshing. **Drinkability:** If I wanted lemonade, then I would have bought that.

Actual review

Poured from a 12oz bottle into a 16oz Samuel Adams Perfect Pint glass. **Appearance:** Very pale golden color with a thin, white head that leaves little lacing. **Smell:** Very mild and inoffensive aromas of citrus. **Taste:** Starts with the same tastes of the citrus and fruit flavors of orange and lemon and the orange taste is all there. There is a little bit of wheat that is pretty weak, but it is sort of harsh (in a good way) and ends with a slightly bitter aftertaste. Mouthfeel: Light body with a little alcohol burn. Finish is slightly dry with some lingering spice. **Drinkability:** A decent beer, but not great. I don't think I would rate this anytime soon as it says that there are other Belgian beers out there, but this is a good choice for a warm day when it's always available in the North Coast Brewing Company party.

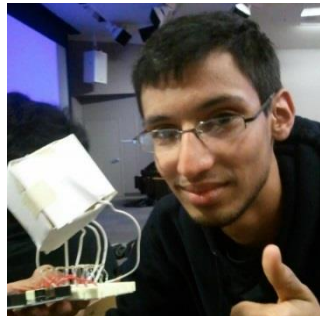
Synthetically generated review

Thanks!

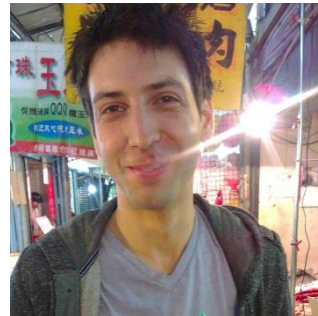
- *Addressing complex and subjective product-related queries with customer reviews.* McAuley, Yang, **WWW 2016**
- *Generative concatenative nets jointly learn to write and classify reviews,* Lipton, Vikram, McAuley **arXiv**



Alex Yang



Sharad Vikram



Zachary Lipton

Code and data: cseweb.ucsd.edu/~jmcauley/