CSE 255 – Lecture 6

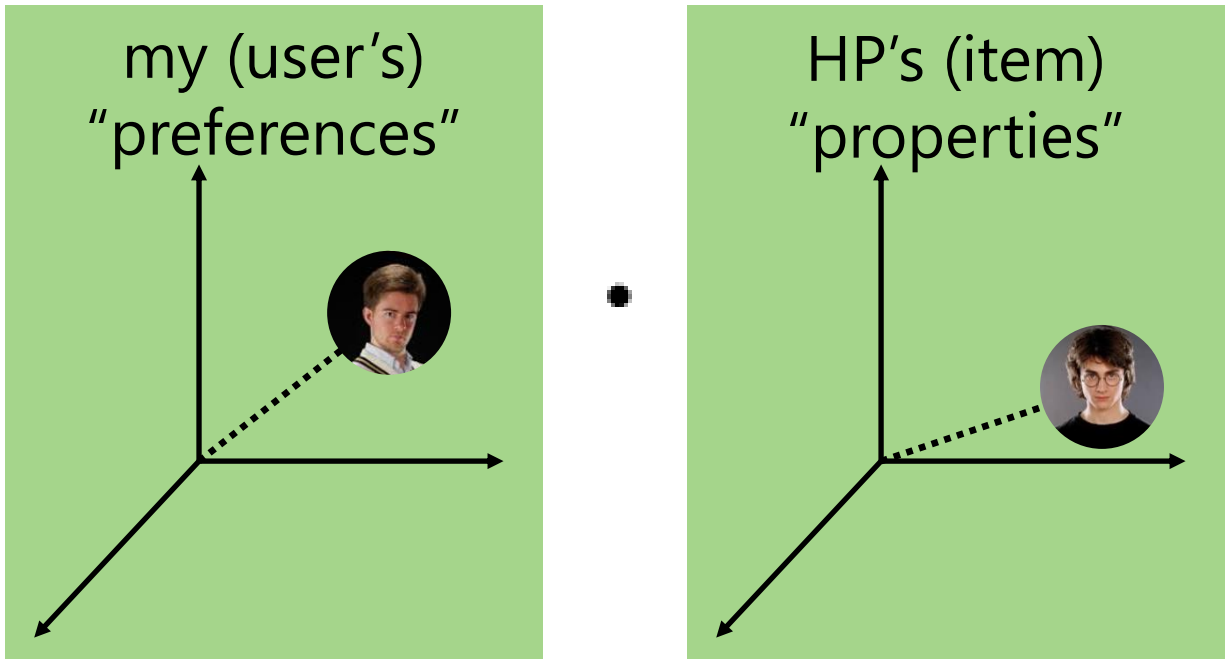
Data Mining and Predictive Analytics

Combining models of ratings and reviews

Ratings – Latent Factor Models

Two models we've seen so far:
1: Latent Factor Models (Lecture 5)

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



e.g. Koren & Bell (2011)

Text – Latent Dirichlet Allocation

Two models we've seen so far:
2: Topic models (Today!)

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

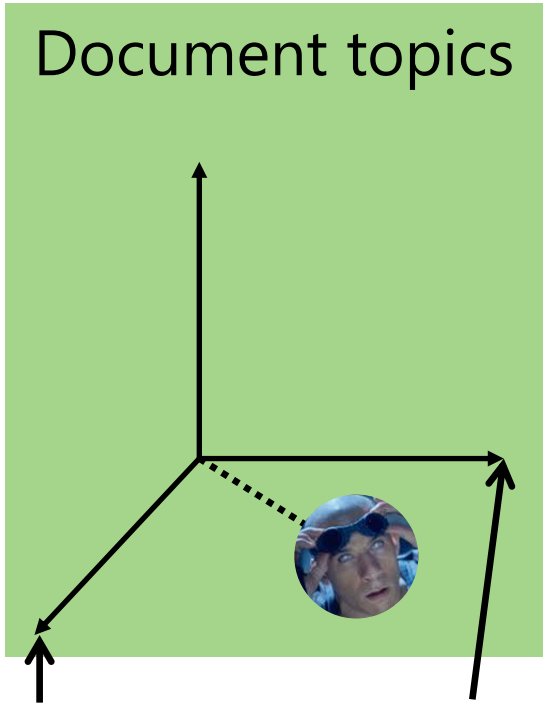
This review is from: **The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)**

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



Blei & McAuliffe (2007)

Low-dimensional representations

- Both of these models try to summarize **complex data** into **low-dimensional** representations
- If both of these models are based on the same principle (project high-dimensional data into low-dimensional spaces), can we combine them?
- In other words, can we come up with low-dimensional representations that capture the **common structure** present in both types of data simultaneously?

Why combine ratings and text?

Reason 1 (modeling):
it takes **lots of ratings** to estimate high-dimensional models of users and items – we might get away with **fewer** reviews

Reason 2 (understanding):
standard rating models have no **interpretations** – text might help us **explain** opinion dimensions

Combining ratings and reviews

The parameters of a "standard" recommender system

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

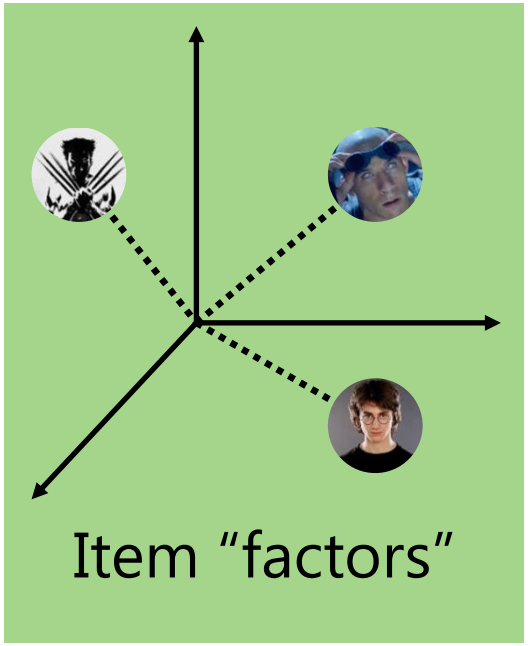
user/item — offset — user/item bias — latent factors

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

Combining ratings and reviews



transform

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

Our approach:

find topics in reviews that **inform us** about opinions

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

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

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

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

Step 2:
identify topics that "explain" the ratings

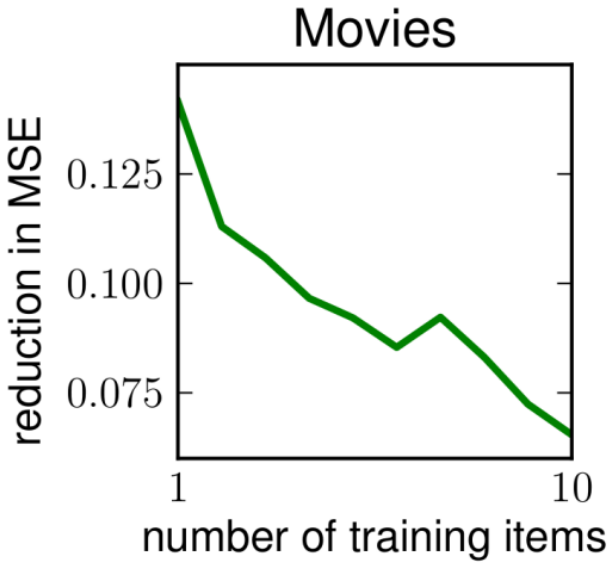
Outcomes – rating prediction

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:



Outcomes – interpretation

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

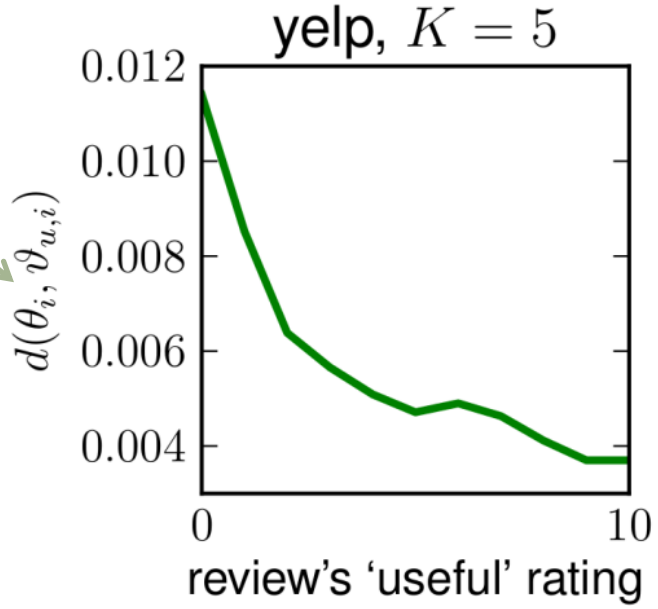
drums	strings	wind	mics	software
cartridge	guitar	reeds	mic	software
sticks	violin	harmonica	microphone	interface
strings	strap	cream	stand	midi
snare	neck	reed	mics	windows
stylus	capo	harp	wireless	drivers
cymbals	tune	fog	microphones	inputs
mute	guitars	mouthpiece	condenser	usb
heads	picks	bruce	battery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

Outcomes – usefulness prediction

What makes a review useful?

"Useful" reviews discuss topics in proportion to their importance

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



Questions?