

Joke Recommendations

...

Alejandro Alvarez
Brenda Palma

Agenda

1. Objective
2. Dataset
3. Models
 - a. Content-based filtering
 - b. Collaborative filtering
 - c. Ensemble
4. User Stories
5. Takeaways

Objective

Objective

In a nutshell...

To create a recommendation system for users to find new and hilarious jokes.

Objective

Using data of users, jokes, and ratings, the goal is to have personalized recommendations of new jokes (jokes from the same set that the user has not seen).

Data

Jester Dataset for Recommender Systems

- 100 jokes
- 73,421 users
- 4.1 million ratings
- Collected between April 1999 - May 2003
- Ratings ranging from -10 to 10

Models

Models

We explored 3 approaches:

1. Content-based Filtering:

- Based on **characteristics of items** and a record of the user's preferences
- Best suited when there is sufficient information on the **items**

2. Collaborative Filtering:

- Based on **past behaviour of users**¹
- Finds an association between the users and the items

3. Ensemble: **weighted rating** using outputs from 1 & 2

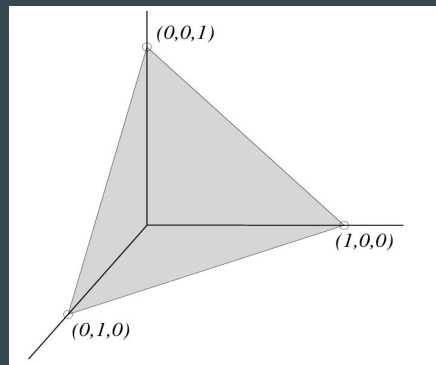
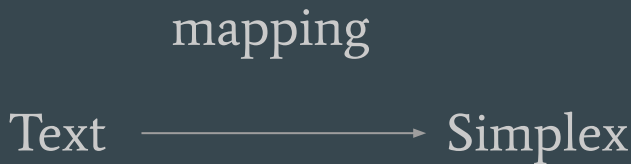
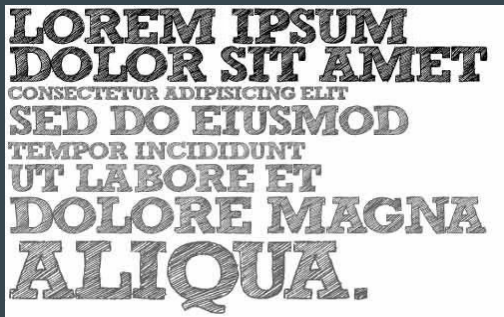
¹ Assumption: people who liked an item in the past will like it in the future

Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave to **some** of these **jokes**:

1. Send jokes to an embedding space (topic modeling)


“What do you call a fish with no eyes? A fsh” $\rightarrow [0.3, 0.5, 0.2]$



Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave **to some** of these **jokes**:

- Based on the topics (embeddings) of each of the jokes that a user rated, build the user's preferences

Joke (text)	Joke (embedding)	User ratings		User's preferences
Why is the obtuse triangle always upset? Because he's never right	[0.1, 0.9, 0.0]	7		$(0.7 \times 0.1 + 0.3 \times 0.4 + -0.2 \times 0.3) / 3 =$ 0.04333
Why don't scientists trust atoms? Because they make up everything.	[0.4, 0.1, 0.5]	3		$(0.7 \times 0.1 + 0.3 \times 0.4 + -0.2 \times 0.3) / 3 =$ 0.18
How do you drown a hipster? Throw him in the mainstream.	[0.3, 0.6, 0.1]	-2		$(0.7 \times 0.1 + 0.3 \times 0.4 + -0.2 \times 0.3) / 3 =$ 0.04333

Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave **to some** of these **jokes**:

3. Frame our problem as a regression problem
 - a. $\underbrace{\text{Rating for joke } j \text{ given by user } i}_{\text{User} \times \text{Joke space}} \sim \text{Joke } j \text{ topics} \& \text{ User } i \text{ preferences}$
 - b. Predict the rating that **user i** would give to unseen **jokes**
 - c. **Recommend the top-rated unseen jokes** (based on the predicted ratings)

Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave **to some** of these **jokes**:

3. Frame our problem as a regression problem
 - a. $\underbrace{\text{Rating for joke } j \text{ given by user } i}_{\text{User} \times \text{Joke space}} \sim \text{Joke } j \text{ topics} \& \text{ User } i \text{ preferences}$
 - b. Predict the rating that **user i** would give to unseen **jokes**
 - c. **Recommend the top-rated unseen jokes** (based on the predicted ratings)

Individual
user
information

Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave **to some** of these **jokes**:

3. Frame our problem as a regression problem

a. Rating for **joke j** given by **user i** ~ Joke j topics & User i preferences

User \times Joke space

b. Predict the rating that **user i** would give to unseen **jokes**

c. **Recommend the top-rated unseen jokes** (based on the predicted ratings)

Individual
user
information

Content-related
information

Content-based Filtering

Given **100 jokes (text)**, and the **ratings** each of the 70K+ users gave to **some** of these **jokes**:

3. Frame our problem as a regression problem

a. Rating for **joke j** given by **user i** ~ **Joke j topics & User i preferences**

User \times Joke space

b. Predict the rating that **user i** would give to unseen **jokes**

c. **Recommend the top-rated unseen jokes** (based on the predicted ratings)

a. k. a. content-based
recommendation

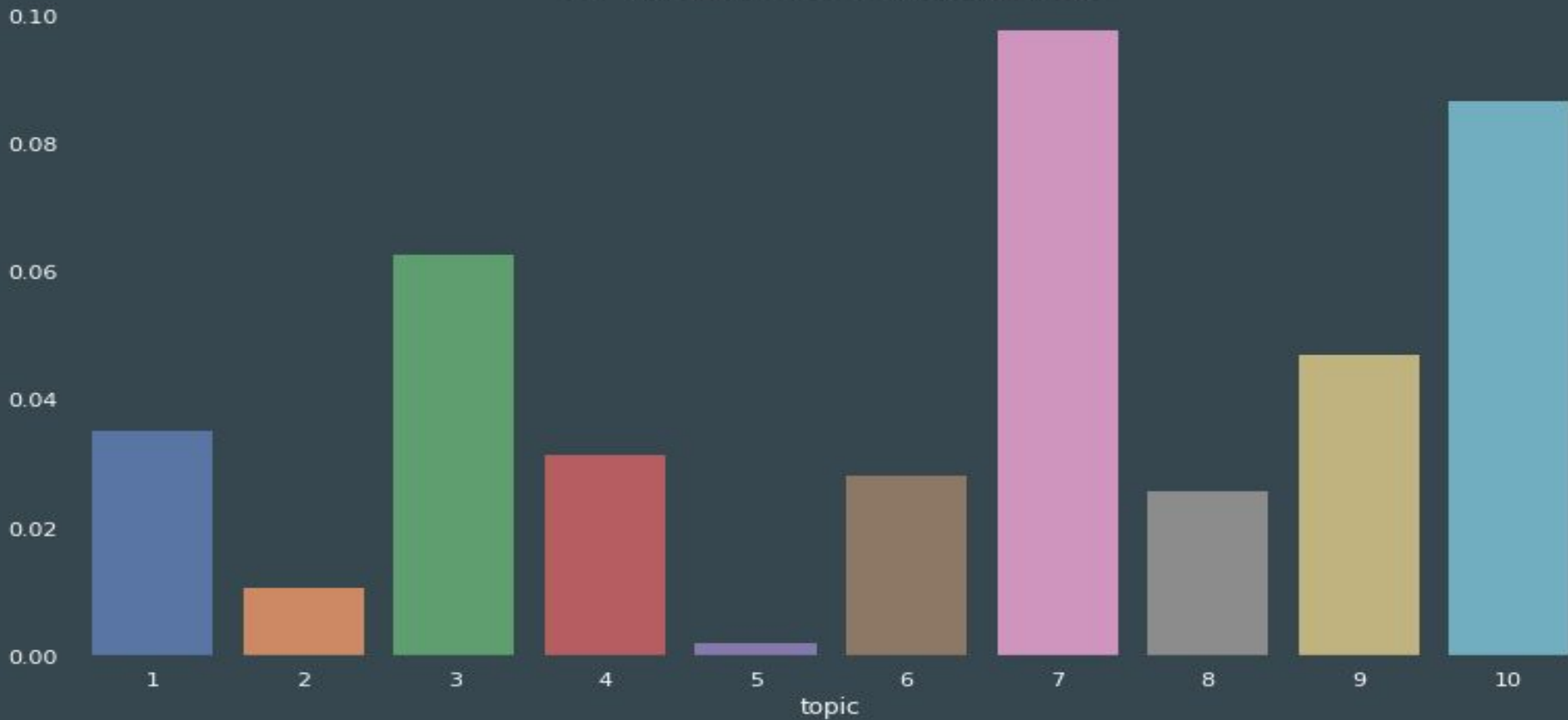
Individual
user
information

Content-related
information

User 4493

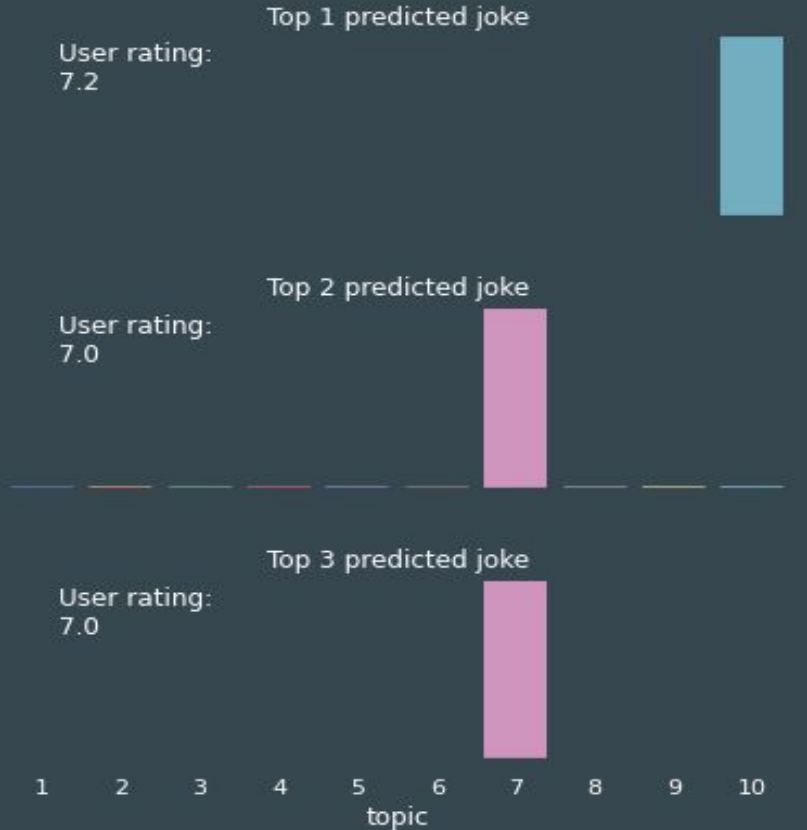
Content-based filtering

User 4493 preferences distribution per topic



User 4493

Jokes' topic distribution (top 3 rated by User 4493)



Content-based Filtering

- d. (Boring) regression problem details:
 - i. Topic modeling done with Latent Dirichlet Allocation model
 - ii. Users' preferences based only on training split (to avoid leakage)
 - iii. Histogram Gradient Boosting (🌲 🌳 🌴 ⚡ ▽ 📊)
 - ↪ Decision Trees 🌳
 - ↪ Ensembles 🌲 🌳 🌴 🌲 🌳 🌴 🌲 🌳 🌴
 - ↪ Boosting ⚡
 - ↪ Gradient Boosting ▽
 - ↪ Gradient binning (continuous gradient → discrete gradient) 📊
 - iv. Automatic hyperparameter optimization (using Optuna)

Collaborative Filtering

Given a matrix structure where each row represents a **user**, each column represents a **joke**, and values are **ratings** given by users to jokes,

1. Apply **Singular Value Decomposition**: find latent factors that describe the jokes using matrix factorization

$$A = USV^T$$

Matrix U: singular matrix (n users, r 'concepts')

Matrix S: diagonal matrix (shows strength of each 'concept')

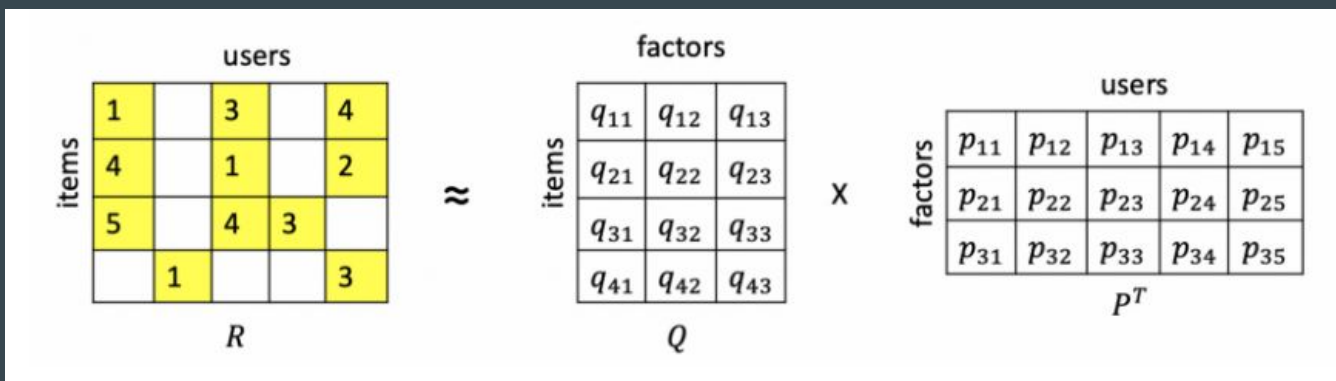
Matrix V: singular matrix (m jokes, r 'concepts')

a. k. a. singular values

→ Keep top k 'concepts'

Collaborative Filtering

2. Predict ratings of unseen jokes(using projections into concept space)



3. Recommendations will be the top rated items

Ensemble

Predict ratings by **combining** the outputs from the Content-based and Collaborative filtering models.

We chose to use a linear regression model with L2 regularization:

$$r_i = w_{cb} * r_{cbi} + w_{cf} * r_{cfi}$$

The diagram shows the equation $r_i = w_{cb} * r_{cbi} + w_{cf} * r_{cfi}$ with three annotations: a pink circle around r_i , a yellow bracket under $w_{cb} * r_{cbi}$, and a cyan bracket under $w_{cf} * r_{cfi}$. Below the equation, three labels are aligned with these annotations: 'Weighted rating' under the pink circle, 'Content-based Filtering term' under the yellow bracket, and 'Collaborative filtering term' under the cyan bracket.

Weighted
rating

Content-based
Filtering term

Collaborative
filtering term

Model Comparison

Algorithm	MAE
Content-based Filtering	3.47
Collaborative Filtering	3.33
Ensemble	3.25 ★

*All models were evaluated using the same test set

User Stories

User 4493

Top rated

A man piloting a hot air balloon discovers he has wandered off course and is lost. He descends and locates a man on the ground. He shouts "Excuse me, **can you tell me where I am?**" **The man says:** "Yes, you're in a hot air balloon, about 30 feet above this field." "You must work in IT," says the balloonist. "Yes," **replies the man.** "**How did you know?**" "Well," says the balloonist, "what you told me is technically correct, but of no use to anyone." The man says, "You must work in management." "I do," replies the balloonist, "how did you know?" "Well," says the man, "you don't know where you are, or where you're going, but you expect my immediate help. You're in the same position you were before we met, but now it's my **fault!**"

Content-based filtering

Recommended

Bill & Hillary are on a trip ... Bill pulls into a service station ... The attendant runs out ... Hillary realizes it's an old boyfriend ... **Bill ... says, 'Now aren't you glad you married me and not him ?** You could've been the wife of a grease monkey !' ... **Hillary replied,** 'No, Bill. If I would have married him, you'd be pumping gas and he would be the **President!**

User 4493

Top rated

A group of managers were given the assignment to measure the height of a flagpole. They go out to the flagpole with ladders and tape **measures**, and they're falling off the ladders, dropping the tape measures. An **engineer** comes along and sees what they're trying to do, walks over, pulls the flagpole out of the ground, lays it flat, measures it from end to end, gives the measurement to one of the managers and walks away. After the engineer has gone, one manager turns to another and laughs. "Isn't that just like an engineer, we're looking for the height and he gives us the length."

Collaborative filtering

Recommended

An **engineer**, a physicist and a mathematician are sleeping in a room. There is a fire in the room. The engineer wakes up, sees the fire, picks up the bucket of water and douses the fire and goes back to sleep. Again there is fire in the room. This time, the physicist wakes up, notices the bucket, fills it with water, **calculates** the optimal trajectory and douses the fire in minimum amount of water and goes back to sleep. Again there is fire. This time the mathematician wakes up. He looks at the fire, looks at the bucket and the water and exclaims, "A solution exists" and goes back to sleep.

User 4493

Top rated

One morning William burst into the living room and said, "I am getting married to the most beautiful girl in town. She lives a block away and her name is Susan." After dinner, William's dad took him aside. **"Son, your mother and I have been married 30 years. She's a wonderful wife but she has never offered much excitement in the bedroom, so I used to fool around with women a lot.** Susan is your half-sister, and you can't marry her." A year later William came home and announced, "Dianne said yes! We're getting married in June." Again his father insisted on another conversation. "Dianne is your half-sister too. I'm sorry." William was furious! He went to his mother with the news. His mother just shook her head. "Don't pay any attention to what he says, dear. He's not really your father."

Recommended

A guy goes into confession and says to the priest, "Father, I'm 80 years old, widower, with 11 grandchildren. Last night I met two beautiful flight attendants. **They took me home and I made love to both of them. Twice.**" The priest said: "Well, my **son**, when was the last time you were in confession?" "Never Father, I'm Jewish." "So then, why are you telling me?" "I'm telling everybody."

Takeaways

Takeaways

- Our content-based approach uses **explicit individual preferences** based on **observed characteristics of the content** the users rated to predict ratings for new content.
- Our collaborative-filtering approach uses **implicit characteristics** of the content (does not observe the jokes themselves) on which the user preferences are based.
- **Content-based** approach recommendations are similar in content (**common words**), while **collaborative filtering** are based on abstract **concepts**.
- Both approaches yield good and similar results in terms of rating predictions.
- Ensembling both approaches yielded the best results, since it emphasizes both individual preferences and collective user-preferences.

Questions?

Thank you

Thank you :)