Sokovia News
Hybrid News Recommender

Beatriz Leitão
Louis Ritz
Sachin Nair
Vasco Oliveira
Y Chi Cindy Lange

Agenda

1 **STATUS QUO & PROBLEM**

What is Sokovia News' current situation and the project goal?

2 **SYSTEM DEFINITION**

What are common recommender systems and how do they work?

3 **DATA SCIENCE PIPELINE**

What data did we use and how did we create our own recommender model?

4 **EVALUATION& OPTIMIZATION**

How does our model perform and how could it be improved?

5 **IMPLEMEN-TATION**

Which IT architecture would we use to deploy and scale up our model?

6 **BUSINESS CASE & ROI**

What would be the incremental revenue and return on investment?

Status Quo & Problem: We aim to increase customer engagement through using a recommendation system

300,000 readers
assumed user base

3.5 minutes
time spent on news sites

2-3 articles
read per visit by a user

40%
average churn rate

Current Challenges

- Lack of personalization
- Minimal user engagement
- Limited revenue growth
- Uncompetitive in the market

Definition: Recommender systems combine user info to offer personalized news recommendations

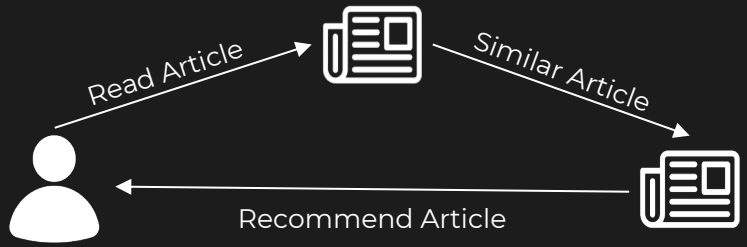
Most Known Use Cases

“Customers who bought this item also bought...”

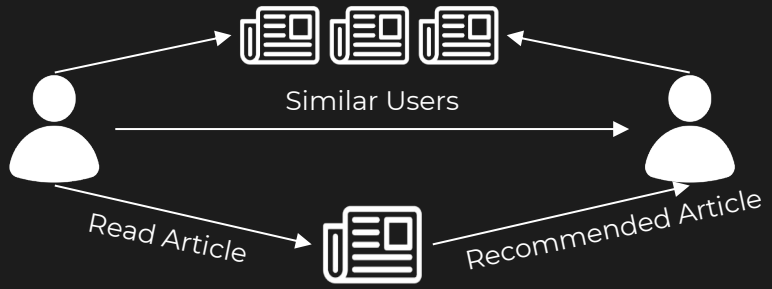
“Because you watched...”

“You might also like...”

Content-Based Filtering



Collaborative Filtering



Project Structure: Microsoft *MIND* “small” as a basis for developing the proof of concept recommender

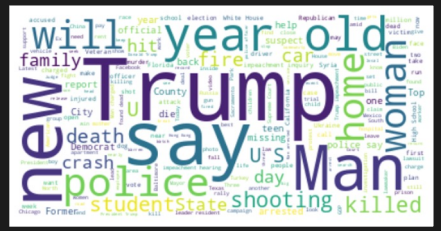
Datasets

50,000 news

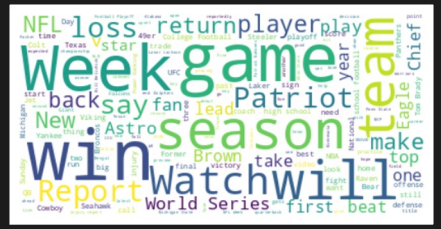
50,000 users

160,000 impressions

News
US News



Sports
NFL



Neutral Sentiment

Most Popular Hours

5:00 - 8:00 | 10:00 - 11:00

Data Science Pipeline (1/2): Data cleaning and NLP transformations used to extract relevant information

News

1. Removal of duplicate values
2. NLP text preprocessing
3. Content consolidation
4. Embeddings addition
5. Addition of release date

Category	Sub Category	Title	Abstract	Title Entities	Abstract Entities
Health	weightloss	50 Worst Habits For...	These seemingly...	[{"Label": "Adipose...	[{"Label": "Adipose...

Category	Sub Category	Title	Abstract	Content	Avg Vector	Date
health	weight loss	50 worst habits For...	these seemingly...	health weight...	[-0.97, -0.85,...	2019-11-1 12:02:12

Behaviors

1. Removal of duplicate values
2. Removal of unclicked articles
3. History & impressions
4. Average vector calculation

User ID	Timestamp	History	Impressions
U13740	11/11/2019 9:05:58 AM	N55189 N42782 N34694	N20678-1 N39317-0 N58114-0

User ID	Timestamp	History	Impressions	History & Impressions	Avg Vector
U13740	11/11/2019 9:05:58 AM	N55189 N42782 N34694	N20678	N55189 N42782 N34694 N20678	[-0.08, -0.84, 0.01.,...

Original Dataframe

Pre-Processed Dataframe

Data Science Pipeline (2/2): Three models provide relevant recommendations to new & existing users

RECOMMEN-DER TYPE	USE CASE	INPUT	MODEL STEPS	OUTPUT
Baseline Model *Random Recommender*	User with & without Article History	**User ID** @ Interaction Time T	Obtain 10 news articles at **random** in last two weeks	Top 10 **Random** Articles
Model 1 *Frequency & Category-Based*	User with & without Article History	**User Interests** (categories, optional)	Check for **most frequently read** news in selected **categories** in last two weeks	Top 10 **Most Read News** in Categories
Model 2 *Content-Based Filtering*	User with Article History	**Article History** @ Interaction Time T	Compare with **all news released** in last two weeks	Top 10 **Most Similar News**
Model 3 *Collaborative Filtering + Content-Based Filtering*	User with Article History	**Article History** @ Interaction Time T	Compare with articles read by **other users** in last two weeks (user-2-user) → Get 5 **most similar users** → Get **reduced news pool** of most similar users (excl. already read) → Compare **articles read** with **reduced news pool**	Top 10 **Most Similar News** from Top 5 **Most Similar Users**

Model Evaluation: Our best-performing model achieves +14x improvement on the current baseline

Baseline *Random Recommender*		**Model 3** *Collaborative + Content-Based Filtering*	
0.03%	**x14** →	**0.42%**	**Precision@10** Relevant Predictions
0.03%	**x14** →	**0.44%**	**Recall@10** Impressions Predicted
0.08%	**x17** →	**1.34%**	**Mean Reciprocal Rank** Ranking Quality
0.03%	**x16** →	**0.48%**	**nDCG@10** General Relevance

Performance Optimization: Model combination and re-weighting of metrics for future performance boost

Part 1: Model Combination

"People also liked"

Model 3
Collaborative Filtering + Content-Based Filtering

0.3 0.4 **0.5** 0.6 0.7

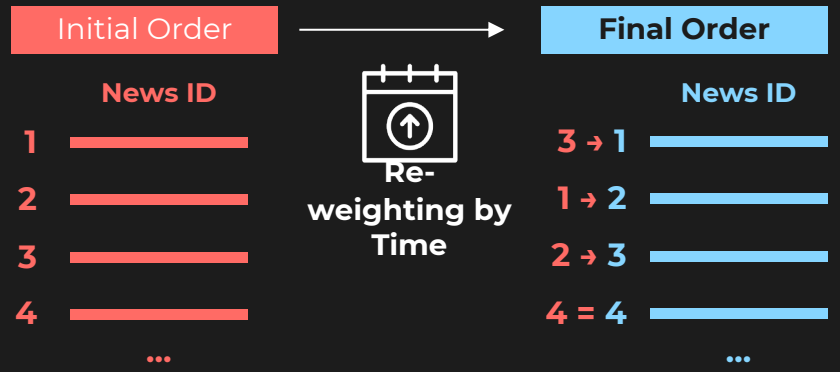
X

"Unseen News"

Model 2
Content-Based Filtering

0.3 0.4 **0.5** 0.6 0.7

Part 2: Recency Consideration

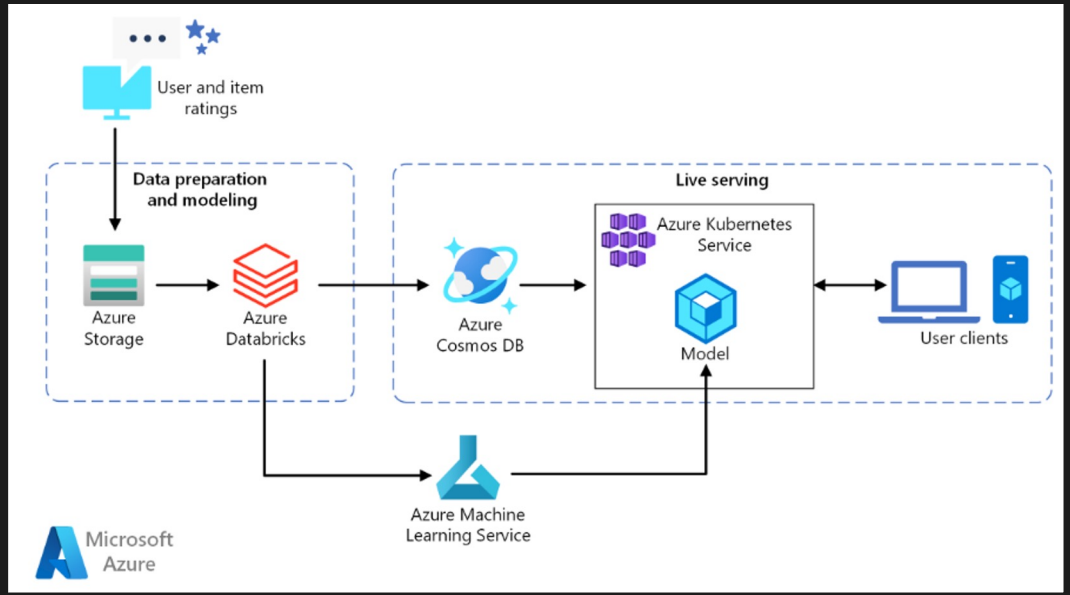


Improved Top 10 News Recommendations

Deployment, Legal & IP: Streaming data architecture required, compliance with regulations to be ensured

Architecture

Microsoft Azure Cloud Environment
Big Data Architecture for News Recommendations



Legal & Intellectual Property

- *Dataset Licensing*
- *Intellectual Property*
- *Data Privacy*
- *Ethical Considerations*
- *Attributions & Citations*

Business Case & ROI: 220% ROI, exponential growth expected with yearly revenue increases at steady cost

Incremental Revenue

1. Ad Revenue Growth
2. Premium Subscriptions
3. Reduced Churn

+ 20 % Yearly Readers

First-Year

EUR 200,000

Implementation Costs

1. Data Storage
2. Recommendation System
3. Production and Scaling

First-Year

EUR 60,000

First-Year

220%

ROI

Additional Features to Further Drive Growth

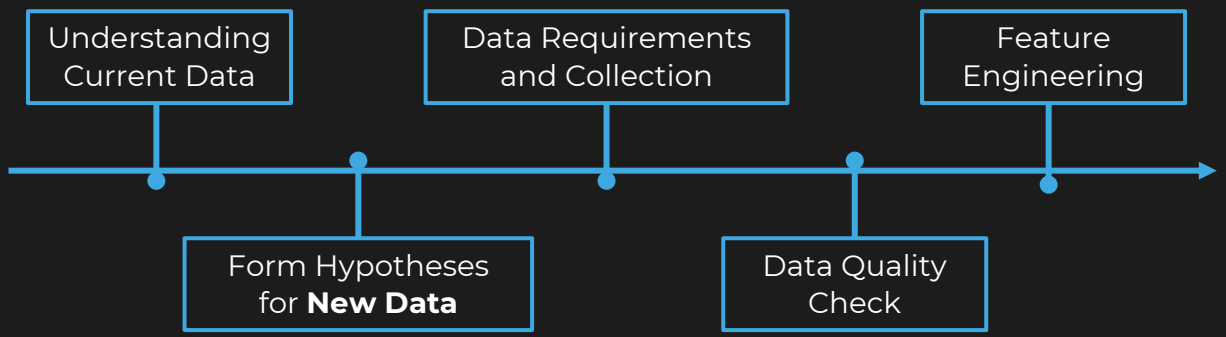
Split Testing

Personalized Email Recommendations

Push Notifications

Future Outlook & Next Steps: After PoC, focus on developing own dataset and improving model

In-House Dataset Methodology



New Data

User Ratings & Comments

User Demographics

External & Trendy Events

Reinvest incremental income to **iteratively improve data quality** and **model performance.**

Thank You!

Do you have any questions?

Beatriz Leitão

Louis Ritz

Sachin Nair

Vasco Oliveira

Y Chi Cindy Lange