0 -

In just a few words, this project investigates a variety of recommender systems and
approaches for making Amazon software recommendations in an offline setting
these are just a small preview of what is to come

0 -

I will follow this structure, explaining and discussing non-trivialities and decisions for each of the implemented systems, and then gradually build a table of results

1 -

The review data contains several columns, but we are only interested in this subset
further, a collection of metadata complements the product IDs, describing each product

1 -

From a histogram we quickly see that the distribution of item rating counts follows a long-tailed distribution: here, 40 products (of 800) correspond to >20% of ratings

The violin plots summarize different statistics of the reviews.
First, we have the distribution of mean rating per reviewer
Second, we have the same per item
This gives some qualitative intuition about the data, how users give ratings, and how items receive them
The median for user means is rather high (4) and displays greater variance, while more agreement occurs per item

2 -

off-the-shelf surprise implementations
5 folds used (fewer wrongly encourages smaller neighbour counts, and provides unrealistic bad setting
training time
kNN gives simple intuitive + interpretable
SVD robust, efficient, high coverage but less interpretable
NB error analysis: few ratings, actual k = 0 for most,
when RR != 0 then rating count is larger

3 -

two different vector spaces: TF-IDF and Word2Vec

I will talk about por

World Wide Web and Web crawling

The Web is part of the Internet (protocol to share info on Internet (: massive network of communicating computers))

Web as a Graph (Directed)

V is the set of pages;

E is the set of hyperlinks (inlinks & outlinks);

- ~10-20 hyperlinks per Web page on average

many more graphs can be defined…

Power law distribution of hyperlinks:

Very few pages have the most hyperlinks;

Vast majority of pages have very few hyperlinks

bow-tie structure

tendrils + in -> **SCC core** -> **out+tendrils**

reachable from in, can reach core, strongly connected component,

can be reached from core, can reach out

Largest human artifact ever created

Ss/Vs of data:

Volume, substantially large-scale & increasing

Variety, human/non-human generated, new forms of digital data, sensors etc

Veracity, uncertainty, noise-prone, bias, typos, fake news, corrupt source

Velocity, dynamic, time-series (near real-time or periodic)

Situation: context of data measurement

Scale: data with limited range vs wide range

Semantics: circa 80% of data is unstructured, extracting is a challenge

Sequence: velocity, dynamic

Crawlers

- Selection policy, re-visit policy, politeness policy
- discovery queue/refreshing queue
- METRICS: coverage/freshness (quality), PERFORMANCE: throughput

Geographically Distributed Web Crawling

Sentiment or thematic web crawling

Introduction to Recommender Systems with Collaborative Filtering (CF)

Long-tail Distribution of ratings among items

- Small fraction of items are rated frequently (popular items);
- Low frequency items:
- Larger profit;
- More difficult to recommend;
- Potential bias.

CF Idea:

users shared the same interests in the past → will have similar tastes in the future

no need of knowledge about items

user-based

- identify peers with similar preferences
- for each non-rated item, make prediction based on peers
- assumes preferences are stable and consistent over time

Pearson accounts for differences in interpreting the ratings scale.

Evaluation of Recommender Systems

Text representation and content-based Recommender Systems

Content-based and knowledge-based Recommender Systems

Hybrid Recommender Systems

Explanations in Recommender Systems

Collective intelligence and crowdsourcing

Industry talks