Text Classification with Deep Learning

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute

Speaker: D. L. Lazarenko

Supervisor: Ph.D. in Physico-mathematical Science, docent A. Maltsev

Kyiv, 2018

Introduction

Aim of this thesis is building an effective model which have high accuracy and an appropriate speed for classification of advertisements at the e-commerce platform Jiji.ng.

Object of study is advertisements at e-commerce platform

Subject of study is classification model for advertisements:

Relevance of the problem

- e-commerce sales are quickly increasing
- large online e-commerce websites serve millions of users' requests per day
- processes of registrations and purchases as much convenient and fast as possible
- users have to make a choice from more than hundred categories
- automatic category prediction is very important in terms of saving moderators' time and as a result, decreasing the number of necessary moderators to process them

Structure of the data files

lvl2	titles	descriptions
29	Clean Toyota Camry 2008 Silver	Fairly used Toyota 08 Camry with no problems V4 engine fabric seats and interior
25	Look Unique	Nice, quality, adorable,unique dress available now, whatsapp me

Content-based

- title
- description
- images
- ...

Algorithms: Latent Dirichlet allocation(LDA), relevance feedback(RF), TF-IDF

Collaborative Filtering

Оставим в векторах только те элементы, для которых нам известны значения в обоих векторах, т.е. оставим только те продукты, которые оценили оба пользователя, или только тех пользователей, которые оба оценили данный продукт. В результате нам просто нужно определить, насколько похожи два вектора вещественных чисел.

Collaborative Filtering

Подсчитаем коэффициент корреляции:

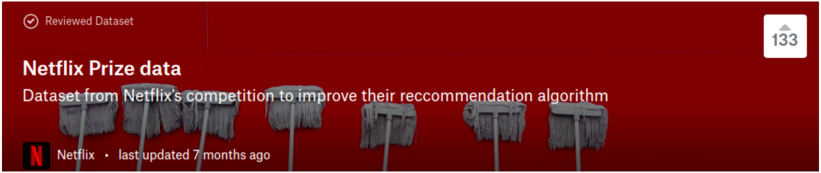
$$w_{ij} = \frac{\sum_a (r_{ai} - \overline{r_i})(r_{aj} - \overline{r_i})}{\sqrt{\sum_a (r_{ai} - \overline{r_i})}\sqrt{\sum_a (r_{aj} - \overline{r_j})}}$$

где,

$\overline{r_i}$ - средний рейтинг, выставленный пользователем i

Matrix-factorization

On September 21, 2009, the grand prize of US 1,000,000 was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06



Reviewed Dataset

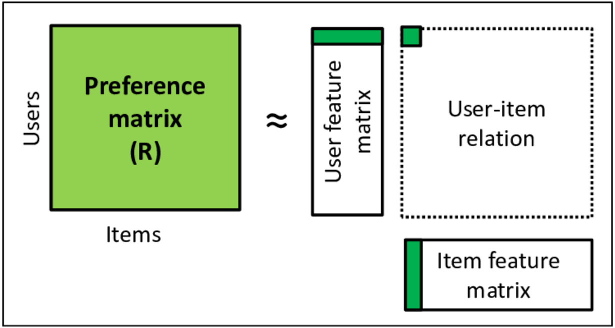
Netflix Prize data

Dataset from Netflix's competition to improve their reccommendation algorithm

Netflix • last updated 7 months ago

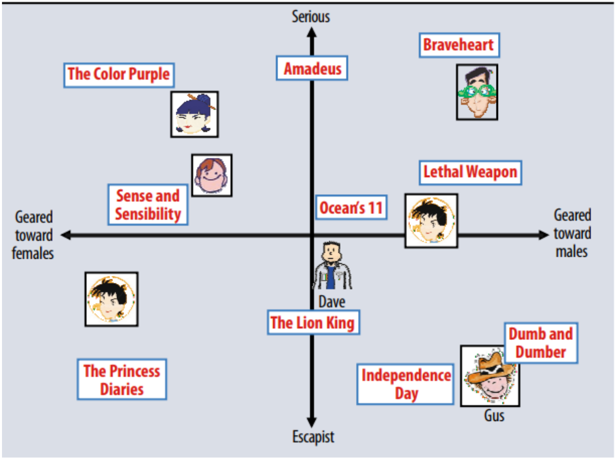
Matrix-factorization

Latent Factor Models



Algorithms: Alternating least squares(ALS), Stochastic gradient descent(SGD)

Matrix-factorization

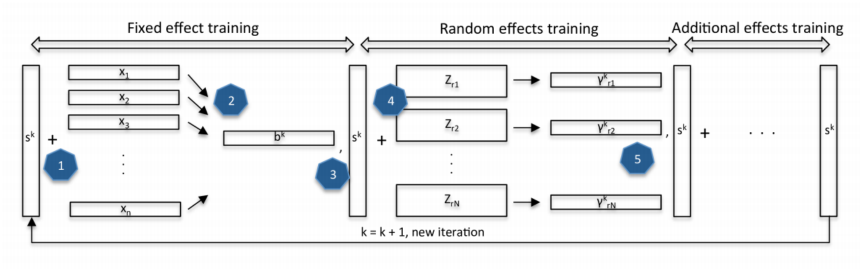


Photon-ml

Photon-ml

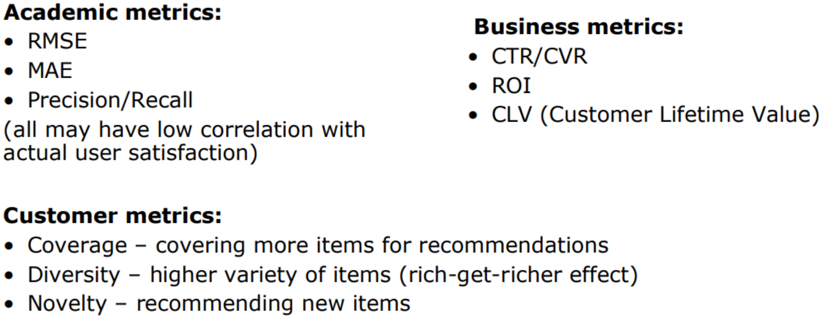
- Generalized Linear Model (GLM)
- Generalized Additive Model (GAM)
- Generalized Additive Mixed-Effect Model(GAME)
- GLMix(Generalized Linear Mixed) $=$ GLM $+$ per-user model $+$ per-item model

Photon-ml



The experiments were conducted on a cluster consisting of 135 nodes managed by Apache YARN 3. Each node has 24 Intel Xeon(R) CPU E5-2640 processors with 6 cores at 2.50GHz each, and every node has 250GB memory.

Evaluation



Academic metrics:

- RMSE
- MAE
- Precision/Recall

(all may have low correlation with actual user satisfaction)

Business metrics:

- CTR/CVR
- ROI
- CLV (Customer Lifetime Value)

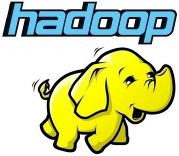
Customer metrics:

- Coverage – covering more items for recommendations
- Diversity – higher variety of items (rich-get-richer effect)
- Novelty – recommending new items

Проблемы

- Товары быстро продаются, не успев даже набрать хорошую историю по просмотрам и запросам контактов. Классические алгоритмы коллаборативной фильтрации устроены так, что объявления с короткой историей не попадают в рекомендации. Чаще рекомендуются долго живущие объявления, которые, как правило, представляют меньший интерес для покупателей.
- Проблемы холодного старта
- Как заставить это все быстро работать

Проблемы



Спасибо за внимание!