Machine Learning on Big Data

Ильнур Шугаепов

About

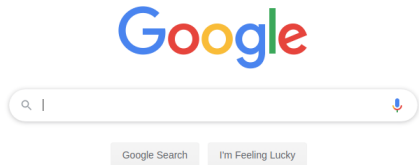
linkedin.com/in/ilnur-shugaepov

- 2018-till now —Senior ML Engineer at VK.com
- 2015-2017 — АУ РАН
- 2014-2016 — Computer Science Center
- 2011-2015 — ITMO Uiversity

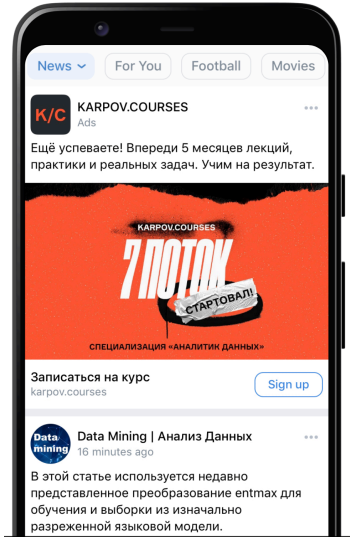
Table of Contents

Web Search

Google



Click Through Rate Prediction



Criteo

Movies Recommendation[1]

Netflix

[1] Yehuda Koren. "The bellkor solution to the netflix grand prize". In: *Netflix prize documentation* 81.2009 (2009), pp. 1–10.

Ключевые особенности

- Большое количество данных и признаков ($> 10^6$)
- Сильно разряженные данные
- Категориальные признаки большой размерности

Подвыборка?

TLDR: Заметно падает качество[2]

Нужны инструменты для работы с большими данными и эффективные алгоритмы ML

[2] Xinran He et al. “Practical lessons from predicting clicks on ads at facebook”. In: *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM. 2014, pp. 1–9.

Table of Contents

Part I

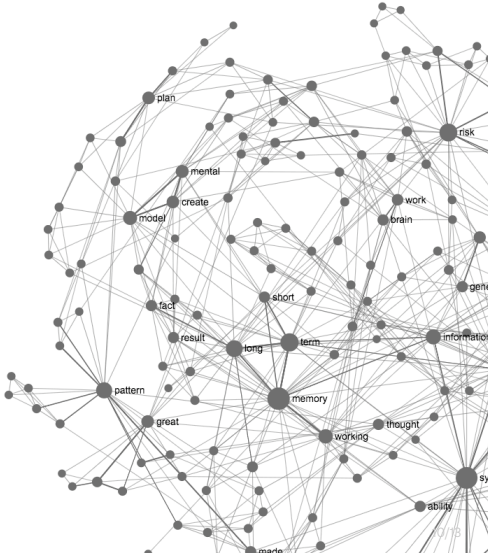
Tools and Systems for Big Data Storage and Processing

1. Hadoop and MapReduce
2. Apache Spark
3. Spark SQL

Part II

Large Scale Machine Learning

1. Distributed ML Introduction
2. Categorical Features in Large Scale ML
3. Gradient Boosting Decision Tree
4. Hyperparameters Optimization
5. DNN Compression and Acceleration
6. RecSys/Nearest Neighbors Search



Part III

Online Controlled Experiments

1. How to conduct AB Tests
2. Results Analysis
3. Heterogeneous Treatment Effect

Практики

Pull Requests with Jupyter notebooks

Домашние задания

Kaggle

1. CTR-prediction (Criteo Ads)
2. Recommendations (MovieLens + TMDB)

Вопросы?

Ильнур Шугаепов