Yandex
Crowdsourcing Practice for Efficient Data Labeling: Aggregation, Incremental Relabeling, and Pricing

Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zerminova, Daria Baidakova
Introduction

Olga Megorskaya,
Head Yandex.Toloka

Yandex.Toloka is a service of Swiss company Yandex Services AG
Majority of ML-based solutions require training data labelled by human
...at a large scale
Crowdsourcing

Specific way to design a business process

A big task → Cloud of performers → Result
Crowdsourcing: require less from performer, more – from manager

Expert approach: rely on an expertise of a particular performer:
- expensive
- unmeasurable
- hard to scale

Crowdsourcing approach:
- measurable
- scalable
- manageable
XX century – style management

- Routine tasks
- Regular work
- No ability to choose tasks

It can be different

- Flexibility to choose from hundreds of tasks
- No requirements in regularity
- Switch to another task when bored
Crowdsourcing can provide maximal flexibility to performers if:

- On a platform side, efficient tools for quality management are available for requester.

- Requester knows how to build smart crowdsourcing pipelines resistant to single performer’s mistakes.
Crowdsourcing applications: examples

<table>
<thead>
<tr>
<th>Task type</th>
<th>Used in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information assessment</td>
<td>Ranking of search results</td>
</tr>
<tr>
<td>Content categorization</td>
<td>Text and media moderation, data cleaning and filtering</td>
</tr>
<tr>
<td>Content annotation</td>
<td>Metadata tagging</td>
</tr>
<tr>
<td>Pairwise comparison</td>
<td>Offline evaluation, media duplication check</td>
</tr>
<tr>
<td>Object segmentation, including 3D</td>
<td>Image recognition for self-driving car</td>
</tr>
<tr>
<td>Audio and video transcription</td>
<td>Speech recognition for voice-controlled virtual assistant</td>
</tr>
<tr>
<td>Spatial crowdsourcing</td>
<td>Verify business information and office hours</td>
</tr>
</tbody>
</table>
Example: binary classification

Is this cat white?

Yes
No
Example: multi classification

If you are a gourmand, I can recommend you the "Real French restaurant", located in the historic cellar, with elements of antique design and quite interesting cuisine. The restaurant is small, but very cozy and romantic. The restaurant is very suitable for romance and even for business meetings.
Example: multi classification with ordered labels
Examples: pairwise comparison
Examples: transcription with textual answers
Examples: object segmentation
Examples: spatial crowdsourcing
A crowdsourcing platform: two-sided market

Performers

Requesters
Crowdsourcing platforms: examples

› Amazon Mechanical Turk
› Yandex.Toloka
› Microworkers
› Gigwalk
› ClickWorker
› CloudFactory
› CrowdSource
› DefinedCrowd
› …
Pros of crowdsourcing platforms

24/7

Variety of skilled performers

Vast region coverage

Ongoing processes
Crowdsourcing growth: Yandex experience

Different projects in Yandex.Toloka

* An extrapolation based on the first 3 months of 2020
Crowdsourcing growth: Yandex experience

Active performers in Yandex.Toloka

2014: 9K
2015: 120K
2016: 270K
2017: 570K
2018: 1.1M
2019: 2.2M
2020: 6.4M*

* An extrapolation based on the first 3 months of 2020
Everyday on Yandex.Toloka

500+ different projects  37K+ performers  13M+ tasks
# Yandex.Toloka: real-life cases

<table>
<thead>
<tr>
<th>Task</th>
<th>Details</th>
<th>Time</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Side-by-side object comparison</td>
<td>1,000 tasks</td>
<td>Done in 10 min</td>
<td>$2.4</td>
</tr>
<tr>
<td>Object classification</td>
<td>1,000 photos</td>
<td>Done in 15 min</td>
<td>$1.2</td>
</tr>
<tr>
<td>Object segmentation</td>
<td>about 1,000 objects in 100 photos</td>
<td>Done in 6 h</td>
<td>$3.6</td>
</tr>
<tr>
<td>Phrase generation for a chatbot</td>
<td>500 phrases for the same topic</td>
<td>Done in 15 min</td>
<td>$1</td>
</tr>
<tr>
<td>Audio transcription</td>
<td>100 recordings 25 minute long</td>
<td>Done in 20 min</td>
<td>$6</td>
</tr>
<tr>
<td>Video ranking</td>
<td>10,000 videos</td>
<td>Done in 2 h</td>
<td>$10</td>
</tr>
</tbody>
</table>
Tutorial overview
Why this tutorial?

Practice
Tutorial schedule

Part I: 20 min
- Main Components

Part II: 10 min
- Introduction to Crowd Platform

Introduction: 15 min

Part III: 15 min
- Brainstorming pipeline

Part IV: 60 min
- Set & Run Projects

Break: 30 min

Part VI: 20 min
- Set & Run Projects cont.

Part V: 25 min
- Theory on Aggregation

Part VII: 10 min
- Results & Conclusions
Thank you!
Questions?

Olga Megorskaya
Head of Crowdsourcing Department

omegorskaya@yandex-team.ru

https://research.yandex.com/tutorials/crowd/sigmod-2020