




The Hunt for Prohibited Content

Predict which ads contain illicit content

\$25,000 · 285 teams · 3 years ago

[Overview](#)[Data](#)[Discussion](#)[Leaderboard](#)[Rules](#)[Team](#)[My Submissions](#)[Late Submission](#)

Competition Data

[Edit](#) sample_submission.cs...**avito_train.zip** 736.83 MB[Download](#) APatK.py avito_ProhibitedCont... avito_test.zip [avito_train.zip](#)

Data Description

Data for this competition consists mainly of Russian text. All files are encoded in UTF-8 and are in tab separated format (.tsv). To help you transform Russian text into a set of features we have prepared introductory code, where we recommend which modules in Python to use.

Also note that uncompressed training and test data together take ~4GB of space.

Training and Test data sets consist of individual ads that have either been blocked for illicit content or that have never been blocked. All ads that participate in this competition have already been closed.

Using External Data

External data is allowed in this competition with approval. To gain approval for a data set/source, please post your request on this [forum thread](#).

File descriptions

- **avio_train.zip** - training dataset. Contains both ads descriptions and labels. Ads for training has been sampled from Dec 2013 - Mar 2014.

- **avio_test.zip** - testing dataset. Contains only ads description. Ads for testing has been sampled from Apr 2014.
- **sampleSubmission.tsv**- a sample submission file in the correct format produced by our introductory code.

Data fields



- **itemid** - unique identifier of each ad
- **category** - 1st level category of an ad
- **subcategory** - 2nd level category of an ad
- **title** - name of the Ad
- **description** - Full text with ad description
- **attrs** - additional parameters of the ad in JSON format. Each parameter has its name and its value. E.g if you are selling bmw z1 car, you would have {"car brand":"bmw", "car model":"z1"}
- **price** - final price of ad in Russian rubles
- **is_proved** - Additional data column that is available in the training only. Not to be used as a direct modeling attribute. This flag is provided only for blocked ads. It indicates that ad was blocked by an experienced moderator. Because humans do make errors it is likely (though not proven) that ads blocked by experienced moderator who should contain larger % of actually illicit content.
- **is_blocked** - Boolean **target variable**. This is the column to predict.
- **phones_cnt** - Number of contact phones that we found in ad description. Some sellers provide their contact phone numbers in ad description. If it was the case we would replace this phone number with @@PHONE@@ in description.
- **emails_cnt** - Number of emails that we found in ad description. Some sellers provide their emails in ad description. If it was the case we would replace this email with @@EMAIL@@ in description.
- **urls_cnt** - Number of urls that we found in ad description. Some sellers provide urls in ad description. If it was the case we would replace this email with @@URL@@ in description.
- **close_hours** - Available in train only! Number of hours as a real number how long ad was live on avito. The more hours it was live and was not blocked it is more likely that it does not contain illicit content and it was not missed by moderators by error.

To have more understanding about each field please see commented picture from Ad details page on Avito.ru

Все объявления в Москве ▸ **Транспорт** ▸ **Автомобили с пробегом** ▸ **BMW** ▸ **X1**

BMW X1, 2012

Размещено 19 мая в 16:25. ✎ ✖ Редактировать, закрыть, поднять объявление



Цена **1 200 000 руб.** [Онлайн калькулятор Каско](#)

Продавец **Владимир**
[Показать телефон](#)

Город Москва [показать на карте](#)

BMW X1 , 2012 г.
Пробег 30 000 - 34 999 км , 2.0 AT , бензин , полный привод , кроссовер , левый руль , цвет чёрный
2 литра, 184 л.с. Полный привод, 8 ступ. автомат, климат контроль, подогрев сидений, зимняя резина и так далее...

Не бита, не крашена.

Пробег 32 000км.

Машина обслуживается только у официального дилера.

Номер объявления: 335309783