**Cheetah Mobile Transfer Learning Data Documentation**

# Introduction

This document describes the data Cheetah Mobile used for transfer learning between users’ mobile app behavior and content reading behavior.

# Data Source

User actions on news app, User info from tools/system app

# Data Schema

Following is the data schema used, there are 2 sources of data, joined by aid (android device id):

1. User article reading behavior, columns start with prefix “article\_”, each row means 1 user clicked on 1 article
   1. Articles user read are classified into categories using text classification
   2. Article\_l1\_categories are coarse categories that the article belongs to
   3. Article\_l2\_categories are fine grained categories that the article belongs to (contains l1 categories)
   4. Article\_dwelltime is the time user spent reading the article
   5. Article\_lan is article language
   6. Other fields are self-explainable
2. User Info and mobile app behavior, columns start with prefix “user\_”
   1. Age is age group predicted
      1. 1 stands for 21-30
      2. 2 stands for 31-40
      3. 3 stands for 41-50
      4. 4 stands for 50+
   2. Gender is predicted, 1 stands for male, 2 stands for female, weight near 1 means higher probability to be female, near 0 means higher probability to be male, use weight >= 0.7 and weight <=0.3 for 90% confidence level
   3. User\_brand and user\_model is phone brand and model
   4. User\_all\_pkgs is all packages(apps) user installed
   5. User\_gp\_frequency is the corresponding google play category the apps belongs to. Frequency means the number of apps user installed that belongs to this category
   6. Gp\_keywords is the google place description keywords of the apps user installed

**Note**:

1. All user ids are anonymized with a PBKDF2 combination
2. All article categories are number coded

root

|-- user\_id: string (nullable = true)

|-- article\_contentid: string (nullable = true)

|-- article\_l1\_categories: array (nullable = true)

| |-- element: struct (containsNull = true)

| | |-- name: string (nullable = true)

| | |-- weight: double (nullable = true)

|-- article\_l2\_categories: array (nullable = true)

| |-- element: struct (containsNull = true)

| | |-- name: string (nullable = true)

| | |-- weight: double (nullable = true)

|-- article\_dwelltime: integer (nullable = true)

|-- article\_lan: string (nullable = true)

|-- article\_publisher: string (nullable = true)

|-- article\_title: string (nullable = true)

|-- article\_update\_time: long (nullable = true)

|-- dt: string (nullable = true)

|-- user\_age: integer (nullable = true)

|-- user\_gender: integer (nullable = true)

|-- user\_gender\_weight: double (nullable = true)

|-- user\_model: string (nullable = true)

|-- user\_brand: string (nullable = true)

|-- user\_country: string (nullable = true)

|-- user\_maxmind\_country\_iso\_code: string (nullable = true)

|-- user\_maxmind\_state\_iso\_code: string (nullable = true)

|-- user\_maxmind\_city: string (nullable = true)

|-- user\_all\_pkgs: array (nullable = true)

| |-- element: string (containsNull = true)

|-- user\_gp\_frequency: map (nullable = true)

| |-- key: string

| |-- value: integer (valueContainsNull = true)

|-- user\_gp\_keywords: map (nullable = true)

| |-- key: string

| |-- value: integer (valueContainsNull = true)