

Articles RecSys

Валентин Копылов

March 5, 2020

Content

Описание задачи

Основные проблемы

Модель

Признаки для модели

Лайфхаки

Articles RecSys

**Валентин
Копылов**

Описание
задачи

Основные
проблемы

Модель

Признаки для
модели

Лайфхаки

Описание задачи

Есть:

1. История кликов пользователей по статьям
2. Описание всей имеющихся статей

Необходимо:

Из списка `random_benchmark` отранжировать статьи в порядке релевантности

Основные проблемы

- ▶ Все пользователи видели не все статьи
- ▶ Много данных – проблема памяти
- ▶ Матрица взаимодействий - огромная и не хранится в памяти в dense формате
- ▶ Colab лагает

Модель LightFM

Гибридная рекомендательная система:

Сочетает в себе плюс **Memory-based** моделей и **Content-based** моделей

- ▶ Предоставляет класс для создания признаков
- ▶ `no_components` - latent dimensionality
- ▶ Много разных лоссов – работает `logloss`

Признаки для модели

1. Статьи

- ▶ tfidf для названия статьи (**lemmatized**)
- ▶ tfidf для самой статьи (**lemmatized**) + (в сокращенном виде)
- ▶ (Эмбединги картинок?)

2. Пользователи

- ▶ Количество статей, которые пользователь увидел
- ▶ Количество статей, на которые пользователь кликнул
- ▶ Rate кликнувших статей ко всем увиденным

1. Kaggle API
2. Скрипт для прокликивания Colab в браузере
3. Telegram Bot API

Описание
задачи

Основные
проблемы

Модель

Признаки для
модели

Лайфхаки

Главная проблема: Разные выдаваемые ресурсы!

Описание
задачи

Основные
проблемы

Модель

Признаки для
модели

Лайфхаки

Спасибо за внимание!