Отчет по домашней работе от DATA

Михаил Михайлов, Ельцов Даниил, Гавриленко Дмитрий

10 ноября 2020 г.

Содержание

Резюме

Был построен классификатор, который по количеству дипломов от школы по каждому предмету определяет ее профиль. Источник данных — датасет победителей и призеров олимпиад московских школ (источник).

На вход программы подается название школы, после чего строится вектор данных, содержащий информацию о количестве дипломов (диплом призера и победителя считаются равноценными), который прогоняется через классификатор.

Постановка задачи

Построить классификатор который определит к какому из профилей принадлежит школа:

- физико-математический,
- естественно-научный,
- обществоведчский,
- языковый,
- гуманитарный,
- многопрофильный/без профиля/не определенно.

Используемый датасет

Датасет взят с kaggle.com — источник.

Состоит из одной csv таблицы, содержащей около 70 тысяч строк, формата:
`Полное название школы, Краткое название школы, Тип олимпиады (ММО/Всеросс), Этап, Класс, Предмет, Год, ID`

Описание решения

Разметка предметов по профилям

- *Категория физико-математических профилей*: астрономия, физика, информатика, математика, информатика и информационно-коммуникационные технологии (ИКТ).
- *Категория естественно-научных профилей*: биология, географияя, химия, экология.
- *Категория обществоведчских профилей*: изобразительное искусство, история, мировая художественная культура (МХК), обществознание, право, экономика, искусство (МХК), бюджетная грамотность.
- *Категория языковых профилей*: французский язык, итальянский язык, китайский язык, английский язык, испанский язык, немецкий язык, латынь.
- *Категория гуманитарных профилей*: филология, русский язык, литература, лингвистика.
- *Категория других предметов*: технология, основы безопасности жизнедеятельности, физическая культура, робототехника, информационные технологии в профессиональной деятельности.

Обработка данных

Строится таблица:
`Полное имя школы -- вектор: (Количество дипломов категории 1, ..., Количество дипломов категории 6)`

Построение классификатора

Для классификатора зафиксированы константы:

- `THRESHOLD` — порог числа дипломов, необходимого для того чтобы можно было определить профиль.
- `MAIN_SUBJECT_THRESHOLD` — минимальный процент дипломов, необходимый для выделения главного профиля.
- `MIDDLE_SUBJECT_THRESHOLD` — минимальный процент дипломов, необходимый для установки многопрофильности школы.
- `EPS` — порог разницы процентов между дипломами, необходимый для установки многопрофильности школы.

Если значение какой-то координаты в пересчете на проценты превосходит `MAIN_SUBJECT_THRESHOLD`, то профиль школы соответствует профилю этой координаты.

Если значение каких-то двух координат в пересчете на превосходит `MIDDLE_SUBJECT_THRESHOLD`, то школа имеет несколько профилей.

Результаты

Обработка данных — Михаил Михайлов
Построение классификатора — Даниил Ельцов
Отчет — Михаил Михайлов

Код