Определение категории видеозаписи на основе текстовых метаданных

Остапец Андрей Александрович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: aostapec@mail.ru

Автоматическое определение категорий видеозаписей, размещенных в интернете, становится все более важной задачей с ростом числа различных видеохостингов (YouTube, Vimeo и т.п.). Информация о категориях видеозаписей может быть использована для улучшения решений задачи поиска релевантных запросам пользователей видеозаписей и задачи показа таргетированной рекламы пользователям видеохостингов.

В данной работе рассматривается задача классификации видеозаписей, размещенных на видеохостинге YouTube. Каждое видео на видеохостинге принадлежит одной из 15 категорий (примеры категорий: Наука и Образование, Музыка, Новости и Политика). Необходимо правильно определять релевантную категорию видео используя метаинформацию о видеозаписи, такую как заголовок, описание, продолжительность видеозаписи и т.д.. В качестве функционала качества рассматривалась точность классификации (Accuracy). Поставленная задача решалась в рамках международного конкурса [1].

Обучающая выборка состояла из более чем 240,000 видеозаписей. Для тренировочных данных организаторами конкурса были выбраны видеозаписи, загруженные на хостинг в период с 01 января 2013 года по 31 декабря 2014 года. Тестовая выборка включала в себя более 115,000 записей. В тестовые данные входили видео, загруженные на YouTube в период с 01 января 2015 года по 30 апреля 2015 года. Каждая видеозапись описывалась 15 предикторами различных типов и категорией записи в качестве целевой переменной. Видеозаписи выбирались из загружаемых на видеохостинг случайным образом. Вследствие этого в выборках были представлены видеозаписи, загруженные из различных стран мира. В результате, среди текстовых предикторов встречалось множество различных языков (английский, немецкий, русский, арабский и т.д.).

Для работы алгоритма использовались четыре поля:

- заголовок видеозаписи
- описание видеозаписи

- набор тем, напрямую связанных с видеозаписью
- набор тем, релевантных данной видеозаписи

В двух последних полях каждая тема была закодирована строкой (хэшом). Истинные значения тем не раскрывались.

В описываемом алгоритме предварительная обработка данных включала в себя последовательное применение следующих шагов:

1. Удаление html-тегов.
2. Удаление стоп-слов (только английских).
3. Приведение всех слов к нижнему регистру.

После предварительной обработки данных для каждого объекта генерируется один числовой вектор $x \in \mathbb{R}^d$. Для этих целей использовался следующий подход: текстовая информация о видеозаписи представлялась в виде «мешка слов» (bag of words), а затем осуществлялось TF-IDF преобразование. Для формирования вектора использовались униграммы, биграммы и триграммы. После этого классификация полученных векторов осуществлялась с помощью библиотеки Liblinear [2]. Итоговое решение является линейной комбинацией нескольких линейных моделей, обученных на разных подвыборках тренировочной выборки.

Данная комбинация повысила точность классификации на скрытой тестовой выборке до 73.9%, что позволило автору в составе команды Московского Государственного Университета победить в международном конкурсе [1]. В комбинации с другими решениями участников команды данное решение достигало точности в 75.4%. Команды, занявшие второе и третье место, достигли точности классификации в 73.9% и в 73.7% соответственно.

Литература

1. Страница конкурса «Data Science Game 2015»: `http://www.datasciencegame.com/`
2. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin Liblinear: A library for large linear classification // The Journal of Machine Learning Research, 2008, 9:1871–1874