УДК 004.85

В.С. Мальчиц, А.Н. Гетман

ОБРАБОТКА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ И ПРИМЕНЕНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ ДЛЯ РЕАЛИЗАЦИИ КЛАССИФИКАТОРА НОВОСТЕЙ

Рассматривается способ применения методов классического машинного обучения с учителем для решения задачи, связанной с реализацией классификатора новостей. Для реализации классификатора подобран набор данных, предварительно обработанный при помощи кластеризации и сэмплирования в целях большей точности определения. Обучение классификатора построено на основе метода опорных векторов.

Ключевые слова: машинное обучение, классификация, определение категории «новости», сэмплирование, метод опорных векторов.

Введение

Алгоритмы автоматического обучения стали частью современного мира. Множество сервисов, приложений и прочих цифровых ресурсов работает благодаря новейшим алгоритмам искусственного интеллекта и машинного обучения. Главный принцип машинного обучения заключается в том, что решаемая задача выполняется не напрямую, а путем систематического обучения алгоритмов и систем, в результате чего их знания или качество работы возрастают по мере накопления опыта. В качестве примера можно привести систему фильтрации спама, поступающего на почту. Такая система адаптируется под каждого пользователя индивидуально, ведь одно и то же сообщение для одного человека может оказаться спамом, а для другого нет [1].

Машинное обучение

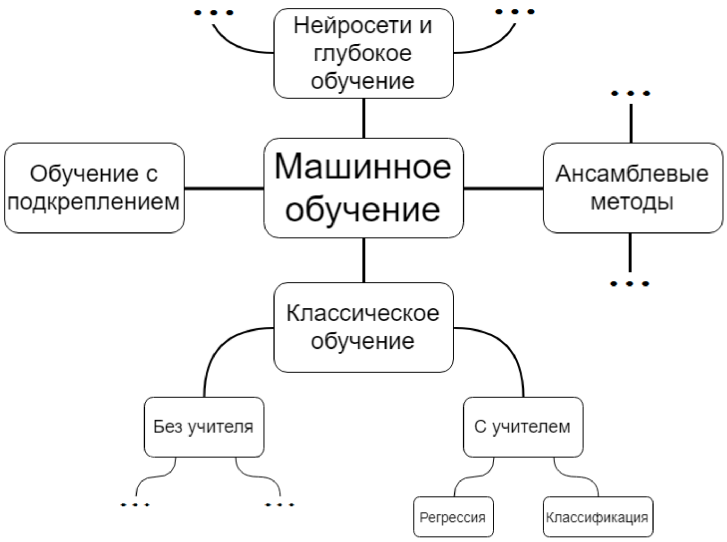

Рис. 1. Машинное обучение.

Машинное обучение – это систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы возрастают по мере накопления опыта [2]. В машинном обучении можно выделить несколько основных веток: классическое обучение, нейросети с глубоким обучением, обучение с подкреплением и ансамблевые методы. В данной статье рассматривается классическое обучение, которое, в свою очередь, можно разделить на обучение с учителем и обучение без учителя.

На рис. 1 показана иерархия, на которой видно направление машинного обучения для данной работы. Обучение с учителем и обучение без учителя происходит с использованием некоторых наборов данных (датасетов). Если обучающие данные размечены, то это обучение с учителем, если не размечены, то без учителя. Размеченными можно назвать такие данные, которые выступают в качестве конкретных примеров

для машины. В примере с фильтром спама такими данными может послужить набор сообщений, каждое из которых помечено спамом или нет. На размеченных данных машина обучается быстрее и точнее. В случае обучения без учителя модели приходится самой искать какие-либо закономерности [2]. В данной работе рассматривается более простой для понимания и реализации вариант – обучение с учителем.

Задачи обучения с учителем можно разбить на два типа: классификация и регрессия. В первом случае машина будет распределять данные по заранее известному признаку («спам» или «не спам»). Нами используется один из нескольких популярных алгоритмов обучения с учителем, решающих задачу классификации. Этот алгоритм носит название «Метод опорных векторов» и изложен в [3].

Суть метода опорных векторов заключается в следующем. Объекты представляются в виде векторов (точек) в n-мерном пространстве. Каждый из векторов может принадлежать только одному классу. Решение задачи данным методом сводится к ответу на вопрос: можно ли разделить точки гиперплоскостью размерности (n-1)? Также стоит учитывать, что гиперплоскостей может быть много, поэтому максимизация зазора между классами способствует более уверенной классификации [4].

Регрессия очень схожа с классификацией, за исключением того, что вместо категорий прогнозируется число.

Решение задачи требует большой работы с текстом, так как необходимо получить набор признаков, по которым классификатор сможет в дальнейшем определять категорию. Чтобы решить эту задачу, стоит обратиться к одному из направлений искусственного интеллекта – обработке естественного языка. Это общее направление искусственного интеллекта и математической лингвистики, которое изучает проблемы компьютерного анализа и синтеза естественных языков [5].

К популярным понятиям обработки естественного языка можно отнести токенизацию, лемматизацию, а также стемминг. Токенизация – процесс разбиения исходного текста на его отдельные составляющие (слова, знаки препинания). Лемматизация – процесс приведения словоформы к ее словарной форме (лемме). Стемминг – более простая форма лемматизации, в которой в основном просто удаляются суффиксы [5].

Описание задачи классификации новостей

Необходимо с помощью методов машинного обучения разработать и обучить модель, которая бы занималась сортировкой новостей почти без участия человека. Идея задачи взята из [6].

Модель требует на вход только готовый текст, который она «изучит» и покажет, к какой категории он относится. Это позволит значительно уменьшить объем работы, который возлагается на людей. Можно увеличить точность классификатора путем добавления данных или обучить ее на других данных для решения иных задач. Качество модели также зависит от набора данных, на которых она построена, поэтому необходимо тщательно подойти к выбору такого набора.

Решение задачи

В работе использовался набор данных, взятый с сайта www.kaggle.com [7]. Датасет, содержащий примерно 200 тысяч записей, каждая из которых, в свою очередь, имеет следующие свойства: категория, к которой относится новость (category), заголовок новости (headline), авторы новости (authors), ссылка на полный текст новости (link), краткое описание (short_description), дата публикации (date).

Для обучения модели необходимыми свойствами являются категория и текст новости. Так как в датасете не содержится полных текстов новостей, вместо них в работе используем краткое описание.

Для написания классификатора использовался высокоуровневый язык программирования общего назначения Python [8].

На рис. 2 показана диаграмма для изначального набора данных. Диаграмма построена с использованием сторонней библиотеки matplotlib. Matplotlib – это библиотека Python для построения 2D графиков, которая генерирует показатели качества публикаций в различных печатных форматах и интерактивных средах на разных платформах. По оси абсцисс расположены названия всех категорий, а по оси ординат – количество новостей, которые помечены той или иной категорией.

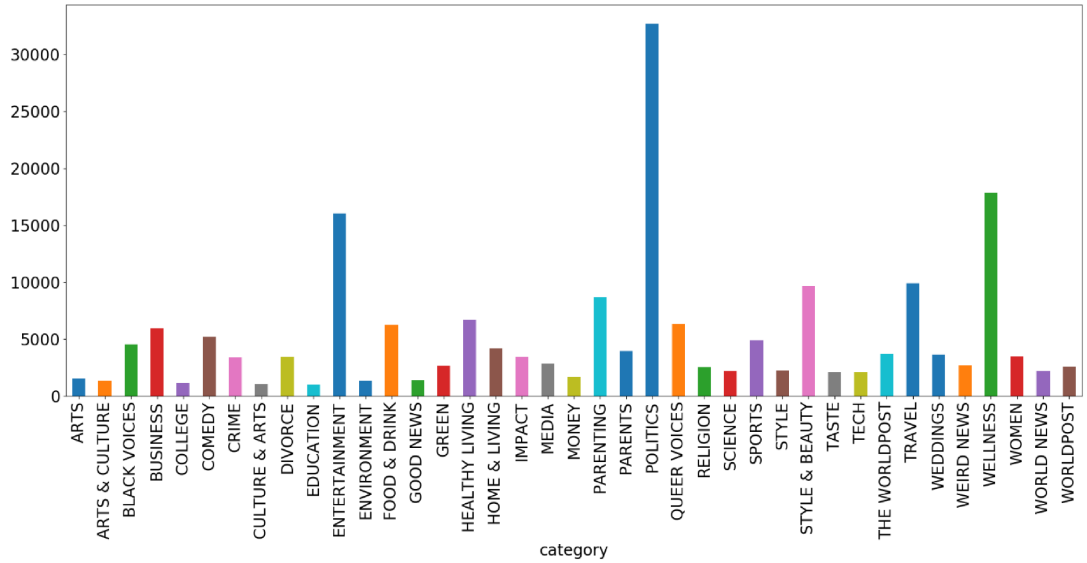

Рис. 2. Диаграмма с количеством новостей для каждой категории.

В данных были замечены очень похожие категории. Например, есть категории «ARTS & CULTURE», «CULTURE & ARTS» и «ARTS». Для того чтобы избавиться от лишних категорий, была проведена кластеризация. Кластеризация – задача группировки данных в отсутствие априорной информации о группах [1]. Если вернуться к трем вышеописанным категориям, то их можно объединить в одну общую категорию и назвать ее «CULTURE». В табл. 1 представлены все категории, которые были подвержены кластеризации.

Таблица 1

Кластеры

1	2	3
№	Категории	Общая категория (кластер)
1	BUSINESS	BUSINESS
2	MONEY	
3	ARTS	CULTURE
4	ARTS & CULTURE	
5	CULTURE & ARTS	
6	EDUCATION	EDUCATION
7	COLLEGE	
8	PARENTS	FAMILY
9	PARENTING	
10	DIVORCE	
11	WEDDINGS	
12	HEALTHY LIVING	HEALTH
13	WELLNESS	
14	STYLE	STYLE
15	STYLE & BEAUTY	
16	TASTE	TASTE
17	FOOD & DRINK	
18	TECH	TECH & SCIENCE
19	SCIENCE	
20	WEIRD NEWS	WEIRD NEWS
21	QUEER VOICES	

Продолжение табл. 1

1	2	3
22	WORLD NEWS	WORLD
23	GREEN	
24	GOOD NEWS	
25	ENVIRONMENT	
26	THE WORLDPOST	WORLDPOST
27	WORLDPOST	

После кластеризации данных количество категорий уменьшилось с 39 до 21. Это показано на рис. 3.

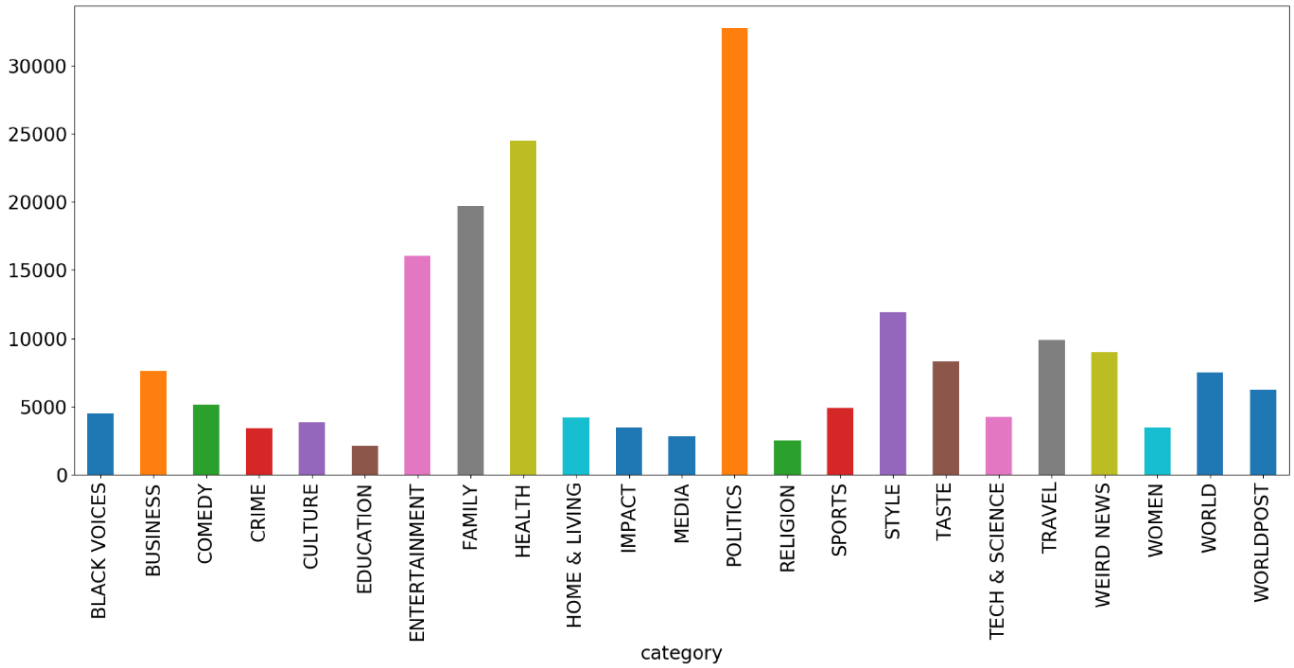

Рис. 3. Диаграмма с количеством новостей после кластеризации.

Также стоит отметить некоторые отличия в количестве новостей среди категорий. Есть категории, имеющие более 20 тыс. новостей, а есть такие, количество новостей которых не превышает и 5 тыс. Это говорит о том, что баланс данных сильно нарушен, что в свою очередь повлияет на точность модели.

Рассмотрим так называемую матрицу смежности. Она изображена на рис. 4, и ее суть заключается в следующем. По оси ординат отмечены все актуальные категории новостей, а по оси абсцисс – определенные моделью. На пересечении двух категорий находится число, которое показывает, сколько раз актуальная категория была определена как та, что находится по оси абсцисс.

Для необработанных данных видно, что качество определения новости моделью достаточно плохое. Об этом говорит то, что значения главной диагонали почти ничем не отличаются от остальных. В идеальном случае значения по главной диагонали должны принимать максимальные значения, а все остальные минимальные (нули). Такой результат будет говорить о 100% точности классификатора.

Проведем баланс данных с помощью так называемой выборки данных (сэмплирования). Суть сэмплирования заключается в том, что для выбранной категории берутся не все новости, а лишь часть, которая выбрана случайно и задается процентным соотношением. В данной работе процентные соотношения были подобраны вручную и приведены в табл. 2. То есть, например, значение 15 для категории «HEALTH» говорит о том, что из всех имеющихся новостей данной категории случайно отберется лишь 15%.

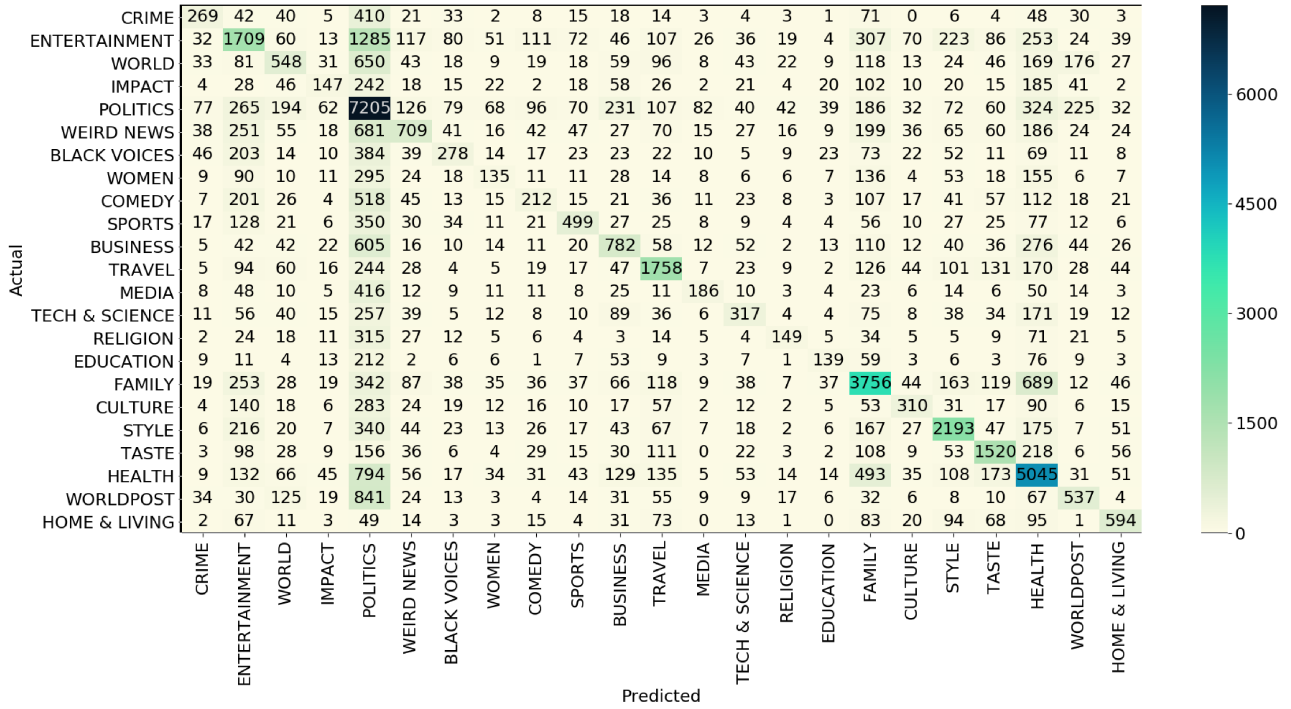

Рис. 4. Матрица смежности.

Таблица 2

Результат сэмплирования данных

№	Категория	Выборка от общего количества, в процентах
1	POLITICS	15
2	ENTERTAINMENT	30
3	HEALTH	15
4	FAMILY	20
5	STYLE	30

В данной работе сэмплирование было проведено при помощи функции sample(), которая включена в библиотеку pandas. Pandas – это библиотека с открытым исходным кодом, предоставляющая высокопроизводительные, простые в использовании структуры данных и инструменты анализа данных для языка программирования Python. После проведения сэмплирования количество новостей всех категорий стало близко к одному уровню (рис. 5).

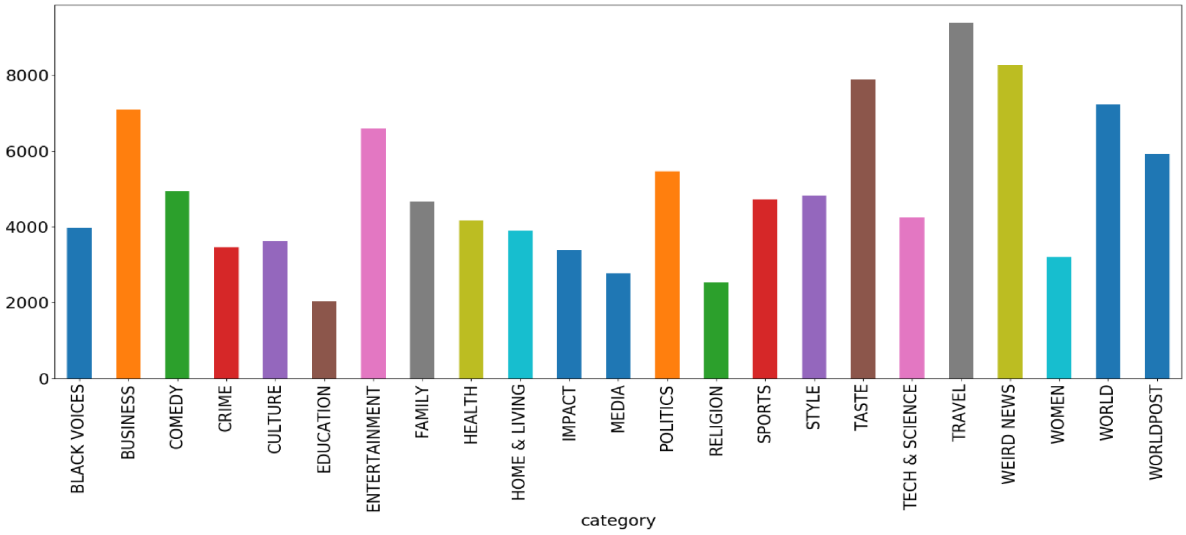

Рис. 5. Диаграмма с количеством новостей после сэмплирования.

Матрица смежности после проведения сэмплирования изображена на рис. 6. Данная матрица говорит о том, что точность стала заметно выше. Точность определения после проведения сэмплирования увеличилась до 90%.

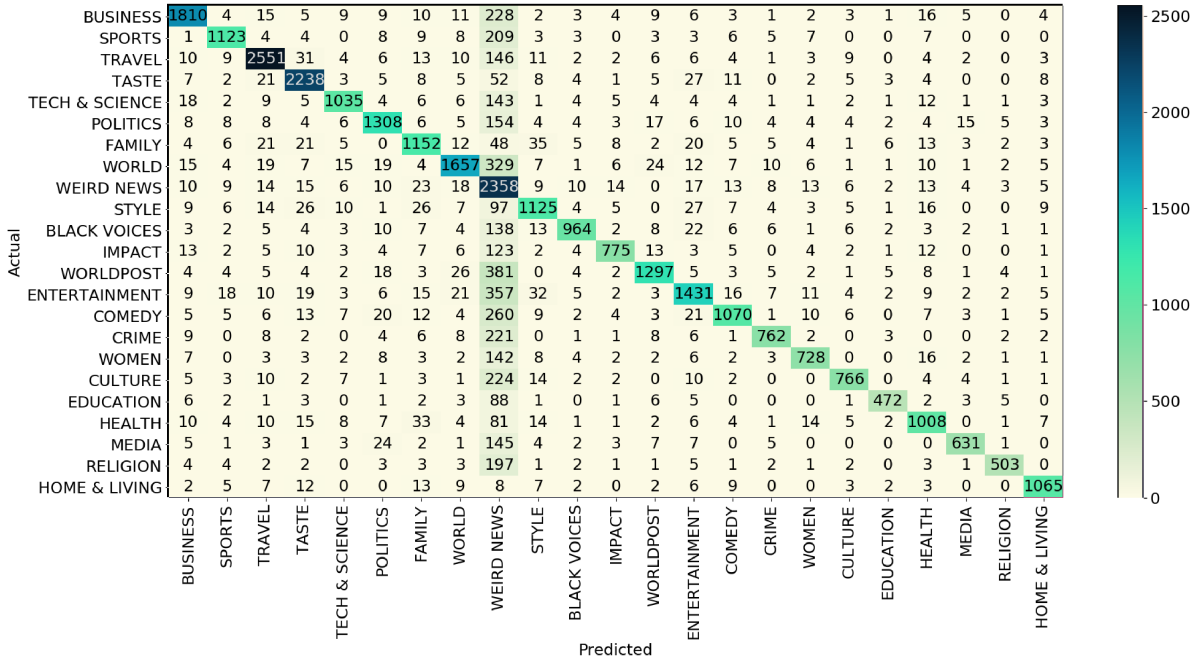

Рис. 6. Матрица смежности после сэмплирования.

Выводы

В работе на основе метода опорных векторов был построен классификатор новостей, точность которого по мере разработки возросла до 90%. Для обучения классификатора использовался «сырой» набор данных, который в дальнейшем был обработан с помощью кластеризации и сэмплирования. На примере было показано, как несбалансированные данные влияют на качество обучения, а следовательно, и на точность классификатора.

---

1. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – Электрон. дан. – М.: ДМК Пресс, 2015. – 400 с. – Режим доступа: https://e.lanbook.com/book/69955. – Загл. с экрана.

2. Бринк, Х Машинное обучение / Х. Бринк, Д. Ричардс, М. Феверолф. – СПб.: Питер, 2017. – 336 с.

3. Nefedov, Alexey. Support Vector Machines: A Simple Tutorial – 2016. – URL : https://svmtutorial.online/ (дата обращения: 13.02.2019).

4. Лозинская, А.М. Прогнозирование индекса ММВБ. Предсказательная сила метода нейросетевого моделирования и метода опорных векторов / А.М. Лозинская, В.А. Жемчужников // Вестник Пермского университета. Серия «Экономика». – Электрон. дан. – 2017. – № 1. – С. 49-60. – Режим доступа: https://e.lanbook.com/journal/issue/301052. – Загл. с экрана.

5. Близнюк, Б., Васильева, Л., Стрельников, И., Ткачук, Д. Современные методы обработки естественного языка. // Вестник Харьковского национального университета им. Каразина. Серия «Математическое моделирование. Информационные технологии. Автоматизированные системы управления». – 2017. – № 36. – С. 14-26. – Режим доступа: https://periodicals.karazin.ua/mia/article/view/10084.

6. FASAM – NLP Competition. Predict News Category / Kaggle: Your Home for Data Science. – URL: https://www.kaggle.com/c/fasam-dl-nlp (дата обращения: 15.10.2018).

7. News Category Dataset. Identify the type of news based on headlines and short descriptions / Kaggle: Your Home for Data Science. – URL: https://www.kaggle.com/rmisra/news-category-dataset (дата обращения: 20.10.2018).

8. Официальный сайт Python. – URL: https://www.python.org/ (дата обращения 20.11.2018).