Доклад

1. Абстракт

В тази статия представям резултатите и използваните методи за предсказване на категорията на новини на база тяхното заглавие и кратко описание. Успеваемостта на използвания подход, предсказвайки три категории новини с равен брой примери от всяка категория, е 67 процента. Методите на предсказване се базират на използването на ключови думи.

2. Въведение

2.1. Въведение в контекста

Задачата за определяне категорията на текст (в нашия случай – новини) възниква в множество задачи от реалния свят. В тази статия е показан един възможен подход за справянето с този проблем.

2.2. Проблем

Необходимостта от класификацията на текст възниква в множество проблеми от реалния свят. Част от тях са: класификация на спам, определяне на категорията на новини, персонализиране на реклами за даден потребител на база неговите постове в социалните мрежи, подобряване на SEO чрез автоматично категоризиране на съдържанието на уебсайтове и др.

2.3. Цел

Целта на този проект е да подпомогне и улесни заинтересованите страни със справянето с гореизложените поблеми, предоставяйки им подход за автоматизирано класифициране на текст, в частност – новини.

2.4. Дейности

- Изтегляне на данни необходими за тренирането и определянето на точността на използваните модели.
- Предварителна обработка на данните с цел представянето им в подходящ формат за използваните модели, което включва:
  - Семплиране на данните – селектиране на равен брой новини от всяка категория с цел по-добра оценка на точността на предсказването.
  - Определяне на ключови думи за всяка от наличните категории. Процесът по екстракция на ключовите думи не е автоматизиран – ключовите думи се подбират на общ принцип според съответната категория. Част от данните се използват за обогатяване на списъка с ключови думи.
  - Създаване на таблица, чиито колони представляват наличните категории към които можем да причислим съответната новина, а всеки ред от таблицата съответства на точно една новина. Всички налични новини се разполагат по редовете на тази таблица.
- Пускане на различни алгоритми за класификация и определяне на точността от предсказването.
- Обобщаване на резултатите и определяне на най-добре представилия се алгоритъм за класификация.

3. Материали и методи

Основните инструменти, използвани за реализирането на този проект, са програмният език R и средата за разработка RStudio заедно с външни библиотеки за обработка на текст (“rsjon”) и използване на алгоритми за машинно самообучение (“caret” и “e1071”). Към статията се прилага и кодът реализиращ целта (решението на проблема).

Основните методи за предсказване използвани в проекта са: Latent Dirichlet allocation (LDA), Classification and Regression Trees (CART), k-Nearest Neighbours (kNN), Support Vector Machines (SVM), Random Forest (RF) и Naive Bayes classifier.

4. Резултати и дискусии

След прилагането на методите: Latent Dirichlet allocation (LDA), Classification and Regression Trees (CART), k-Nearest Neighbours (kNN), Support Vector Machines (SVM), Random Forest (RF) и Naive Bayes classifier, получените резултати по метриката Accuracy, използвайки 10-fold cross validation са съответно следните: 66.63 %, 64.65 %, 66.95 %, 67 %, 66.83 %, 66.93 %. Както се вижда от резултатите, най-добра успеваемост в предсказването на категорията на новини дава алгоритъмът Support Vector Machines (SVM) – 67 %. След него се нареждат kNN с успеваемост 66.95 % и Наивният Бейсов класификатор с успеваемост 66.93 %. Трябва да се отчете фактът, че произволен генератор, при наличието на три категории новини, ще има успеваемост от едва 33.33 %.

5. Заключение

Класифицирането на текст по описания в текущата статия начин постига до голяма степен заложената цел, а именно – улесняване на заинтересованите страни в решаването на проблемите за класификация на спам, определяне категорията на новини, персонализиране на реклами за даден потребител, подобряване на SEO и др.

Проектът може да продължи своето развитие в посока автоматизиране на процеса по екстракция на ключови думи, което може да бъде постигнато чрез използване на tf-idf (term frequency-inverse document frequency) индекс на думите заедно с хи-квадрат тест за определяне значимостта на всяка дума и избиране на най-значимите измежду всички уникални думи в нашите данни. Възможно е да се използва n-грамен модел, който да включва униграми в комбинация с биграми.

6. Литература

[1] https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
[2] https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/