Классификация текста

Классификация текста – задача компьютерной лингвистики, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Классификация текстов как правило применяется для разделения сайтов по тематическим каталогам, борьбы со спамом, распознавания эмоциональной окраски текстов, персонификации рекламы и т. д. В данной работе рассматривается именно тематическая каталогизация.

Формализовано задачу классификации можно поставить следующим образом:

- X – множество описаний объектов,
- Y – конечное множество номеров (имён, меток) классов,
- $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$ – конечная обучающая выборка,
- y^* - целевая зависимость, отображение значения которой известны только на объектах обучающей выборки.

Требуется построить алгоритм $\alpha: X \to Y$, способный классифицировать произвольный объект $x \in X$.

В задаче агрегации новостных статей необходимо провести классификацию документов для определения тематики содержания. В общей сложности было выявлено и размечено 7 классов.

Предобработка и векторизация текста

Для очистки документа от возможных шумов и аномалий, которые могут помешать построению достоверного вектора, производится предобработка текста. Предобработка включает в себя: приведение слов к нижнему регистру, токенизацию, удаление стоп-слов, стемминг и лемматизацию.

Приведение слов к нижнему регистру необходимо для устранения помех, связанных с расстановкой одинаковых слов в начале и середине предложения, что меняет их регистр написания.

Токенизация- процесс разбиения текста на текстовые единицы, например, слова или предложения.

Удаление стоп-слов- извлечение слов с большой частотой встречаемости, которые не несут большой смысловой нагрузки.

Стемминг- нахождения основы слова.

Лемматизация- процесс приведения слова к нормальной (словарной) форме.

После предобработки документов необходимо выбрать алгоритм векторизации текста. В данной работе будут рассматриваться несколько векторизаторов, а именно: TF-IDF, Word2Vec, Doc2Vec, FastText, GloVe, Universal-Sentence-Encoder и Bert. Будут рассмотрены принципы их работы и произведено сравнение их влияния на точность классификаторов. Суть векторизации документов заключается в построении векторов для каждого текста. Близкие по смыслу с точки зрения модели документы (похожие по смыслу слова используются в сходных контекстах) будут близки по косинусной мере.

Косинусное сходство- это мера сходства между двумя векторами, которая используется для измерения косинуса угла между ними.

Word2Vec

Word2Vec- общее название для совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке. Архитектура Word2vec подразделяется на два вида– Skip-gram и Continuous Bag of Words (CBOW). Оба эти метода изучают веса, которые действуют как представления слов в векторе. В корпусе модель CBOW предсказывает текущее слово из окна окружающих контекстных слов, в то время как модель Skip-gram предсказывает окружающие контекстные слова по текущему слову.

Принцип работы Word2Vec заключается в нахождение связей между контекстами слов согласно предположению, что слова, находящиеся в схожих контекстах, имеют тенденцию значить похожие вещи, т.е. быть семантически близкими. Более формально задача стоит так: максимизация косинусной близости между векторами слов (скалярное произведение векторов), которые появляются рядом друг с другом, и минимизация косинусной близости между векторами слов, которые не появляются друг рядом с другом. Рядом друг с другом в данном случае значит в близких контекстах.

Word2vec строит вектор не для документа, а для отдельного слова, опираясь на контекст. С целью нахождения вектора цельного текста был выбран подход суммирования векторов слов, с учетом весов их частоты встречаемости.

Частоту встречаемости слов можно взять из TF-IDF модели:

$$DV = \frac{\sum_{i=1}^{n} w2v_i \times tfidf_i}{n},$$

где DV – итоговый вектор документа, $w2v_i$ – векторное представление i-ого слова в документе, $tfidf_i$ – частота встречаемости i-ого слова во всем корпусе, n – количество слов в документе.