УДК 004.67

ПОДХОД К ОПРЕДЕЛЕНИЮ ТЕМАТИЧЕСКОЙ БЛИЗОСТИ ТЕКСТОВ

Бодунов Г.А. (к-т ВКА имени А.Ф. Можайского
Тельбух В. В. (адъюнкт 61 кафедры ВКА им. А.Ф. Можайского)
Научный руководитель –
Бирюков Д.Н. (д.т.н., доцент, начальник 61 кафедры ВКА им. А.Ф. Можайского)

Аннотация:

В статье приведен сравнительный анализ подходов к представлению естественного языка в задаче определения тематической близости публикуемых новостных статей в популярных интернет-медиа СМИ. Рассмотрены основные алгоритмы перевода текстовой информации в векторное пространство и применены алгоритмы кластеризации к полученным данным (k-means, k-minibatches, DBSCAN). Исследуются методы, основанные на статистической мере появления слов (TF-IDF), мешке слов (CBOW), тематическом моделировании (LDA), а также алгоритмы перевода слов и документов в векторное пространство, полученных на основе работы моделей нейронных сетей (word2vec). Работа направлена на поиск наиболее результативных алгоритмов представления текстовой информации и кластеризации, а также их комбинаций, исходя из оцениваемых метрик качества.

Введение.

На сегодняшний день во всемирной информационной сети зарегистрировано более 1,9 миллиардов веб-сайтов, из них активных более 200 миллионов и ежедневно их количество увеличивается, что обусловлено глобальной информатизацией общества. Эффективная деятельность современного человека все больше зависит от его информированности. Ориентироваться и разбираться в огромном количестве и объеме информации становится всё труднее, тем самым возникает потребность в совершенствовании подходов обработки массивов текстовой информации – установление тематической близости, включающих в себя обработку текстов на естественном языке и их кластеризацию.

Основная часть.

Кластеризация текстовой информации является одной из фундаментальных задач информационного поиска. Решения этой задачи позволяют разбивать множества документов на близкие по смыслу подмножества, что особенно важно в поиске тематической близости новостных статей. Данный процесс можно разделить на 2 этапа:

- на первом этапе текстовые представления документов по определенным правилам переводят в векторные представления;
- на втором этапе к полученным векторным представлениям применяются различные алгоритмы кластеризации.

Используемые в данной работе алгоритмы реализованы в пакетах nltk, genism и sklearn написанных на языке Python.

Основным источником данных для обучения алгоритма является текстовая информация, получаемая с порталов основных российских интернет-медиа СМИ (Meduza, Lenta.ru, Коммерсантъ). В результате был сформирован датасет, предварительно обработанный с помощью средств библиотеки nltk. Далее полученные данные подавались на вход алгоритмам векторизации. В ходе работы алгоритмов word2vec, tfidf, cbow, lda были построены векторные представления каждого образца документа. Оценить качество полученного векторного представления можно исходя из косинусной меры расстояния между векторами одинаковых тематик документов. Полученные данные были использованы для обучения алгоритмов кластеризации.

В ходе эксперимента были исследованы алгоритмы DBSCAN, k-means, k-minibatches, hierarchical-clustering, реализованные в библиотеке scikit-learn, что позволило

провести сравнительный анализ и выявить наилучшую связку алгоритмов векторного представления текста и их кластеризации, исходя из оценивания метрик Inertia, Чистота, Энтропия, Silhouette Coefficient и AMI.

Параметры обучения модели были подобраны с помощью программного средства - GridSearchCV, встроенного в библиотеку scikit-learn. Все используемые алгоритмы также были собраны по средствам Pipline в одну последовательность и обучены. В последствии исходя из параметра Smart Voting был выбран наилучший алгоритм кластеризации с подобранным в GridSearch параметрами обучения модели.

Выводы:

В результате проведенного эксперимента среди выбранных алгоритмов кластеризации и векторизации текстов на естественном языке лучшие результаты показала комбинация DBSCAN+word2vec. Сравнительный анализ алгоритмов и их комбинаций позволил сделать вывод, что повышение результативности алгоритмов представления и кластеризации текстовой информации зависит от количества данных в используемой выборке, подбора параметров и алгоритма обучения, а также комбинаторики алгоритмов.