Отзыв

на выпускную квалификационную работу бакалавра

Логачева Михаила Максимовича

“Детектирование общественно значимых новостей в потоке сообщений”

В выпускной квалификационной работе бакалавра оценивается применимость существующих методов машинного обучения и кластеризации к решению задачи отделения уникальных кластеров сообщений, а также разработка программного комплекса для выделении общественно значимых событий на основе изученных методов.

Реализованная система состоит из нескольких частей: сбор сообщений из сети Twitter, обработка собранных данных, анализ результатов применения методов кластеризации к поставленной задаче. Следует отметить, что автор рассматривает сообщения на русском языке, что усложняет работу.

Проведены эксперименты, оценивающие работу системы в различных условиях. Для двух видов векторного представления собранных сообщений (Bag-of-Words и Word2Vec), полученных в результате преобразования данных, была проведена кластеризация всеми четырьмя способами. Подбор оптимального количества кластеров производился поиском числа кластеров, разбиение на которое давало наибольшее значение silhouette coefficient. Для проведения кластеризации были использованы библиотеки scikit-learn и pandas для языка Python.

По результатам проведенного анализа был сделан вывод, что для рассматриваемой задачи наиболее удачным представлением является векторное представление, учитывающее только именованные сущности в тексте сообщения. Наиболее эффективном методом кластеризации из рассмотренных оказался Affinity Propagation, так как он показал наибольшее значение silhouette coefficient

К недостаткам работы можно отнести:

1. В работе явным образом не описано, что автор понимает под общественно значимой новостью, и как это понятие соотносится с реализованной кластеризацией новостей.
2. Не приводится сравнение с существующими системами определения новостей.

В целом работа заслуживает оценки “отлично”.

Научный руководитель,
ст. преподаватель Малинина М.А.