УДК 004.896

РАЗРАБОТКА СИСТЕМЫ ПОСТРОЕНИЯ ОЦЕНКИ КАЧЕСТВА ЖИЗНИ РАЙОНОВ ГОРОДА НА ОСНОВЕ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ СОЦИАЛЬНОЙ СЕТИ

Замиралов А. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),
Ходорченко М.А. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н., доцент ИДУ, старший научный сотрудник НЦКР Бутаков Н.А.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе рассматриваются такие модули системы как «Тематическое моделирование» и «Извлечение аспектов предметной области и локационной информации», производится оценка качества модулей, происходит анализ перспективы пространственной агрегации для решения задачи оценки качества жизни.

Введение. В настоящее время существует немало способов оценить качество жизни районов, однако, большинство из них основывается либо на сухих числах о количестве развлекательных, культурных или иного рода точек в пешей доступности (данный способ не может быстро отследить динамику изменений в районе и сильно пренебрегает мнением самих жителей), либо на картах вручную составленных активистами города (данный способ имеет явный потолок человеческих сил и плохо масштабируется). Зачастую, качественный анализ проводится единожды и без открытой публикации.
В данной работе предлагается создать систему, способную автоматически и в реальном времени оценивать реакцию жителей на состояние района и проводимые в нем акции, визуализировать полученные результаты на карту города, выделять неадминистративные локальные-экономические образования.
Предполагается, что данное решение будет полезно людям, которые ищут или собираются сдавать в аренду жилье с лучшими для себя условиями; городским активистам, которым станет легче координироваться и отслеживать проблемные места; местному правительству, которое сможет быстрее отслеживать реакцию жителей. Система может являться частью умного города.

Основная часть. Данная работа посвящена двум основным модулям, ответственным за выявление семантического ядра из неструктурированного текста.
Тематическое моделирование. Идея в том, что если бы нам была известна релевантная нашим исследованиям тема каждого сообщения\поста из социальной сети, то мы смогли бы количественно оценить, какие темы наиболее волнуют жителей определенных районов. Трудность в том, что нет четкого списка интересующих нас тем, и построить идеально точную модель нетривиальная задача.
Предлагаемое решение: трехгруппная многошаговая модель аддитивной регуляризации.
На каждом шаге выделяется три группы: размеченные релевантные темы – топики, темы которых мы задали сами и точно хотим найти; общесодержательные темы – топики, которые образуются сами, могут иметь любую одну ярко выраженную тему; фоновые темы.
К каждой группе применяется свой набор из регуляризаторов. Относящиеся к жилищным проблемам темы из второй группы на новом шаге переносятся в группу релевантных тем.
Шаги повторяются пока в корпусе все еще находятся новые релевантные темы.
Полученная модель тестировалась с разным количеством тем. Лучшим решением оказалась модель со 100 темами, но несмотря на отлично найденные ключевые слова, качество на

общей тестовой выборки не доходило выше 70%. Основная часть ошибок на коротких сообщениях, размера которых недостаточно для четкого ответа тематической модели. Для обработки этих сообщений лучше подходит второй модуль.
Извлечение аспектов предметной области. Одним из видов этой задачи является NER (выделение именованных сущностей), где аспектами являются адреса и имена. Данная задача хорошо решена для русского языка. Идея в том, чтобы натренировать нейросетевую модель на выделение аспектов, которые соответствуют основным релевантным темам, полученным предыдущим модулем. Проблемы заключаются в том, что для качественного обучения моделей, показывающих стабильно высокое качество, требуются большой корпус размеченных данных, которого на русском пока еще нет. На данный момент работа ведется в трех направлениях, для последующего сравнения качества. Первый – использовать нестабильные модели с обучением без учителя, и соответственно без необходимости иметь размеченный корпус. Второй – использовать автоматические переводчики и пробовать анализировать короткие сообщения на русском, как короткие сообщения на английском. Третий – сделать автоматическую разметку небольшого корпуса длинных сообщений, с которыми хорошо справляется тематическая модель.
Предполагается, что данных решений будет достаточно для качественного определения семантического ядра сообщения.

Выводы. В ходе работы был разработан макет системы пространственной системы оценки качества жизни. Был разработан, протестирован и оценен подход по модулю тематического моделирования. Был произведен анализ литературы связанной с выделениями аспектов, и предложены подходы для частного решения этой проблемы. Дальнейшая работа будет производиться в области трекинга пространственных событий и дизайна визуализации полученной информации.

Замиралов А. (автор)	Подпись
Ходорченко М.А. (соавтор)	Подпись
Бутаков Н.А. (научный руководитель)	Подпись