ТЕХОЛОГИЯ NLP: СЦЕНАРНОЕ ПРОЕКТИРОВАНИЕ И ГОРОДСКАЯ СРЕДА

В настоящее время возможности для исследования городских пространств в России существенно ограничены узким спектром качественных и открытых источников геоинформационных данных, но развитие средств информатизации способствует росту количества хранимой информации, в первую очередь - текстовой. Терминология Big Data (большие данные), которая сейчас является обозначением для огромных массивов информации, появилась двадцать лет назад, в полной мере отражает экспоненциальный рост объемов информации. Чаще всего эти данные не являются структурированными, не относятся к одной тематике. Все эти факторы затрудняют информационный поиск релевантной информации. Поэтому современным и востребованным направлением исследований являются методы анализа и генерации текста, обработки естественного языка NLP (Natural Language Processing).

Методология NLP применяется обычно в задачах классификации с использованием структурирования, при этом определяются параметры, которые позволяют быстрее проводить обработку корпусов обучения. Многие из существующих алгоритмов работают с контекстом и тематическим моделированием, но каждый из них имеет как положительные стороны, так и недостатки. Методы строятся на допущениях, свойствах обобщения, распространении существующих методологий, что не всегда приводит к положительному результату при работе с данными, отличными от обучающего множества.

Функции потерь при NLP часто носит общий характер и не учитывает контекст. Также при работе с данными, например, звуковыми, возникает сложность в распознавании фонем, появляется шум и фон, значит, необходимо также учитывать контекст, взаимное влияние данных. Возникает потребность в расширении методологий, формализаций и реализаций, которые могут помочь в выделении контекста. **Поэтому существует необходимость в разработке новых методов поиска контекста и обработки информации.**

Методы работы с информацией о городе должны быть легко экстраполированы для работы с различной территорией. Анализ текста как раз является универсальным методом

работы, прежде всего. Контент-анализ не является типичным инструментом в сфере обработки городских данных, несмотря на то, что именно текстовые данные составляют наиболее полную характеристику территории.

Задачи на Мастерскую

1. Подготовка и агрегация текстовых данных о городе
2. Векторизация текста. Применение частотного анализа
3. Применение методов тематического моделирования для выделения востребованных тем при благоустройстве
4. Модификация метода pLSA для ситуативного моделирования

В нашей мастерской будут решаться задачи соучаствующего проектирования. Соучаствующее проектирование означает вовлечение горожан в городское планирование либо путем их включения непосредственно в формировании города, например, в реконструкции районов, либо путем сбора и выявления соответствующих данных для градостроительного процесса. Такое участие устраняет разрыв между гражданами, экспертами и политиками, в том смысле, что граждане не рассматриваются как простые испытуемые или потребители городского пространства.

Тематическое моделирование как один из методов машинного обучения, именно тот метод, которого не хватает в исследовании контекста градостроительства. Семантический анализ необходим для подтверждения достоверности тематического распределения. Например, в разное время дня мы можем увидеть разные частоты использования разных объектов (слов), которые можно обоснованно связать с характерными действиями в эти периоды времени. Для визуализации можно использовать гистограммы, облака слов, кластера. Большую ценность это приносит, если ещё использовать картографические данные и графы транспортной доступности. Утром, когда люди обычно ездят на работу, мы можем встретить слова, связанные с транспортом, такие как «вокзал», «аэропорт» и «метро», а также утренние занятия, такие как «пить кофе». Полученные тематические кластеры могут быть использованы для дальнейшего анализа мнения граждан и прогнозирования новых востребованных тем, а также при принятии решений. (см. рисунок 1, этапы 2, 3 и 4).

Рисунок 1 - Порядок разработки проекта благоустройства

Одним из возможных подходов к решению проблемы учета общественного мнения является разработка инструментов для создания “дизайн-игры”. Суть этой методики заключается в следующем: каждый пользователь (участник) размещает значки-индикаторы различных зон на карте, и затем на основе общей картины голосования формируется карта, поделенная на зоны активности, в зависимости от количества голосов и их расположения.

Для реализации алгоритма за основу возьмем вероятностный латентно-семантический анализ pLSA, в котором рассматривается множество всех документов D и всех термов W, причем термы ранее получены предобработкой с помощью «мешка слов» и имеют квантовую запутанность. Использование механизмов квантовой теории поможет более точно использовать пространственность городских данных, также есть вариант использовать алгоритмы тематического моделирования на графах.

Так как встроенные библиотеки тут бессильны, то необходима собственная реализация, поэтому решение задачи оптимизации сводится к использованию EM-алгоритма.

- на E-шаге по найденным на предыдущем шаге параметрам $\varphi_{wt}, \theta_{td}$ определяется количество слов в документе, которые относятся к теме, количество слов, порожденных темой в документе;

- на M-шаге по значениям количества слов в документе, которые относятся к теме

происходит переход к новому базису тем с использованием матрицы поворота и пересчет оценок $\varphi_{wt}, \theta_{td}$ по формуле.

Итерационный процесс продолжается, пока не будет решена задача максимального правдоподобия.

Подобная модификация алгоритма опирается на процесс итерационного вычисления скрытых параметров на основе представления документов и слов по принципу суперпозиции с использованием условий нормировки, учитывающих вероятность. Геометрическая составляющая модели опирается на проекции векторных представлений слов и документов на базис из скрытых тем, а также на ортонормированность векторов скрытых тем, которые пересчитывается на каждом шаге EM-алгоритма.

Основные задачи, которые будем решать на мастерской:

- Агрегация и обработка больших данных (текстовые данные с различных источников)
- Изучение библиотек, для работы с картографическими данными, извлечение необходимых городских кластеров, сети дорог
- Модификация алгоритма тематического моделирования для работы с городскими данными
- Объединение результатов текстовой обработки и карт
- Визуализация результатов

Внимание!

За время мастерской у нас вряд ли получится сделать полную реализацию алгоритма, но зато точно получится применить методы извлечения контекста текста к городским данным и изучить влияние кластеров на принятие решения и развитие городских территорий