УДК 004.042

Сравнительный анализ алгоритмов определения геолокации по текстовым сообщениям о возможных сбоях в организации, размещенных на веб-сайтах социальных сетей.

Сивков Д.И., Университет ИТМО
Научный руководитель – к.т.н., доцент ФБИТ, Воробьева А.А. Университет ИТМО

На основе сравнительного анализа алгоритмов определения геолокации по текстовым сообщениям, разработан модуль для точного и быстрого выявления городов в тестовых сообщениях с выводом координат.

Введение. В докладе представлен обзор алгоритмов и методов, таких как, простой поиск подстроки в строке, нечеткий текстовый поиск, префиксное дерево, алгоритм Рабина-Карпа для задачи определения геолокации по текстовым сообщениям. Целью является практическая реализация каждого из методов и выявления наилучшего, по значениям скорости работы и точности.

Основная часть. Приводится обзор и последующий анализ научно-технической и методической литературы, для последующей разработки экспериментального образца, основанного на самом точном методе выявления названия городов и последующем выводе координат, с целью выявления текстовых сообщений, размещенных на веб-сайтах и социальных сетях Интернета.

Задача, выявлять названия городов, например: «Москва», «Санкт-Петербург», «Тюмень». Также все осложняется тем, что в России более 1000 городов, а также имеются не официальные название городов, такие как: «Мск», «СПб», «Питер» и т.д. Также все города могут использоваться в тексте в разных падежах, что дополнительно усложняет поиск. Поиск будет проводиться по базе данных городов и их координат. Каждому сообщению пользователя будет присвоена метка, отражающая наличие города и соответсвенного координат в тексте. Все определенные города в сообщениях будут выведены на экран.

Таким образом задача определения геолокации в сообщении сводится к последовательному решению следующих подзадач:

1) предобработка текста;

2) выявление названия населенного пункта;

3) вывод координаты в базу данных.

Вывод. По результату данного исследования сделан следующий вывод, что для решения задачи поиска городов по тексту лучше использовать структуру данных префиксного дерева (trie), для реализации словаря (ассоциативного массива), ключами в котором являются строки, поскольку данный метод показал, что способен определять 99% городов за самое короткое время.

Сивков Д.И. (автор)

Воробьева А.А. (научный руководитель)