LSA-АЛГОРИТМ ПРЕДВАРИТЕЛЬНОГО АНАЛИЗА ТЕКСТОВ

Дудник А.В., Евсина Н.А.
Национальный технический университет «Харьковский политехнический институт», г. Харьков

К числу вспомогательных задач, решаемых при построении сложных систем управления и систем поддержки принятия решений, относятся задачи анализа текстов. Среди них кластеризация, выбраковка «синтезированных» текстов и т.п. Для этих целей широко применяются алгоритмы латентно-семантического анализа (LSA). В [1] предложен LSA-алгоритм, использующий сингулярное разложение предварительно преобразованного корпуса текстов. В данной работе продолжены исследования и модификация алгоритма, начатые в [2].

Из 10 текстов, каждый величиной до 1000 слов, после предварительной обработки, была сформирована матрица размерностью 3378x10, содержащая вес каждого слова в соответствующем тексте. Подвергнув её сингулярному разложению, получили три матрицы, смысл которых интерпретируется следующим образом: темы текстов (матрица разложения), слова и темы, тексты и темы. Матрица текстов и тем имеет размерность 10x10, т.е. предполагается, что каждый текст посвящён одной основной теме, но также содержит дополнительные темы, характеризующиеся меньшим весовым значением. Это предположение подтверждается результатом разложения: в каждом столбце матрицы — для каждого текста — есть своё максимальное значение, которое также является максимумом в соответствующей строке — для каждой темы.

Если принять за оси координатной плоскости две темы и разместить на этой плоскости тексты сообразно их тематическим весам, то большая часть текстов сформирует облако вблизи начала координат — это тексты, в которых данные темы представлены слабо. Два текста, имеющие максимальные веса для одной из выбранных тем, будут наиболее удалены от начала координат, имея смещение к оси основной темы. Также будет заметно 1-3 текста, которые дистанцируются от «нулевого» облака, в то же время их весовые значения существенно меньше максимума — это тексты, в которых данные темы имеют второстепенное значение. Для анализа следует выбирать темы, сингулярное число которых наибольшее.

Исследования были выполнены в среде MATLAB. Дальнейшие исследования ориентированы на предварительную обработку текстов.

Литература:

1. Алгоритм LSA для поиска похожих документов. // [Электронный ресурс] – URL: https://netpeak.net/ru/blog/algoritm-lsa-dlya-poiska-pohozhih-dokumentov/

2. Дудник А.В. Модуль предварительного анализа текстов / А.В. Дудник, Н.А. Евсина, Е.В. Клевцова / Актуальні проблеми автоматики та приладобудування: матеріали III Міжнародної науково-технічної конференції 3-4 грудня 2020 – Харків: ФОП Панов А.М., 2020, с. 13-14.