|  |
| --- |
| Секция: Программирование |
| ГБОУ "Лицей № 1533 (информационных технологий)", г. Москва |
| 119296, Москва, г. Москва, Ломоносовский проспект 16 |
| тел.: (495) 133-20-00; E-mail: [info@lit.msu.ru](mailto:info@lit.msu.ru) |
| Кластеризация результатов поиска |
| Гайдамашко Даниил Олегович |
| тел.: +7 985-258-27-38; E-mail: [dan-gai@yandex.ru](mailto:dan-gai@yandex.ru) |
| Класс: 11 |
| 121351, Москва, ул. Кунцевская 4 корп.1, кв. 20 |
| Карпенко Максим Дмитриевич |
| тел.: +7 925-618-10-63; E-mail: [maxim020998@gmail.com](mailto:maxim020998@gmail.com) |
| Класс: 11 |
| 119571, Москва, Академика Анохина 9, кв. 25 |
| Научный руководитель: Завриев Николай Константинович, преподаватель ГБОУ ЛИТ 1533 |

Проблема большинства интернет-поисковиков заключается в том, что поисковая выдача представляет собой набор ссылок на статьи, описывающие разные объекты, что существенно усложняет поиск необходимой информации.

Проблема решается с помощью кластеризации результатов поиска – разделения результатов поиска на смысловые группы с возможностью для пользователя искать в той группе, которая ему нужна. При этом алгоритм кластеризации, который направлен на выявление семантической сущности текста, имеет потенциал использования и в других интернет-сервисах.

Для решения проблемы кластеризации текстов был выбран алгоритм, получивший название алгоритма латентно-семантического анализа [3], позволяющий выявить зависимости между документами и терминами, в них встречающихся. Результатом выполнения алгоритма является граф, вершины которого соответствуют отдельным документам и тегам, относящимся к этим документам. О семантической близости документов и тегов можно судить при развертке графа в пространстве и взвешивании рёбер.

Для анализа полученного графа был разработан собственный алгоритм выделения кластеров в наборе вершин-текстов и вершин-тегов.

Таким образом, разработанный проект состоит из двух частей. Первая – это графический исследовательский инструмент, реализующий и демонстрирующий кластеризацию введенных текстов, а также позволяющий настраивать параметры алгоритма и улучшать тем самым качество кластеризации. Вторая часть (далее *кластерный-поисковик*) – непосредственно веб-инструмент, выполняющий кластеризацию результатов интернет-поиска, в основе которого лежат алгоритмы, отработанные при помощи графического инструмента.

К функциям графического инструмента относятся:

* Получение на вход некоторое количество текстов, введенных пользователем;
* Выполнение латентно-семантического анализа и формирование кластеров;
* Отображение результатов группировки текстов на графике, где каждый кластер представляет собой связанный граф, вершинами которого являются теги и тексты.

К функциям веб-кластеризатора относятся:

* Получение поискового запроса пользователя и отправление его в существующий веб-поисковик;
* Получение страниц результатов поиска, извлечение из нее необходимой информации и сохранение в базе данных;
* Обработка полученных данных, выполнение латентно-семантического анализа и алгоритма выделения кластеров;
* Формирование кластеров и их вывод на веб-страницу в виде папок, описанных набором тегов и содержащих набор поисковых результатов, относящихся к данному кластеру.

Программная логика веб-кластеризатора реализована на языке C# в среде Microsoft Visual Studio с использованием фреймворка ASP.NET MVC 4, а для написания интерфейса сайта использовались языки HTML, CSS и JavaScript.

Графический инструмент также был реализован на языке C# в среде разработки Microsoft Visual Studio 2010 с помощью .NET Framework 4.5.

В качестве дальнейших разработок планируется дальнейшая оптимизация и улучшение качества работы алгоритма кластеризации, а также визуализация кластеров в веб-поисковике.

Использованные материалы:

* Фридл Дж. Регулярные выражения, 3-е издание. СПб.: Символ-Плюс, 2008 г.
* Edunov «Латентно-семантический анализ» [Электронный ресурс]:  
  [http HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/":// HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/"habrahabr HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/". HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/"ru HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/"/ HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/"post HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/" HYPERLINK "http://habrahabr.ru/post/110078/"/110078/](http://habrahabr.ru/post/110078/)
* andreycha «Обзор алгоритмов кластеризации данных» [Электронный ресурс]:  
  [http HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/":// HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/"habrahabr HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/". HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/"ru HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/"/ HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/"post HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/" HYPERLINK "http://habrahabr.ru/post/101338/"/101338/](http://habrahabr.ru/post/101338/)
* «SVD-разложение прямоугольной матрицы» [Электронный ресурс]:  
  [http HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php":// HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"alg HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"l HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"ib HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php". HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"sources HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php". HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"ru HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"/ HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"matrixops HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"/ HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"general HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"/ HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"svd HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php". HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php" HYPERLINK "http://alglib.sources.ru/matrixops/general/svd.php"php](http://alglib.sources.ru/matrixops/general/svd.php)
* «Учебник HTML» [Электронный ресурс]:  
  <http://webremeslo.ru/html/glava0.html>
* «Учебник CSS» [Электронный ресурс]:  
  [http://webrem HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"e HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"slo.ru/ HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"css HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"/ HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"glava HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"0. HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html" HYPERLINK "http://webremeslo.ru/css/glava0.html"html](http://webremeslo.ru/css/glava0.html)