МОСКОВСКИЙ КОМИТЕТ ОБРАЗОВАНИЯ

ЛИЦЕЙ №1533 (ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ)

**Выпускная работа**по специальности «Прикладное программирование»  
учеников группы 11.4  
Гайдамашко Даниила Олеговича,  
Карпенко Максима Дмитриевича

Кластеризация  
результатов веб-поиска

**Научный руководитель:**   
Завриев Николай Константинович  
преподаватель лицея1533 (информационных технологий)

**Москва, 2015 г.**

Оглавление

[Введение 3](#_Toc426410013)

[Постановка задачи 4](#_Toc426410014)

# Введение

В настоящее время Всемирная сеть важна и полезна практически для каждого. Любой в ней может найти огромное количество разнообразнейшей информации, которой накопилось так много, что актуальна необходимость использования специальных сервисов, позволяющих ориентироваться в информационных ресурсах и находить нужные сведения. Ими являются поисковые системы -  компьютерные системы, предназначенная для поиска информации, одним из наиболее известных применений которых и есть веб-сервисы для поиска текстовой или графической информации во Всемирной паутине.

Для поиска информации с помощью поисковой системы пользователь формулирует поисковый запрос, по которому система находит документы, содержащие указанные ключевые слова или слова, как-либо связанные с ключевыми словами. При этом поисковая система генерирует страницу результатов поиска, которая может содержать различные типы результатов: веб-страницы, изображения, аудиофайлы. А некоторые поисковые системы могут извлекать информацию из баз данных и каталогов сайтов в Интернете - структурированных по темам наборов ссылок на сайты с кратким их описанием. Очевидно, что поисковая система тем лучше, чем больше она возвращает результатов поиска, релевантных, т.е оказывающихся важными и полезными для пользователя.

Порядок каждого документа в поисковой выдаче зависит от таких факторов, как внутренняя структура, число и расположение ключевых слов и количество внешних ссылок на документ. Поэтому, к сожалению, недостаток поисковых систем в том, что при неточно сформулированном запросе результаты оказываются в большем количестве и не сгруппированными тематически, что существенно затрудняет поиск. А формулировать поисковый запрос более подробно у пользователя часто не хватает времени или терпения.

# Постановка задачи

В нашем дипломном проекте мы поставили перед собой две задачи: первичную и вторичную. Первичная задача является обязательной, в то время как вторичную необходимо выполнить по мере возможностей, на основе полученного в ходе решения первичной опыта. Ниже представлены описания обеих задач.

Задача №1

Создать приложение, реализующее латентно-семантический анализ нескольких текстов и графически представляющее результаты анализа. В связи с этим, приложение должно:

1. Получать на входе несколько текстов.
2. Выполнять обработку текстов, предписанную методом ЛСА, с последующим формированием графа, множество вершин которого представлено тегами слов и текстами.
3. Выполнять выделение кластеров вершин при помощи метода минимального остовного дерева с возможностью регулировки коэффициента выделения пользователем.
4. Представлять граф, полученный в пункте (2) в графическом виде и иллюстрировать разделение вершин по кластерам.

Задача №2

Создать веб-сайт, устроенный по принципу стандартных интернет-поисковиков, который реализует кластеризацию результатов веб-поиска на запрос, введенный пользователем. Для этого ресурс должен:

1. Отправлять запрос в уже существующий поисковик (Google, Yandex, Yahoo, Rambler и пр.) и получать оттуда ответы на запрос.
2. Составление кластеров происходит путем обработки и тегирования веб-страниц — присваивания каждой набор тэгов. Чем больше тэгов совпадает, тем больше вероятность текстов веб-страниц быть на одну и ту же тему и находиться в одном кластере.
3. Выдавать результаты не линейно, как в изначальных поисковиках, а по группам в виде открывающихся папок. В каждой папке находится множество ссылок на веб-страницы, сходные по тематике и относящиеся к одному “главному объекту”.

# Обзор аналогов

**Полных аналогов ни графического инструмента кластеризации текстов, ни веб-инструмента кластеризации результатов интернет-поиска найдено не было, он в настоящий момент уже существуют веб-ресурсы, реализующие кластеризацию интернет-поиска, и, как следствие, упрощение поискового процесса, а также инструменты для обработки семантики текстов.**

|  |  |  |
| --- | --- | --- |
| Название инструмента | Достоинства | Недостатки |
| **Yyppi (ранее Clusty)** | * Результаты кластеризации по наиболее часто встречаемым в результатах поиска ключевым словам представляются в виде списка пунктов меню, по которым возможна пересортировка результатов выдачи. * При отображении кластеры упорядочиваются по статистике найденных слов. | * Поиск с учетом словоизменения (учет морфологии) реализован только для английского языка. * Поиск по энциклопедиям реализован только на основе использования метапоиска по англоязычной части энциклопедии wkipedia, даже если запрос пользователя введен на русском языке. * Метапоисковая система не использует ни синтаксического, ни семантического анализа текстов. Поиск проводится по всем словам запроса пользователя. |
|  |  |  |
|  |  |  |