Оглавление

[Введение 5](#_Toc327657599)

[1. Аналитический раздел 6](#_Toc327657600)

[1.1 Постановка задачи 6](#_Toc327657601)

[1.2 Обзор существующих техник кластеризации текстов 8](#_Toc327657602)

[1.2.1 Методы, основывающиеся на метрике близости 8](#_Toc327657603)

[1.2.2 Алгоритмы Single / Complete / Average Link 10](#_Toc327657604)

[1.2.3 Алгоритмы KMeans и EM (Expectation Maximization) 11](#_Toc327657605)

[1.3 Кластеризация с помощью суффиксного дерева (STC – Suffix Tree Clustering) 13](#_Toc327657606)

[1.3.1 Построение дерева 13](#_Toc327657607)

[1.3.2 Виды суффиксных деревьев 14](#_Toc327657608)

[1.3.3 Достоинства и недостатки метода STC 16](#_Toc327657609)

[1.4 Метод латентно семантического анализа (LSA) 17](#_Toc327657610)

[1.4.1 Этапы подготовки данных для метода LSA 18](#_Toc327657611)

[1.4.2 Основные этапы работы метода LSA 18](#_Toc327657612)

[2. Конструкторский раздел 20](#_Toc327657613)

[2.1 Метод латентно-семантического анализа (ЛСА) 20](#_Toc327657614)

[2.1.1 Удаление стоп - символов 20](#_Toc327657615)

[2.1.2 Операция стемминга 21](#_Toc327657616)

[2.1.3 Удаление слов, встречающихся в единственном экземпляре 21](#_Toc327657617)

[2.1.4 Матрица употребляемости 22](#_Toc327657618)

[2.1.5 Сингулярное разложение 23](#_Toc327657619)

[2.1.6 Улучшения, применяемые к частотной матрице 27](#_Toc327657620)

[2.2 Алгоритм Fuzzy K-Means 28](#_Toc327657621)

[2.3 Разработка программного продукта 30](#_Toc327657622)

[2.3.1 Модуль предварительной обработки текстов 33](#_Toc327657623)

[2.3.2 Вычислительный модуль 35](#_Toc327657624)

[2.3.3 Модуль визуализации 37](#_Toc327657625)

[3. Технологический раздел 39](#_Toc327657626)

[3.1 Выбор языка программирования и среды разработки 39](#_Toc327657627)

[3.2 Требования к вычислительной системе 41](#_Toc327657628)

[3.3 Пользовательский интерфейс 42](#_Toc327657629)

[3.3.1 Главное окно программы 42](#_Toc327657630)

[3.3.2 Основные настройки приложения 43](#_Toc327657631)

[3.3.3 Настройки LSA 44](#_Toc327657632)

[3.3.4 Настройки кластеризации 45](#_Toc327657633)

[3.3.5 Результаты кластеризации. Общий вид документов 46](#_Toc327657634)

[3.3.6 Результаты кластеризации. Общий вид кластеров 47](#_Toc327657635)

[3.3.7 Детализированное представление кластеров 48](#_Toc327657636)

[3.3.8 Детализированное представление документа 49](#_Toc327657637)

[4. Исследовательский раздел 50](#_Toc327657638)

[5. Организационно – экономический раздел 50](#_Toc327657639)

[6. Промышленная экология и безопасность 51](#_Toc327657640)

[6.1 Опасные и вредные факторы при работе с ПЭВМ 51](#_Toc327657641)

[6.1.1 Освещенность 52](#_Toc327657642)

[6.1.2 Низкочастотные электрические и магнитные поля 53](#_Toc327657643)

[6.1.3 Статическое электричество 55](#_Toc327657644)

[6.1.4 Воздух рабочей зоны 55](#_Toc327657645)

[6.1.5 Опасность поражения электрическим током 58](#_Toc327657646)

[6.1.6 Опасность возникновения пожара 60](#_Toc327657647)

[6.1.7 Негативное воздействие шума 61](#_Toc327657648)

[6.1.8 Негативное воздействие недостатка освещения 62](#_Toc327657649)

[6.1.9 Утомление и травматизм кистей рук 62](#_Toc327657650)

[6.2 Расчет системы пожарной сигнализации 63](#_Toc327657651)

[6.2.1 Выбор типа пожарного извещателя и расчет защищаемой площади 64](#_Toc327657652)

Введение

В настоящее время количество и общий объем текстовых документов, используемых в научных работах, бизнесе и других областях человеческой деятельности неуклонно растет. Этот рост выражается как в увеличении общего объема текстовых коллекций, так и в значительном росте количества документов, требующих обработки и сортировки.

Следствием столь бурного роста объема текстовых данных в мире, а также стремительного развития сети Интернет, стала необходимость предварительной систематизации текстовых массивов в автоматическом режиме: в первую очередь, появилась актуальная задача поиска нужной информации в больших и сверхбольших коллекциях текстовых данных (например, выдаваемых поисковыми системами в ответ на запросы пользователей).

Кроме того, без систематизации текстов невозможно решение и ряда других задач, в частности:

* реферирования больших документальных массивов;
* определения взаимосвязей между группами документов;
* упрощения визуализации текстовой информации;
* выявления дубликатов или близких по содержанию документов;
* контент-анализа и др.

В этих условиях разработка специальных методов кластеризации текстовых документов, т.е. разбиения текстовых массивов на систему (возможно иерархических) схожих подмножеств, становится важной задачей, без решения которой эффективная работа с текстовой информацией невозможна. Кластеризация текстового массива позволяет систематизировать коллекцию и сузить область рассматриваемых документов. Человек получает инструмент для решения задач, связанных с анализом больших массивов текстовой информации.

1. Аналитический раздел
   1. Постановка задачи

Целью данной работы является разработка программного продукта для проведения многомерной кластеризации корпуса текстовых документов и визуализации результатов. Для достижения поставленной цели необходимо решить следующие основные задачи:

* анализ алгоритмов кластеризации текстов;
* выбор наиболее подходящего алгоритма для реализации;
* разработка программного продукта для осуществления кластеризации корпуса документов и визуализации результатов;

Входные данные:

* Коллекция текстовых новостных документов в формате XML. (Документ получен из новостного источника ROMIP News);

Выходные данные:

* Набор кластеров, построенных по данной новостной коллекции;
* Интерактивное визуальное представление результатов кластеризации;

Требования, предъявляемые к программному продукту:

* возможность ручной настройки основных параметров предварительной обработки документов и кластеризации;
* автоматическая генерация и присваивание имен полученным в результате вычислений кластерам;
* возможность навигации по построенным кластерам и отдельным документам, просмотра содержимого документа, просмотра основных тэгов кластеров;
* кластеризация должна быть нечеткой (один документ может принадлежать к нескольким кластерам с разным весом принадлежности);
* наглядное представление полученных результатов (визуальное выделение групп связанных документов (кластеров), степеней принадлежности каждого документа к построенным кластерам);
  1. Обзор существующих техник кластеризации текстов

При рассмотрении в самом общем виде методы кластеризации могут быть разбиты на два типа: представляющие тексты в виде векторов в многомерном пространстве признаков (и использующие метрику близости между векторами) и методы, основывающиеся на других представлениях кластеризуемых текстов. Первая группа методов представлена алгоритмами иерархической кластеризации, такими как Single / Complete / Average Link, неиерархическими алгоритмами K-Means / EM, а также большим числом других базирующихся на них методов. Примером алгоритмов второй группы может выступать алгоритм Suffix Tree Clustering (STC).

* + 1. Методы, основывающиеся на метрике близости

В методах кластеризации, основывающихся на метрике близости, документ представляется в виде многомерного вектора в пространстве признаков. Подходы к формированию вектора признаков документа могут существенно различаться. В простейшем случае каждый признак соответствует наличию в тексте одной из словоформ, встречающейся в рассматриваемом наборе текстов. Величина компоненты вектора тоже может определяться по-разному: например, компонента может быть равна единице, если данный термин присутствует, и нулю в противоположном случае. Она может быть равна числу вхождений термина в документ (вектор частот) или вычисляться по несколько более сложным формулам, учитывающим среднюю встречаемость словоформы по всему набору текстов или по внешнему корпусу текстов (т.н. интегральная значимость, часто обозначаемая tfidf). Вектор, отвечающий данному тексту, нормируется на единицу. Мерой близости между текстами считается скалярное произведение между соответствующими векторами.

Большинство алгоритмов кластеризации оперирует в качестве входных данных прямоугольной матрицей T, составленной из векторов документов, и симметричной квадратной матрицей S=T·Tt, именуемой матрицей близости. Матрица T, составленная из векторов tfidf, обозначается TFIDF. Основной проблемой методов, основанных на такой матрице, оказывается слишком большая размерность пространства признаков, большая часть которых является избыточными и даже вредными. Например, при кластеризации текстов в данной предметной области термины, не относящиеся к этой области, могут маскировать сходство между документами.

Для уменьшения размерности пространства признаков могут применяться, в частности, следующие приемы:

* Использование стоп-листов, содержащих списки «несмысловых» слов.
* Использование методов лингвистики:
  + - Использование словарей и тезаурусов для группировки словоформ по нормальным формам и объединения нормальных форм в синонимические группы.
    - Более развитый вариант того же метода может использовать семантическую сеть и группировать термины с использованием отношений более сложного типа (например, голонимии и гипернимии). Некоторые кластеризационные алгоритмы используют в качестве элементов вектора tf не отдельные словоформы, а выделяемые именные или глагольные группы, имена собственные, устойчивые словосочетания. В качестве вспомогательных, для уточнения величин компонент вектора tf могут применяться разнообразные методы разрешения омонимии и полисемии.
* Использование лингво-статистических методов, использующих информацию об априорных вероятностях встречаемости слов для включения в вектор tf только статистически значимых терминов, например, определив величину доверительного интервала для частот, и включая только термины, частота которых в каком либо из текстов выходит за верхнюю границу интервала.
* Использование алгебраических и вероятностных методов для разложения векторов в пространстве признаков по минимальному числу главных линейно-независимых компонент, с учетом корреляционных связей между терминами.

При использовании описанного представления текстов кластер оказывается набором векторов в пространстве признаков. Функция близости между кластерами выводится тем или иным образом из функции близости между векторами. Единственное ограничение – требование монотонности функции близости. Важной характеристикой кластера является положение его центроида (т.е. вектора центра масс, среднего арифметического векторов).

* + 1. Алгоритмы Single / Complete / Average Link

Наиболее популярными иерархическими алгоритмами, основанными на близости текстов в пространстве признаков, являются алгоритмы Single/Complete/Average Link. Есть варианты этих алгоритмов, работающие как сверху-вниз (рассматривающие сначала все тексты как один кластер), так и снизу-вверх (начинающие работу с кластеров, состоящих из единственного документа). Результатом работы этих алгоритмов является дендрограмма (бинарное дерево), связывающее все тексты. При заданном вручную числе кластеров (или предельной величине близости) делается соответствующее сечение бинарного дерева, дающее разбиение текстов на кластеры. Алгоритм кластеризации «снизу-вверх» локально оптимален, и работает на объединение двух ближайших элементов.

Расстояние между кластерами может определяться как:

* Минимальное расстояние между парой объектов в соседних кластерах (для Single Link)
* Максимальное расстояние между парой (для Complete Link)
* Среднее расстояние между элементами двух кластеров (что соответствует скалярному произведению центроидов)

Метод Single Link работает быстрее, чем Complete Link (O(N2) против O(N3), где N – число документов), но построенные кластеры часто оказываются чрезмерно «вытянутыми». Метод Group Average представляет собой компромисс между Single и Complete Link по скорости/точности.

В контексте поставленной задачи данные методы представляются малопригодными, так как они, в сущности, не удовлетворяют требованиям, предъявляемым к программному продукту:

* смысл полученных кластеров трудно формализовать, так как кластер описывается главным образом положением его центроида в многомерном пространстве признаков;
* документы относятся лишь к одному кластеру;
  + 1. Алгоритмы KMeans и EM (Expectation Maximization)

Несколько более подходящими представляются алгоритмы KMeans и EM (Expectation Maximization). Под этим названием объединяется обширная группа отличающихся высокой производительностью (порядка O(N)) неинкрементных методов плоской кластеризации. В простейшем случае (K-Means) требуется задание числа кластеров и начальных положений центроидов кластеров, после чего запускается итеративный процесс, стабилизирующий положение центроидов. На каждом шаге процесса документы приписываются к кластеру с ближайшим центроидом. После того, как все тексты распределены, вычисляется новое положение центроидов. Процесс останавливается, когда центроиды перестают перемещаться, или выполняется критерий остановки. Существуют вариант этого алгоритма, в котором применяется вычисление положений затравочных центроидов при помощи методов Single/Average Link на случайном подмножестве данных. Размер случайной выборки в этом случае выбирается равным , где k – желательное число кластеров, n – число документов. Этот вариант часто применяется поисковыми серверами для кластеризации поисковых запросов.

Применяются также методики, позволяющие оценивать необходимое число кластеров автоматически. Наиболее известной является Minimum Description Length (MDL), применяющаяся в системе AUTOCLASS. Идея этого метода состоит в том, что для каждого кластера и для каждого объекта, входящего в кластер, некоторой процедурой генерируется строка описания. Предполагается, что оптимальное количество соответствует минимальной длине описания.

K-Means, один из самых простых в реализации и, вероятно, самый распространенный из методов кластеризации, является одним из частных случаев общего метода EM. Метод EM оперирует вероятностной моделью отнесения документа к определенному кластеру. Предполагается, что существует k (по числу кластеров) скрытых независимых «генераторов», подчиняющихся многомерному закону нормального распределения. Вектора документов рассматриваются в качестве реализации многомерной случайной величины. Если параметры распределений известны, можно вычислить условную вероятность принадлежности вектора документа к одному из «генераторов».

Самая сильная сторона описанных выше алгоритмов – это линейная зависимость времени их работы от количества документов.

Описанные выше техники кластеризации могут являться базой для построения более эффективных кластеризаторов, будучи скомбинированными с разнообразными методами уменьшения размерности пространства признаков. Эти методы понижают размерность пространства признаков за счет анализа корреляций терминов, позволяя также снимать проблемы синонимии и омонимии без использования тезауруса, так как фактически объединяют в один процесс вариант кластеризации на уровне терминов (которая используется для группировки родственных слов) и кластеризацию на уровне документов. К таким методам можно отнести кластеризацию с помощью суффиксного дерева (STC) и методом латентно семантического анализа (LSA).

* 1. Кластеризация с помощью суффиксного дерева (STC – Suffix Tree Clustering)

Изначально суффиксные деревья – suffix trees - были разработаны и применялись для быстрого поиска подстрок. Суффиксное дерево – дерево, содержащее все суффиксы данной строки. Оно состоит из вершин, ветвей и дополнительных указателей, называемых suffix pointers, с помощью которых добиваются линейной скорости построения дерева. Ветви дерева обозначаются буквами или буквосочетаниями, которые являются частями суффиксов строки. Суффикс, соответствующий определённой вершине, можно получить путем объединения всех букв, которые находятся на ребрах дерева, начиная от корневой вершины и заканчивая данной.

* + 1. Построение дерева

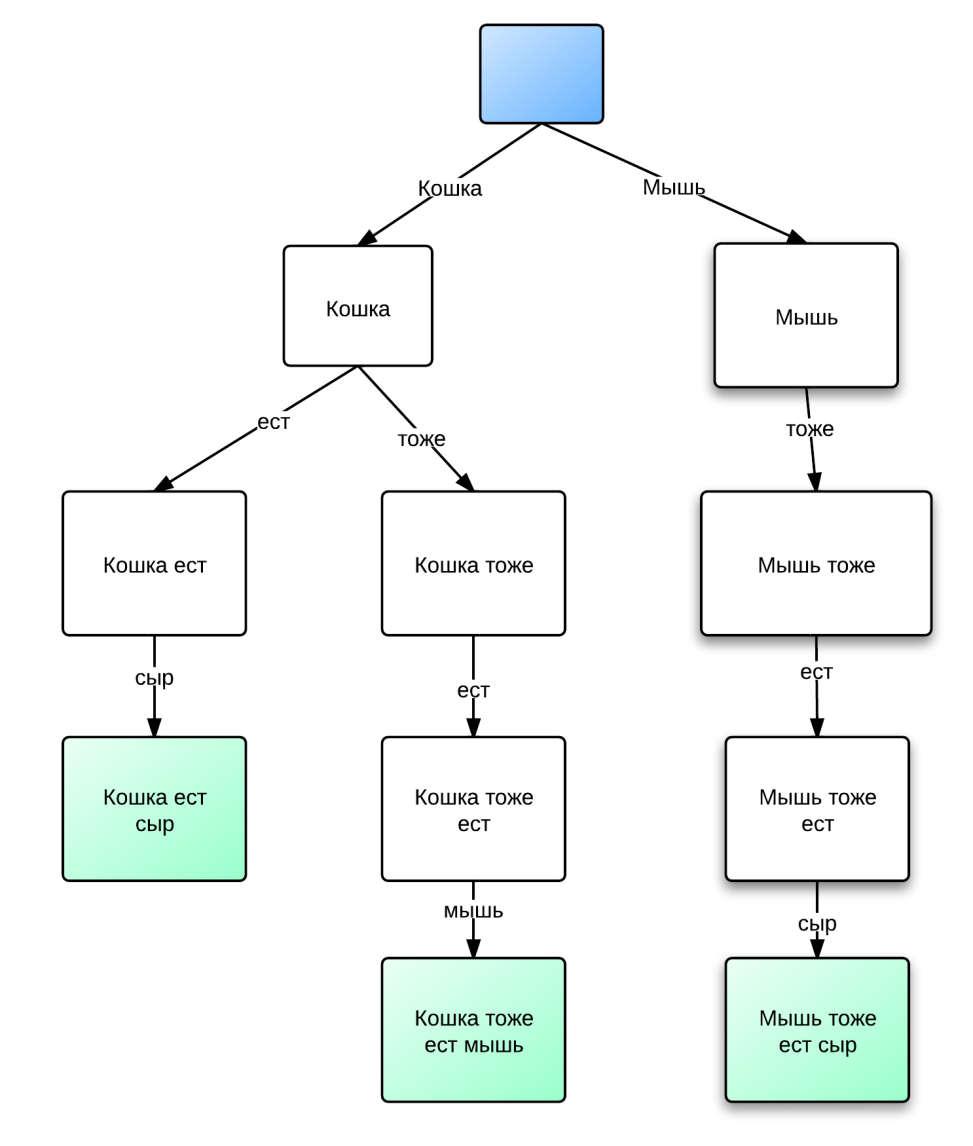
Построение дерева осуществляется поэтапно. Сначала получаемые из сети от поискового сервера документы подвергаются предварительно обработке - отчистка от пунктуации, приведение слов в начальную форму (с помощью алгоритмов морфоанализа или стемминга) и т.д.:

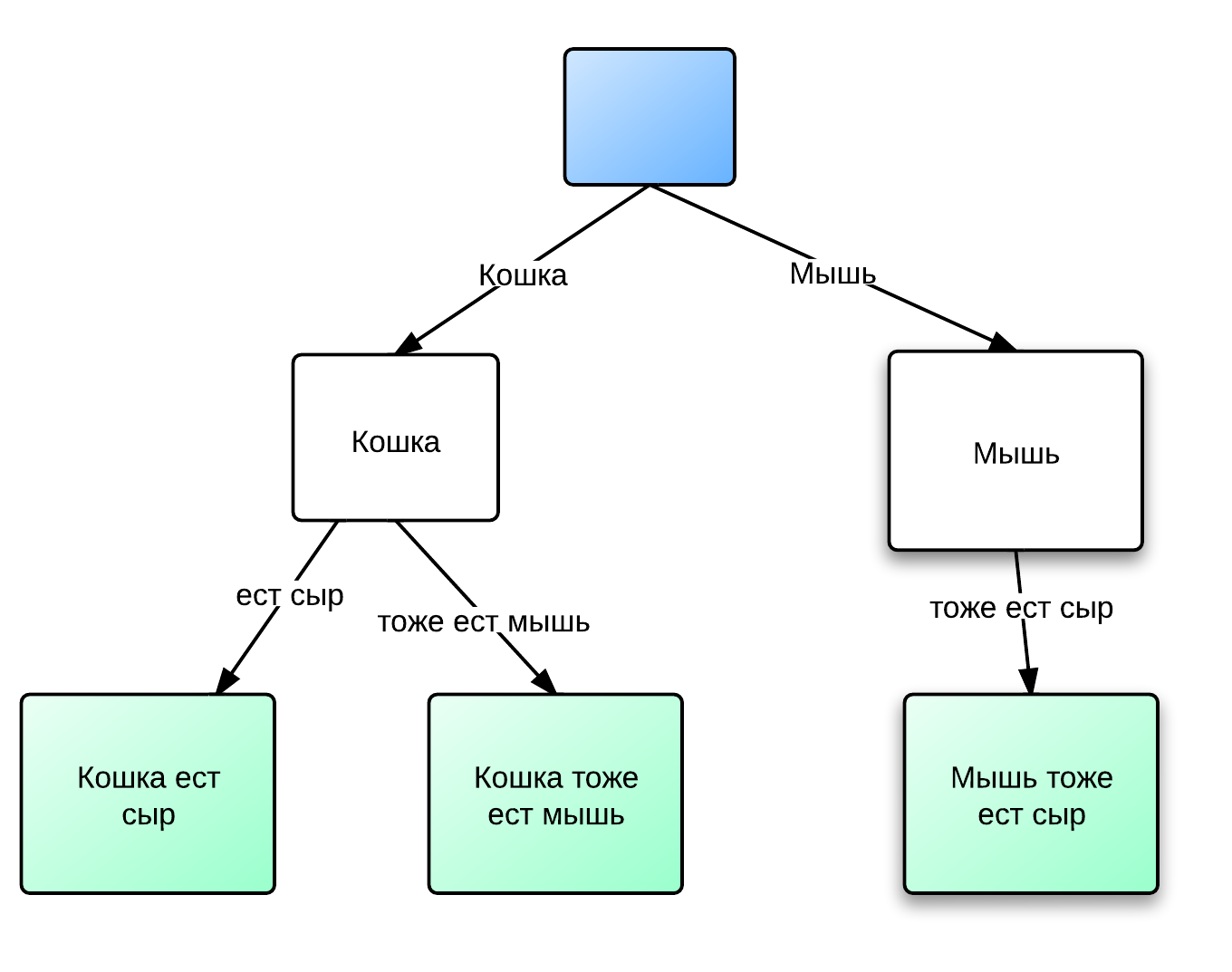
* удаление стоп – слов; (удаление слов, не имеющих смыслового значения: частицы, междометия и т.д.);
* морфологический разбор (приведение слова к нормальной форме: перевод в единственное число, удаление суффиксов);
* синтаксический разбор (выделение связанных компонент (существительное + прилагательное));

Затем для набора документов строится дерево, но единицей, находящейся на рёбрах дерева является слово или словосочетание, а не буква, как в случае с деревьями для строк. Каждой вершине дерева соответствует фраза. Её можно получить, объединив все слова/словосочетания, находящиеся на рёбрах на пути от корня дерева к данной вершине дерева. В тех вершинах дерева, которые имеют потомков, имеются ссылки на документы, в которых встречается фраза, соответствующая вершине. Множества документов, на которые указывают эти ссылки, образуют базовые кластеры.

* + 1. Виды суффиксных деревьев

Различают 2 вида суффиксных деревьев:

* Несжатое суффиксное дерево – бор, составленный из всех суффиксов данной строки (текста)
  1. Несжатое суффиксное дерево для текстов «Кошка ест сыр», «Мышь тоже ест сыр», «Кошка тоже ест мышь».
* Сжатое суффиксное дерево - это некоторая оптимизация несжатого дерева, в котором структуру из нескольких последовательных вершин, выглядящих как односвязный список, заменяют на одно ребро. Каждая из строчек, записанных на рёбрах, является подстрокой исходной строки и, следовательно, задаётся индексом первого и последнего символа.



* 1. Сжатое суффиксное дерево для текстов «Кошка ест сыр», «Мышь тоже ест сыр», «Кошка тоже ест мышь».

После построения суффиксного дерева производится комбинирование (группировка) базовых кластеров и получение набора кластеров.

Комбинируются кластеры по следующей методике:

Пусть Bm и Bn – базовые кластеры, |Bm|, |Bn| - их размеры. |BnBm| - количество общих документов для этих кластеров. Тогда, если: |BnÇBm| / |Bm| > 0.5 и  |BnBm| / |Bn| > 0.5, то базовые кластеры объединяются в один общий, иначе не объединяются.

* + 1. Достоинства и недостатки метода STC

Достоинства метода:

* высокая скорость работы. По времени и занимаемой памяти дерево строится пропорционально количеству документов. Наихудшая теоретическая верхняя граница времени построения - пропорционально квадрату количества документов;
* хорошая наглядность представления результатов. Общие фрагменты текстов и фраз выступают в качестве названия кластеров, – это имеет большой смысл, т.к. не надо затрачивать дополнительных усилий для определения подходящего имени;
* алгоритм не нуждается в обучении и задании порога срабатывания;
* алгоритм инкрементен и допускает пересекаемость областей видимости кластеров;
* если имеется уже построенный индекс по корпусу документов, то можно произвести «арифметизацию», если все тексты документов заменить на некое числовое представление слов, где каждое число ссылается на вход в словаре (или тезаурусе), то сравнивать придётся не слова, а числа, и предварительная обработка текстов становится не нужной.

Недостатки метода:

* необходимость повторной обработки текстов документов;
* не используется уже имеющаяся информация о значения близости документов или значениях tdidf;
* не выявляется скрытая семантика среди документов, которая может присутствовать не только на текстовом уровне;
* проблемы синонимии и омонимии;

Данный метод в гораздо большей степени удовлетворяет поставленным требованиям, т.к. имена для получаемых кластеров создаются автоматически.

* 1. Метод латентно семантического анализа (LSA)

Одним из перспективных методов, позволяющих получать данные о смысле приведенного текста, является метод латентного семантического анализа (ЛСА). ЛСА позволяет выявлять значения слов с учетом контекста их использования путем обработки большого набора текстов. Принцип действия метода заключается в том, что сравнение множества всех контекстов, в которых слова или группы слов употребляются, и контекстов, в которых они не употребляется, позволяет сделать вывод о степени близости смысла этих слов или групп слов.

ЛСА можно рассматривать в двух аспектах:

* как практический прием для получения примерных оценок контекстной связи слов в больших фрагментах по смыслу, либо оценок смысловых корреляций между словом и набором слов (в случае присутствия таких корреляций);
* как компьютерную модель получения и использования знаний человеком, читающим текст.

В качестве практического метода, характеризующего значение слова, ЛСА позволяет измерить корреляции типа «слово-слово», «слово-отрывок» и «отрывок-отрывок». Эти корреляции моделируют механизм мышления человека, сопоставляющего части текста по смыслу. Важно отметить, что результаты, даваемые методом ЛСА, зависят не только от частотности использования слов в отрывках. Метод основывается на выявление более глубоких («латентных») связей и, таким образом, лучше моделирует человеческое восприятия текста, чем простые методы, основанные на частотности употребления слов.

Следует отметить, что у метода ЛСА существуют некоторые ограничения. В нем не используется информация о порядке слов, и, следовательно, метод не учитывает синтаксические отношения, логику или морфологию. Несмотря на это, результаты метода достаточно достоверно отображают смысловые корреляции между словами и отрывками.

***Существуют два основных отличия метода ЛСА от прочих статистических методов обработки текстов:***

* в качестве исходных данных ЛСА использует частоту использования слов в отрывках текста, а не частоту совместного использования слов;
* метод собирает данные не о попарной совместной используемости слов, а об используемости множества слов в большом массиве отрывков.

Таким образом, метод рассматривает влияние выбора, а не порядка слов на смысл отрывка. Можно сказать, что ЛСА представляет значение слова как среднее значений отрывков, в которых оно встречается, а значение отрывка - как среднее значений всех слов, составляющих отрывок.

* + 1. Этапы подготовки данных для метода LSA

Обычно, перед применением метода латентно семантического анализа, исходный корпус документов подвергается предварительной обработке.

Она может включать в себя следующие шаги:

* удаление стоп-символов;
* проведение стемминга слов;
* удаление слов, встречающихся в единственном экземпляре;
  + 1. Основные этапы работы метода LSA

После проведения предварительной обработки текстов составляется частотная матрица индексируемых слов. Данная матрица отражает связи между словами и отрывками.

Следующим шагом после построения частотной матрицы индексируемых слов является ее сингулярное разложение (SVD – Singular Value Decomposition).

Можно заметить, что подавляющее число ячеек частотной матрицы индексируемых слов, созданной на первом шаге, содержат нули. Матрица сильно разрежена, поэтому следующим шагом, обычно, проводится ее нормализация.

1. Конструкторский раздел

Для осуществления кластеризации текстовой коллекции были выбраны 2 метода. Метод латентно семантического анализа (LSA) для построения векторного пространства документов и алгоритм нечеткой кластеризации Fuzzy K-Means для получения результирующих кластеров.

* 1. Метод латентно-семантического анализа (ЛСА)

В данном разделе будет более подробно рассмотрен метод ЛСА, представленный в аналитической части.

Перед применением метода латентно – семантического анализа необходимо провести предварительную обработку корпуса текстовых документов.

Данная обработка включает в себя:

* удаление стоп-символов;
* проведение операции стемминга над словами;
* удаление слов, встречающихся в единственном экземпляре;
  + 1. Удаление стоп - символов

Стоп символы, они же стоп-слова, их еще называют "шумовые слова", это слова, встречающиеся практически во всех текстах и не несущие специальной смысловой нагрузки. Обычно, в большинстве алгоритмов обработки текстов, эти слова удаляются первым шагом. В частности, поисковые машины, вроде Yandex или Google, при индексировании сайтов не обращают на них внимания.

В русском языке, к стоп символам относятся предлоги, суффиксы, причастия, междометия, частицы и т.п.

При реализации программного продукта был составлен список, содержащий стоп – символы русского языка.

* + 1. Операция стемминга

Выделение основы слова это одна из задач, часто возникающая в процессе обработки текста. Русский язык обладает особой сложностью, и если в английском достаточно сделать несколько простых преобразований (вроде удаления ing-ового окончания), чтобы с высокой степенью достоверности получить требуемую основу, то в русском языке эта задача намного сложней. В идеале, она решается путем составления больших морфологических словарей.

Однако существуют алгоритмы стемминга, позволяющие избежать использования дополнительных словарей. Одним из таких алгоритмов является алгоритм Портера.

Главный плюс стеммера Портера заключается в том, что выделение основы осуществляется путем преобразования слова согласно определенным правилам.

Операция стемминга не является обязательной, некоторые источники утверждают, что хорошие результаты получаются и без нее. И действительно, если набор текстов достаточно большой, то этот шаг можно опустить.

* + 1. Удаление слов, встречающихся в единственном экземпляре

Из исходного набора текстовых документов можно также удалить слова, встречающиеся в единственном экземпляре. Данный шаг не влияет на конечный результат и тоже является необязательным, однако он позволяет заметно упросить математические вычисления.

Например, рассмотрим произвольный набор заголовков различных новостей:

1. Британская полиция знает о местонахождении основателя WikiLeaks.

2. В суде США начинается процесс против россиянина, рассылавшего спам.

3. Церемонию вручения Нобелевской премии мира бойкотируют 19 стран.

4. В Великобритании арестован основатель сайта Wikileaks Джулиан Ассандж.

5. Украина игнорирует церемонию вручения Нобелевской премии.

6. Шведский суд отказался рассматривать апелляцию основателя Wikileaks.

7. НАТО и США разработали планы обороны стран Балтии против России.

8. Полиция Великобритании нашла основателя WikiLeaks, но, не арестовала.

9.В Стокгольме и Осло сегодня состоится вручение Нобелевских премий.

После удаление стоп-символов русского языка, проведения операции стемминга и удаления слов, встречающихся в единственном экземпляре, получим следующее (выделено жирным):

1. Британская **полиц**ия знает о местонахождении **основател**я **WikiLeaks**.

2. В **суд**е **США** начинается процесс **прот**ив россиянина, рассылавшего спам.

3. **Церемон**ию **вручен**ия **Нобелевск**ой **прем**ии мира бойкотируют 19 **стран**.

4. В **Великобритан**ии **арестова**н **основател**ь сайта **Wikileaks** Джулиан Ассандж.

5. Украина игнорирует **церемон**ию **вручен**ия **Нобелевск**ой **прем**ии.

6. Шведский **суд** отказался рассматривать апелляцию **основател**я **Wikileaks**.

7. НАТО и **США** разработали планы обороны **стран** Балтии **прот**ив России.

8. **Полиц**ия **Великобритан**ии нашла **основател**я **WikiLeaks**, но, не **арестова**ла.

9.В Стокгольме и Осло сегодня состоится **вручен**ие **Нобелевск**их **прем**ий.

* + 1. Матрица употребляемости

После проведения предварительной обработки текстов составляется частотная матрица индексируемых слов.

ЛСА использует матрицу, которая описывает употребляемость слов в отрывках. Столбцы матрицы соответствуют документам, а строки – словам, встречающимся в документах. Элементы матрицы представляют собой количество употреблений данного слова в данном отрывке. Такой подход стандартен для всех семантических моделей.

Терм – семантическая единица (слово).

Документ – отдельный текст: di = (ti1, ti2, …, tim).

Матрица термы – на – документы: Mmxn = (d1T, d2T,…, dnT).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Документ 1 | Документ 2 | Документ 3 | … | Документ n |
| Терм 1 | 1 | 0 | 0 | … | 1 |
| Терм 2 | 0 | 0 | 1 | … | 0 |
| Терм 3 | 1 | 1 | 0 | … | 1 |
| … | … | … | … | … | … |
| Терм m | 1 | 0 | 0 | 0 | 1 |

* 1. Частотная матрица индексируемых слов.

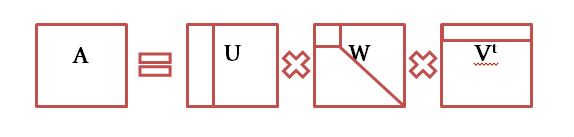
Оригинальная матрица отражает связи между словами и отрывками. При анализе текстов на естественном языке возникает две основные проблемы – синонимия и полисемия:

* синонимия подразумевает, что одно и то же понятие может описываться в естественном языке разными комбинациями слов;
* при полисемии одно и то же слово или комбинация слов могут описывать разные понятия.
  + 1. Сингулярное разложение

Следующим шагом после построения частотной матрицы индексируемых слов является ее сингулярное разложение (SVD – Singular Value Decomposition).

Сингулярное разложение - это математическая операция, раскладывающая матрицу на три составляющих. Т.е. исходная матрица M представляется в виде:

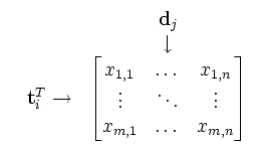
M = U \* W \* Vt,

Где U и Vt – ортогональные матрицы, а W – диагональная матрица. Причем диагональные элементы матрицы W упорядочены в порядке убывания. Диагональные элементы матрицы W называются сингулярными числами.

* 1. Сингулярное разложение частотной матрицы (SVD).

Преимущество сингулярного разложения состоит в том, что оно выделяет ключевые составляющие матрицы, позволяя игнорировать шумы. Согласно простым правилам произведения матриц, верно то, что столбцы и строки, соответствующие меньшим сингулярным значениям, дают наименьший вклад в итоговое произведение. Например, можно отбросить последние столбцы матрицы U и последние строки матрицы Vt, оставив только первые 2. Важно, что при этом гарантируется оптимальность полученного произведения. Разложение такого вида называют двумерным сингулярным разложением. После проведение двумерной декомпозиции (SVD–2) можно легко отметить точки, соответствующие отдельным документам на плоскости.

Рассмотрим математическую модель, лежащую в основе сингулярного разложения матрицы:

Пусть A –матрица, гле элемент (i,j) отображает употребление слова i в отрывке j. Матрица A будет иметь следующий вид:

Элемент xij представляет собой количество употреблений i-го слова в j-том отрывке. Каждая строка в этой матрице – это вектор, соответствующий слову и отражающий его употребляемость в каждом документе.

Аналогично, каждый столбец представляет из себя вектор, cответствующий отрывку и отражающий употребляемость слов в этом отрывке:

Пусть . Тогда – эрмитова неотрицательно определенная матрица:

;

Поэтому все ее собственные значения неотрицательны.

Неотрицательные квадратные корни из собственных значений матрицы A\* A называются сингулярными числами матрицы A. Сингулярные числа (wi = wi(A)) принято нумеровать по не возрастанию:

w1 ≥ w2 ≥ … ≥ wr ≥ wr+1 = … = 0.

Будем считать, что A имеет r ненулевых сингулярных чисел.

Пусть u1 … un - ортонормированный базис собственных векторов матрицы , такой, что

Положим vi = Aui / wi, 1 ≤i ≤r. Тогда (vi, vj) = 0 при i≠ j и (vi,vi) = 1. Дополним систему v1,…vr векторами vr+1,…vm до ортонормированного базиса в Cm. Заметим также, что при j ≥ r +1:

В итоге получаем

где U = [u1, …, un] и V = [v1, …, vn] – унитарные матрицы, а W – диагональная прямоугольная матрица тех же размеров, что и матрица A.

Столбцы матриц U и V образуют сингулярные базисы матрицы A . Столбцы U называются правыми сингулярными векторами, а столбцы V - левыми сингулярными векторами матрицы A . Связь между сингулярными векторами и ненулевыми сингулярными числами устанавливается соотношениями:

, ,

Кроме того,

Итак, мы доказали, что для любой матрицы имеет место равенство:

(\*)

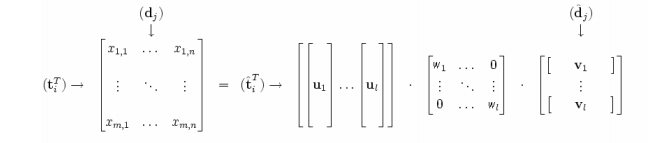
Для некоторых унитарных матриц , и диагональной прямоугольной матрицы размеров mxn с числами wi ≥0 при i = j. Записав (\*) в виде:

(\*\*),

получаем представление матрицы, называемое ее сингулярным разложением.

Если каким-то способом получено разложение (\*\*) с унитарными матрицами U и V, то . Поэтому если W- диагональная прямоугольная матрица с неотрицательными элементами, то ее ненулевые элементы определены однозначно.

Таким образом, разложение имеет следующий вид:



Числа w1, …, wl называются сингулярными числами, а u1, …ul и v1, … vl правыми и левыми сингулярными векторами соответственно.

Если из всех сингулярных значений отобрать k наибольших, то мы получим аппроксимацию исходной матрицы матрицей ранга k.

* + 1. Улучшения, применяемые к частотной матрице

Можно заметить, что подавляющее число ячеек частотной матрицы индексируемых слов, созданной на первом шаге, содержат нули. Матрица сильно разрежена, и это свойство может быть использовано для улучшения производительности и потребления памяти при создании более сложной реализации.

В различных ситуациях, когда длины исходных текстов документов заметно отличаются, частотную матрицу следует нормализовать. Существует множество способов нормализации матрицы, стандартным является способ TF – IDF.

TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

TF (term frequency — частота слова) — отношение числа вхождения некоторого слова к общему количеству слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

,

где ni  есть число вхождений слова в документ, а в знаменателе — общее число слов в данном документе.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов.

,

где – количество документов в корпусе, – количество документов, в которых встречается ti, (когда ni не равно 0).

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов. Таким образом, мера TF-IDF является произведением двух сомножителей: TF и IDF. Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

* 1. Алгоритм Fuzzy K-Means

Алгоритм Fuzzy K-Means (также, иногда его называют Fuzzy C-Means) является расширением популярного и достаточно простого алгоритма кластеризации K-Means. Алгоритм K-Means строит «жесткие» кластеры (каждая точка множества, над которым проводится кластеризация принадлежит строго одному кластеру), в то время как Fuzzy K-Means работает с «мягкими» кластерами, где каждая точка множества может принадлежать одновременно нескольким построенным кластерам с разными весами принадлежности.

Так же, как и алгоритм K-Means, алгоритм Fuzzy K-Means может работать с объектами, представленными в n-мерном векторном пространстве, и для которых определена мера расстояний.

Алгоритм Fuzzy K-Means разбивает исходное множество объектов на k кластеров Sl (l = 1, 2, …, k). Каждый из кластеров Sl ассоциируется с центром кластера Cl. Отношения между объектами и центрами кластеров являются нечеткими, таким образом, величина используется для обозначения степени принадлежности объекта (точки в n-мерном пространстве) Xi  кластеру с центром Cl. Обозначим множество всех объектов как . Алгоритм Fuzzy K-Means (FKM) основывается на минимизации следующей величины, называемой искажением:

В приведенной формуле N – число объектов (точек в многомерном пространстве), m – параметр «нечеткости» кластеризации, k – количество кластеров, а – квадрат Евклидова расстояния между точкой данных Xi  и центром кластера Cj.

Стоит отметить, что величина должна удовлетворять следующему ограничению:

Основная часть работы алгоритма FKM приходится улучшение построенных кластеров путем разбиения заданного набора векторов (координат точек в многомерном пространстве). Алгоритм начинает свою работу с некоторого начально заданного множества кластеров (множества центров кластеров) и повторяет процесс улучшения этих кластеров – построения новых до тех пор, пока не будет достигнут критерий остановки работы алгоритма. Предполагается, что никакие два кластера не имеют центров, расположенных в одной точке.

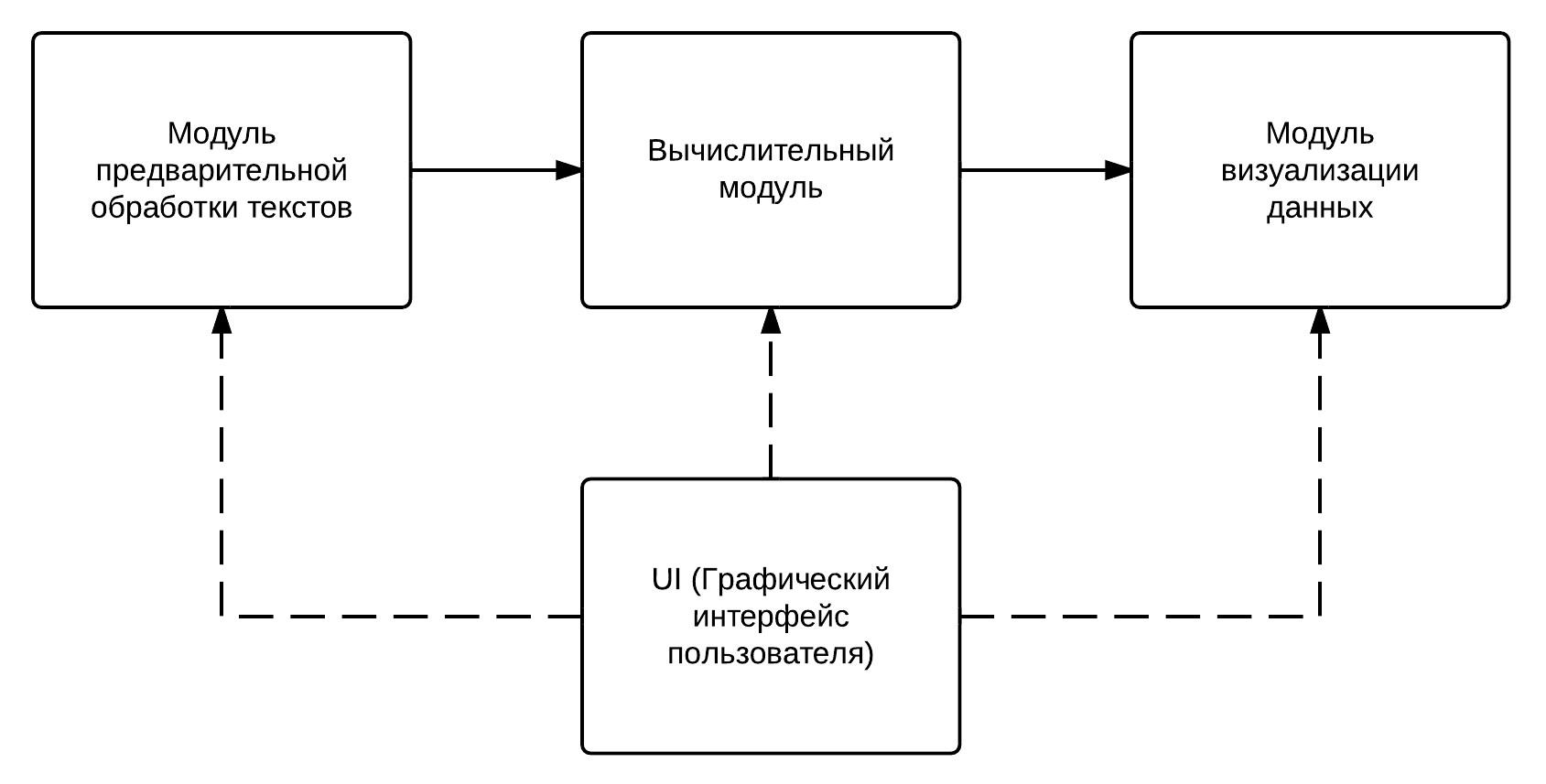
Работа алгоритма Fuzzy K-Means может быть описана следующими шагами:

1. Задается множество начальных центров кластеров , а также величина ε. Устанавливается значение p =1.
2. Для заданного множества центров кластеров вычисляется величина , , . Происходит обновление величины по следующей формуле:

Если < η, то = 1, где η – очень маленькое положительное число.

1. Вычисляется центра для каждого кластера для получения нового множества кластеров
2. Если , завершить работу алгоритма, иначе установить и перейти к шагу 2.
   1. Разработка программного продукта

Программный продукт состоит из нескольких модулей, представленных на рисунке ниже (рис. 2.3).



* 1. Модули программного продукта

Он включает в себя:

* модуль предварительной обработки корпуса текстовых документов и подготовки данных для построения векторного пространства документов;
* вычислительный модуль, выполняющий нечеткую кластеризацию над полученным векторным пространством и присваивающий построенным кластерам текстовые метки;
* модуль визуализации результатов кластеризации – визуализирует построенные кластеры и документы;
* модуль UI – графический интерфейс пользователя, позволяющий взаимодействовать с вычислительным модулем, модулем предварительной обработки данных и модулем визуализации.

На рисунке 2.4 приведена схема черного ящика работы программного продукта:



* 1. Черный ящик программного продукта

На рисунке 2.5 изображена схема IDEF-0 программного продукта:



* 1. Схема IDEF-0 программного продукта

Рассмотрим подробнее каждый из модулей программы.

* + 1. Модуль предварительной обработки текстов

На рисунке 2.6 приведена схема IDEF-0 модуля предварительной обработки текстов:



* 1. IDEF-0 модуля предварительной обработки текстов

На рисунке 2.7 приведена диаграмма классов для данного модуля:



* 1. Диаграмма классов модуля предварительной обработки текстов

Основным классом данного модуля является класс SSpaceLSA. Он содержит методы для проведения предварительной обработки текстов. Для этих целей используются следующие классы:

Класс DocParser предоставляет методы для загрузки текстовых документов. Он реализует интерфейс IParser, который должны реализовывать все классы – загрузчики.

Класс LoadedDocuments используется для хранения в памяти загруженных документов, а также предоставляет методы доступа к документам и работы с ними.

При проведении операции стемминга слов используются класс NoStemming или PorterStemmer. Первый из них является «заглушкой» и используется, если стемминг проводить не нужно, второй представляет из себя реализацию алгоритма Портера.

* + 1. Вычислительный модуль

На рисунке 2.8 представлена схема IDEF-0 для вычислительного модуля:



* 1. IDEF-0 для вычислительного модуля

На рисунке 2.9 представлена диаграмма классов данного модуля:



* 1. Диаграмма классов вычислительного модуля

Основным классом, содержащим в себе методы кластеризации текстовых документов является ClusteringUtils. Он использует класс SSpaceLSA для получения векторного пространства документов.

Для кластеризации векторного пространства используется класс MahoutImpl, содержащий в себе логику нечеткой кластеризации с помощью алгоритма Fuzzy K-Means и использования библиотеки Apache Mahout.

Класс ClusteringNameGenerator содержит методы для присвоения построенным кластерам текстовых имен.

В процессе вычислений используются дополнительные структуры данных – кластеризованный документ и кластер. За хранение этих структур отвечают, соответственно, классы FuzzyClusteredDocument и DocumentCluster.

FuzzyClusteredDocument хранит в себе идентификатор документа, его текстовое содержимое, список кластеров и весов, с которыми данный документ принадлежит к каждому из этих кластеров, а также координаты документа в многомерном пространстве векторов.

DocumentCluster содержит идентификатор кластера, имя кластера, количество документов в данном кластере, список идентификаторов документов, принадлежащих данному кластеру, а также структуру содержащую имена тегов данного кластера и их частоту.

* + 1. Модуль визуализации

На рисунке 2.10 представлена схема IDEF-0 для модуля визуализации:



* 1. IDEF-0 для модуля визуализации

На рисунке 2.11 приведена диаграмма классов для данного модуля:



* 1. Диаграмма классов для модуля визуализации

Основным классом модуля является FuzzyChartController, он содержит основные методы для визуализации результатов нечеткой кластеризации.

Класс CircleDocumentChartImpl представляет собой реализацию визуализации с помощью разработанного метода «документов – кругов».

Класс ChartUtils хранит основные методы для построения всех изображений.

DocumentCircle служит для отображения документа в виде круга, разделенного на секторы. Каждый сектор – это кластер, к которому принадлежит данный документ, размер каждого сектора – пропорционален весу, с которым документ принадлежит данному кластеру.

CircleDocumentScaledView представляет из себя увеличенный документ, выбранный пользователем, отображает увеличенную картинку разделенного на секторы документа, а также содержимое документа и его заголовок.

ClusterCircle – служит для отображения кластера в виде круга. Отображает цвет кластера и количество документов в кластере, размер круга пропорционален количеству документов.

1. Технологический раздел

В данном разделе обосновывается выбор операционной системы, платформы, среды разработки и языка программирования, которые были использованы для создания демонстрационного программного продукта. Также в данном разделе представлено описание графического пользовательского интерфейса.

* 1. Выбор языка программирования и среды разработки

В настоящее время существует множество языков и сред программирования, многие из которых обладают достаточно высокой эффективностью, удобством и простотой в использовании. Для разработки данного проекта использовалась платформа Java и среда Eclipse Indigo (3.7).

Java – универсальный кросс-платформенный объектно-ориентированный язык программирования, разработанный компанией Sun Microsystems. В настоящий момент собственником технологии является компания Oracle Corporation (сделка по приобретению Sun Microsystems компанией Oracle Corporation была оформлена в 2009 году).

Язык Java предоставляет программисту широкие возможности реализации самых разнообразных алгоритмов, обладая высокой эффективностью, гибкостью и универсальностью. Он позволяет в полной мере реализовать концепцию объектно-ориентированного программирования.

На данный момент платформу Java используют более 9 миллионов разработчиков программного обеспечения. Она используется во всех основных сферах деятельности и применяется в самых различных устройствах, компьютерах и сетях. Универсальность, эффективность, совместимость с разными платформами и безопасность технологии Java делают ее идеальным инструментом для сетевых технологий.

Java протестирована, доработана, расширена и проверена специализированным сообществом. В этом сообществе насчитывается более 9 миллионов разработчиков. Это самое большое и самое активное сообщество на планете. Благодаря своей универсальности, эффективности и кросс-платформенной совместимости Java стала незаменимой для разработчиков.

Основные преимущества платформы Java:

* возможность запуска приложений под управлением большинства современных операционных систем;
* высокая надежность и безопасность;
* Переносимость;
* высокая производительность;
* автоматическое управление выделением памяти;
* простые и удобные возможности разработки web-приложений и распределенных сетевых приложений;
* богатые возможности фильтрации ввода/вывода;
* средства создания многопоточных приложений.

Что предлагает Java пользователям:

* экономия средств при внедрении новых и миграции существующих систем за счет унификации платформы и возможности выбора различных решений, поддерживающих технологию Java.
* отсутствие жесткой зависимости от поставщика программного обеспечения.
* использование Java-технологий позволяет компаниям двигаться быстрее, быть более гибкими и более эффективно реагировать на изменяющиеся условия рынка, создавая возможности для получения прибыли.
* никакая другая современная технология не обладает такими широкими возможностями на настольных компьютерах, серверах и мобильных устройствах одновременно с признанием в промышленности и широтой распространения, как платформа Java.

На данный момент существует множество средств разработки на этом языке. Для разработки данного проекта была выбрана среда Eclipse Indigo (3.7). Она зарекомендовала себя своей стабильностью и широкими возможностями. Дизайн среды ориентирован на продуктивность работы программистов, позволяя им сконцентрироваться на разработке функциональности, в то время как Eclipse берет на себя выполнение рутинных операций.

* 1. Требования к вычислительной системе

Разрабатываемая система требовательна к аппаратным ресурсам вычислительной машины, поскольку в процессе функционирования системы необходимо выполнять различные алгоритмы с высокой вычислительной сложностью.

Разрабатываемый продукт тестировался и показывал приемлемые результаты в системе, отвечающей следующим требованиями:

* процессор класса Intel Core i5-2410M 2.3 GHz;
* 8096 MB оперативной памяти;
* операционная система Windows 7 с установленной платформой Java.

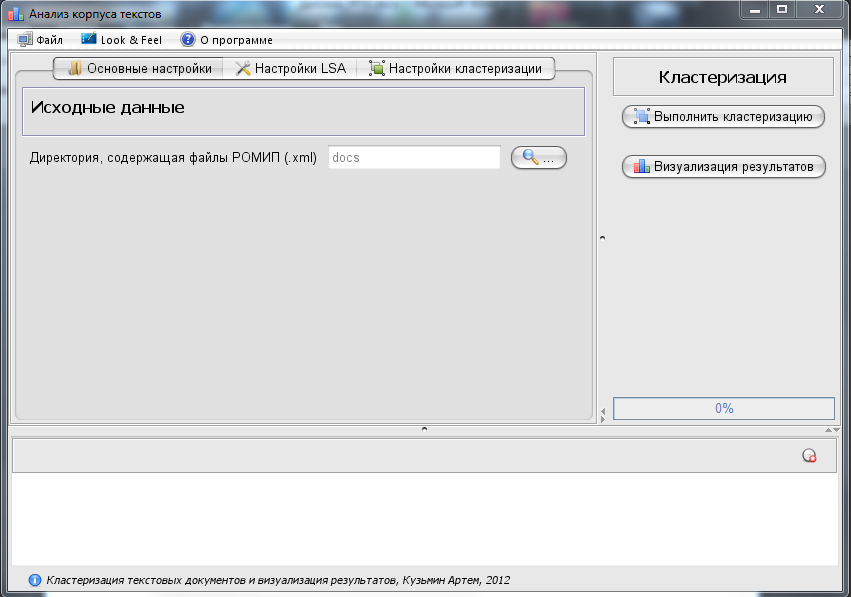
Сама платформа Java, согласно данным, представленным компанией Oracle, предъявляет следующие системные требования:

* поддерживаются процессоры Intel и совместимые с ними на 100% процессоры других производителей;
* рекомендуется к применению процессор Pentium с тактовой частотой 166 МГц и выше;
* оперативная память емкостью не менее 64 МБ;
* минимум 98 МБ свободного дискового пространства.
  1. Пользовательский интерфейс

Программный продукт представляет собой оконное приложение, разработанное для ОС Microsoft Windows. Графический пользовательский интерфейс представляет собой главное родительское окно, в рамках которого открываются все остальные окна, требуемые в процессе функционирования системы.

* + 1. Главное окно программы

На рисунке 3.1 показано главное окно приложения.



* 1. Главное окно приложения

Сверху главного окна приложения располагается панель меню, состоящего из следующих пунктов:

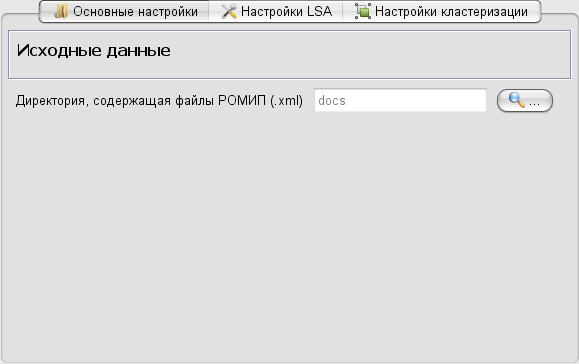
* Файл – общие функции работы с системой;
* Look & Feel – набор стилей интерфейса, позволяет изменять стиль в реальном времени работы системы;
* О программе – информация о программном продукте.

В правой части главного окна приложения располагается меню для проведения кластеризации документов и визуализации результатов.

В нижней части расположена панель для вывода лога работы приложения.

* + 1. Основные настройки приложения

На рисунке 3.2 изображено окно с основными настройками приложения:

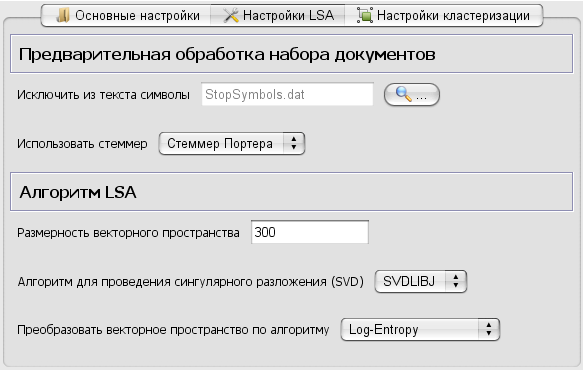


* 1. Основные настройки приложения

Окно содержит пункт для выбора директории, из которой будут загружаться исходные текстовые документы.

* + 1. Настройки LSA

На рисунке 3.3 показано окно с настройками метода латентно – семантического анализа (LSA):

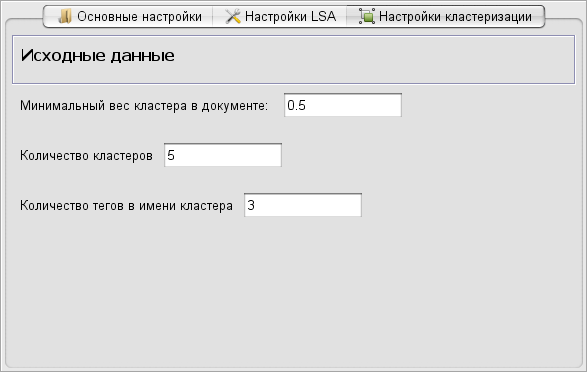


* 1. Настройки LSA

Окно с настройками разделяется на 2 части:

* настройки для предварительной обработки документов: позволяют указать словарь стоп-символов, а также выбрать стеммер для слов;
* настройки метода LSA: позволяют указать размерность получаемого векторного пространства документов, выбрать алгоритм для проведения сингулярного разложения частотной матрицы, а также выбрать метод преобразования полученного векторного пространства.
  + 1. Настройки кластеризации

На рисунке 3.4 показано окно с настройками кластеризации:

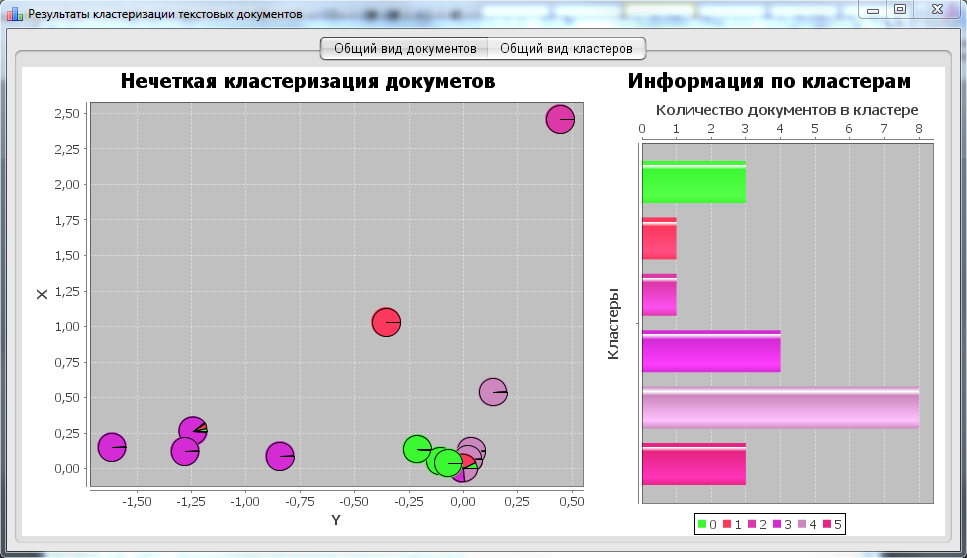


* 1. Настройки кластеризации

Данное окно предоставляет доступ к основным настройками алгоритма нечеткой кластеризации Fuzzy K-Means:

* Минимальный вес кластера в документе – величина, начиная с которой считается, что документ принадлежит кластеру;
* Количество кластеров;
* Количество тегов в имени кластера – используется генератором имен для кластеров.
  + 1. Результаты кластеризации. Общий вид документов

На рисунке 3.5 показано окно с результатами кластеризации текстовых документов, представлен общий вид всех документов:

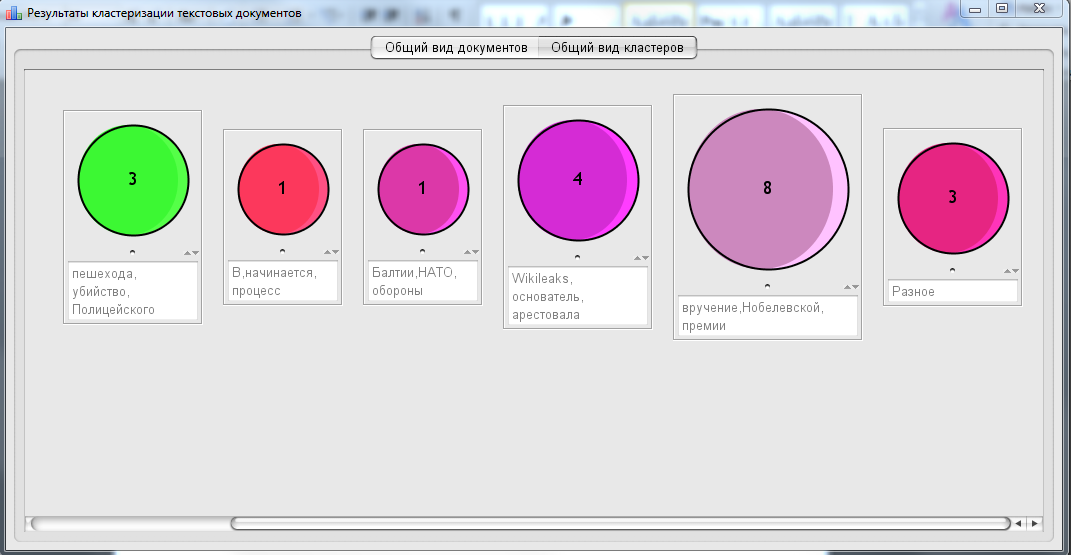


* 1. Результаты кластеризации, общий вид документов

Окно разделяется на 2 части:

* В левой части отображаются документы, над которыми проведена нечеткая кластеризация, документы изображаются разноцветными кругами, каждый цвет соответствует отдельному кластеру. Круги документов разделены на области, размер каждой области пропорционален весу документа в данном кластере.
* В правой части находится общая информация по кластерам, каждый столбец гистограммы – отдельный кластер, длина столбца показывает количество документов в данном кластере, каждый кластер имеет уникальный цвет.
  + 1. Результаты кластеризации. Общий вид кластеров

На рисунке 3.6 показано окно, отображающее результаты кластеризации – общий вид кластеров:

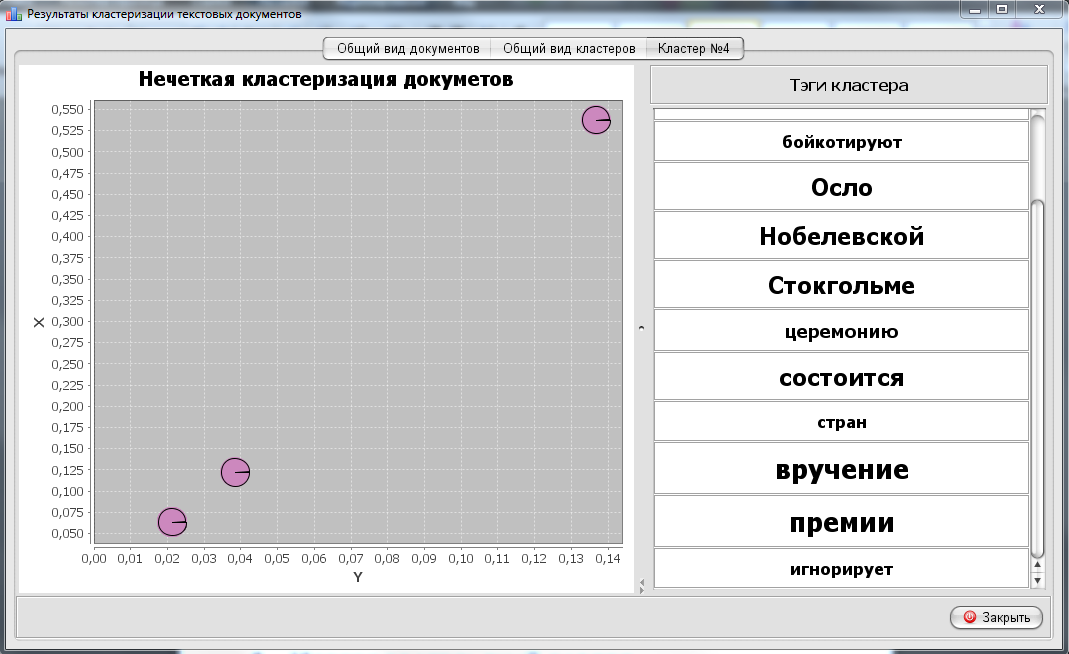


* 1. Результаты кластеризации, общий вид кластеров

На главной панели данной формы располагаются кластеры, в виде разноцветных кругов, каждый кластер имеет свой уникальный цвет. Внутри кластера отображается количество документов, принадлежащих ему. Размер круга пропорционален количеству документов в кластере.

* + 1. Детализированное представление кластеров

На рисунке 3.7 изображена экранная форма с детализированным представлением одного из выбранных кластеров:

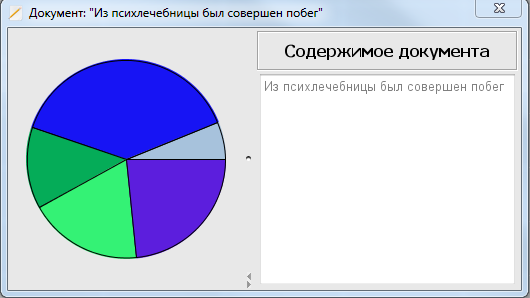


* 1. Детализированное представление кластера

Форма делится на 2 части:

* В левой части располагаются документы, относящиеся к данному кластеру с возможностью выбора документа и более детального просмотра.
* В правой части располагается список тегов кластера. Размер тегов зависит от частоты их появления в документах кластера.
  + 1. Детализированное представление документа

На рисунке 3.8 показано детализированное изображение документа после кластеризации:



* 1. Детализированное представление документа

Окно делится на 2 части:

* В левой части располагается увеличенное изображение документа после нечеткой кластеризации.
* В правой части окна отображается содержимое документа (текст документа).

Заголовок окна соответствует заголовку выбранного документа.

1. Исследовательский раздел
2. Организационно – экономический раздел
3. Промышленная экология и безопасность

Основой проектирования безопасной техники и технологии, а также разработки комплекса мероприятий, обеспечивающего безопасные условия труда на производстве, является учет требований законодательных актов и нормативно-технической документации по охране труда.

* 1. Опасные и вредные факторы при работе с ПЭВМ

Опасный фактор — фактор, воздействие которого на работающего потенциально может привести к травме. Вредный производственный фактор — фактор, воздействие которого на работающего может привести к заболеванию.

Во время работы с ПЭВМ на оператора возможно воздействие следующих опасных и вредных факторов:

|  |  |
| --- | --- |
| **Вредные факторы** | **Опасные факторы** |
| Переменные электрические и магнитные поля | Опасность поражения электрическим током |
| Статическое электричество | Опасность возникновения пожара |
| Негативное воздействие некомфортного микроклимата |  |
| Негативное воздействие шума |  |
| Негативное воздействие недостатка освещения |  |
| Утомление и травматизм кистей рук |  |

* 1. Опасные и вредные факторы
     1. Освещенность

Естественное освещение должно осуществляться через светопроемы, ориентированные преимущественно яна север и северо-восток, и обеспечивать коэффициент естественной освещенности (КЕО) не ниже 1,5% для г. Москвы.

Искусственное освещение должно осуществляться системой общего равномерного освещения. В производственных и административно-общественных помещениях в случае преимущественной работы с документами допускается применение искусственного освещения (и вместе с общим освещением устанавливаются светильники местного освещения).

Освещенность на поверхности рабочего стола должна быть 300-500 лк. Разрешается установка светильников местного освещения для работы с документами, но они не должны создавать блики на поверхности экрана и увеличивать освещенность более 500 лк.

Путем правильного расположения рабочих мест относительно источников освещения должна ограничиваться блёскость от источников освещения.

|  |  |
| --- | --- |
| **Параметр** | **Значение** |
| Яркость светящихся поверхностей, попадающих в поле зрения | < 200 кд/кв.м |
| Яркость бликов на экране ВДТ и ПЭВМ | < 40 кд/кв.м |
| Яркость потолка при применении системы отраженного освещения (Для уменьшения используются матовые неотражающие материалы и защитные колпаки.) | < 200 кд/кв.м |
| Источники света при искусственном освещении | Должны применяться преимущественно люминесцентные лампы типа ЛБ с коэффициентом пульсации не больше 5%. |

* 1. Яркость некоторых объектов
     1. Низкочастотные электрические и магнитные поля

Одним из наиболее вредоносных для здоровья человека факторов при работе с ПЭВМ являются переменные электрические и магнитные поля. Практические все составные части ПЭВМ (системный блок, монитор, клавиатура, дисковые накопители, принтер, сканер) являются источниками электромагнитных полей. Взаимодействие электромагнитных полей различной интенсивности и частоты от разных источников формирует сложную электромагнитную обстановку на рабочем месте оператора.

Продолжительное и систематическое воздействие электромагнитного поля негативно сказывается на здоровье оператора. Наиболее чувствительными к воздействию электромагнитных полей системами организма человека являются: нервная, иммунная, эндокринная и половая. Эти системы организма являются критическими. Биологический эффект ЭМП в условиях длительного многолетнего воздействия накапливается, в результате возможно развитие отдаленных последствий, включая дегенеративные процессы центральной нервной системы, рак крови (лейкозы), опухоли мозга, гормональные заболевания. Особо опасны ЭМП могут быть для детей, беременных (эмбрион), людей с заболеваниями центральной нервной, гормональной, сердечно-сосудистой системы, аллергиков, людей с ослабленным иммунитетом.

|  |  |
| --- | --- |
| **Источник** | **Диапазон частот (первая гармоника)** |
| *Монитор*:  сетевой трансформатор блока питания | 50 Гц |
| *Монитор*:  статический преобразователь напряжения в импульсном блоке питания | 20 - 100 кГц |
| *Монитор:*  блок кадровой развертки и синхронизации | 48 - 160 Гц |
| Монитор:  блок строчной развертки и синхронизации | 15  110 кГц |
| Процессор | 50 Гц- 1000 МГц |
| Другие устройства | 0 Гц - 50 Гц |
| Источники бесперебойного питания | 50 Гц, 20- 100 кГц |

* 1. Частоты электромагнитных полей

В таблице 6.4 приводятся предельно допустимые уровни электрического и магнитного полей в соответствии с СанПиН 2.2.2/2.4.1340-03.

|  |  |  |  |
| --- | --- | --- | --- |
| **Вид поля** | **Диапазон частот** | **Единица измерения** | **ПДУ** |
| магнитное поле | 5Гц- 2кГц | нТл | 250 |
| магнитное поле | 2кГц - 400 кГц, | нТл | 25 |
| электрическое поле | 5Гц- 2кГц | В/м | 25 |
| электрическое поле | 2кГц - 400 кГц | В/м | 2,5 |

* 1. Предельно допустимые уровни электрических и магнитных полей по СанПиН 2.2.2/2.4.1340-03

Стандартом MPRII выдвигаются следующие требования к мониторам (таблица 6.5).

|  |  |  |  |
| --- | --- | --- | --- |
| **Вид поля** | **Диапазон частот** | **Единица измерения** | **ПДУ** |
| магнитное поле | 5Гц- 2кГц | нТл | 200 |
| магнитное поле | 2- 400 кГц, | нТл | 25 |
| электрическое поле | 5Гц- 2кГц | В/м | 25 |
| электрическое поле | 2- 400 кГц | В/м | 2,5 |

* 1. Предельно допустимые уровни электрических и магнитных полей согласно стандарту MPRII

Видно, что требования стандарта MPRII являются более жесткими, чем СанПиН 2.2.2/2.4.1340-03, следовательно, любой монитор, удовлетворяющий MPRII, будет удовлетворять требованиям СанПиН.

Проект предусматривает использование мониторов, удовлетворяющих стандарту MPRII.

* + 1. Статическое электричество

При работе монитора на экране кинескопа накапливается электростатический заряд, создающий электростатическое поле (ЭСтП). Исследования показывают, что люди, работающие с монитором, приобретают электростатический потенциал. Разброс электростатических потенциалов пользователей колеблется в диапазоне от -3 до +5 кВ. Когда ЭСтП субъективно ощущается, потенциал пользователя служит решающим фактором при возникновении неприятных субъективных ощущений.

Стандарт MPRII определяет предельно допустимый уровень электрического потенциала 500В. В соответствии с СанПиН 2.2.2/2.4.1340-03 допустимым является значение 500В. Следовательно, монитор, удовлетворяющий требованиям MPRII удовлетворит и требованиям СанПиН.

* + 1. Воздух рабочей зоны

Одним из необходимых условий здорового и высокопроизводительного труда является обеспечение требуемого газового состава воздуха и комфортного микроклимата в рабочей зоне помещения, т.е. в пространстве до 2-х метров над уровнем пола, где расположены рабочие места.

К нормируемым показателям микроклимата рабочей зоны относятся:

* Микроклимат (температура, влажность, скорость движения воздуха);
* Газовый состав воздуха;
* Ионизация воздуха

Не оптимальные или даже недопустимые параметры микроклимата негативно сказываются на производительности труда операторов и состоянии их здоровья. При воздействии высокой температуры, интенсивного теплового излучения возможен перегрев организма, который характеризуется повышением температуры тела, обильным потовыделением, учащением пульса и повышением частоты дыхания, резкой слабостью, головокружением, а в тяжелых случаях - появлением судорог или теплового удара. Высокая температура оказывает вредное воздействие и на ряд психологических функций, уменьшая объем запоминаемой информации, резко сужая способность к ассоциациям, понижая внимание и т.д. Источником высокой температуры в машинном зале ВЦ являются сами компьютеры, дисплеи, принтеры, а также плохая работа кондиционеров. Влажность воздуха оказывает большое влияние на терморегуляцию организма.

Уровень влажности оказывает важное влияние на терморегуляцию организма. Повышенная влажность, к примеру, снижает скорость испарения пота с поверхности кожи человека, уменьшая таким, образом отвод излишнего тепла от организма.

Скорость движения воздуха также влияет на терморегуляцию организма. Слишком высокая скорость воздуха может привести к ощущению холода.

Газовый состав воздуха определяет, в том числе, и лёгкость дыхания. С атмосферным воздухом в лёгкие человека попадают также различные пары, газы, а также различные жидкие и твёрдые вещества, с ним смешанные. Пары и газы образуют с воздухом смеси, а твердые и жидкие вещества—дисперсные системы, которые делятся на пыль (размер твердых частиц более 1 мкм) и дым (менее 1 мкм).

Допустимыми условиями считаются такие сочетания количественных показателей микроклимата, которые при длительном и систематическом воздействии на человека обеспечивают сохранение нормального функционального и теплового состояния организма, сопровождающееся напряжением механизмов терморегуляции, не выходящим за пределы физиологических приспособительных возможностей.

В таблице 6.6 приведены количественные показатели микроклимата для работы программиста (категория работ – 1а).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Период года** | **Категория работ** | **Температура, С** | **Влажность, %** | **Скорость воздуха, м/с** |
| Холодный | Лёгкая 1а | 21-25 | 75 | 0,1 |
| Тёплый | Лёгкая 1а | 22-28 | 55 | 0,1 |

* 1. Допустимые нормы температуры, относительной влажности и скорости движения воздуха в рабочей зоне производственных помещений

Оптимальными условиями в нашем случае являются:

* Температура от 20оС до 23оС.
* Относительная влажность - 40-60% .
* Скорость движения воздуха не должна превышать 0,1 м/с.

Для обеспечения оптимального соотношения параметров микроклимата в рабочем помещении предполагается:

* использование естественной вентиляции.
* использование системы приточно-вытяжной вентиляции.

Кроме вышеперечисленных параметров важен ещё и уровень ионизации воздуха, т.к. пониженное содержание ионов вызывает сонливость и повышенную утомляемость человека при работе на ПЭВМ. Нормы ионизации согласно СанПиН 2.2.4.1294-03 приведены в табл. 6.7.

Коэффициент униполярности – отношение концентрации аэроионов положительной полярности к концентрации аэроионов отрицательной полярности.

|  |  |  |  |
| --- | --- | --- | --- |
| **Нормируемые показатели** | **Концентрации аэроионов, р (ион/см3)** | | **Коэффициент униполярности, У** |
| **n+** | **n-** | 0,4<У<1,0 |
| Минимально допустимые | 400 | 600 |
| Максимально допустимые | 50000 | 50000 |

* 1. Нормы ионизации

Таким образом, для нормализации параметров микроклимата предлагаются следующие устройства и мероприятия:

* использование естественной вентиляции
* использование системы приточно-вытяжной вентиляции
* использование ионизаторов для приведения в соответствие требованиям СанПиН уровня ионов в воздухе рабочей зоны.
  + 1. Опасность поражения электрическим током

Основным опасным фактором, связанным с электричеством является поражение электрическим током.

Поражение электрическим током может произойти вследствие следующих событий:

* Пробой изоляции.
* Появление напряжения на корпусе.
* Статическое электричество.

Проходя через организм, электрический ток оказывает термическое, электролитическое и биологическое действие.

Термическое воздействие заключается в нагреве тканей и биологических сред организма, что ведет к перегреву всего организма и, как следствие, нарушению обменных процессов и связанных с ним отклонений. Электролитическое воздействие заключается в разложении крови, плазмы и прочих физиологических растворов организма, после чего они уже не могут выполнять свои функции. Биологическое воздействие связано с раздражением и возбуждением нервных волокон и других органов.

Различают два основных вида поражений электрическим током: электрические травмы и удары.

К электротравмам относятся:

* электрический ожог – результат теплового воздействия электрического тока в месте контакта;
* электрический знак – специфическое поражение кожи, выражающееся в затвердевании и омертвении верхнего слоя;
* металлизация кожи – внедрение в кожу мельчайших частичек металла;
* электроофтальпия – воспаление наружных оболочек глаз из-за воздействия ультрафиолетового излучения дуги;
* механические повреждения, вызванные непроизвольными сокращениями мышц под действием тока.

Электрическим ударом называется поражение организма электрическим током, при котором возбуждение живых тканей сопровождается судорожным сокращением мышц.

Основной фактор, обусловливающий ту или иную степень поражения человека, - сила тока.

Для характеристики его воздействия на человека установлены три критерия:

* пороговый ощутимый ток - наименьшее значение тока, вызывающего ощутимые раздражения (0,5... 1,5 мА для переменного тока 50 Гц);
* пороговый неотпускающий ток - значение тока, вызывающее судорожные сокращения мышц, не позволяющие пораженному освободиться от источника поражения (6... 10 мА для переменного тока 50 Гц);
* пороговый фибрилляционный ток - значение тока, вызывающее фибрилляцию сердца (50...100 для переменного тока 50 Гц).

Фибрилляцией называются хаотические и разновременные сокращения волокон сердечной мышцы, полностью нарушающие ее работу.

Напряжение в сети составляет 220В, а частота переменного тока 50Гц, и согласно ПУЭ помещение относится ко II-ой категории, т.е. без повышенной опасности.

Для предотвращения электротравм и снижения опасности поражения электрическим током предлагается применение систем защитного заземления металлических частей ПЭВМ, доступных для прикосновений, которые могут оказаться под напряжением в результате повреждения изоляции. Предлагается использовать заземление выносного типа. В соответствии с ПУЭ сопротивление заземления должно быть не более 4 Ом. Современные розетки имеют заземляющий контакт. Современные корпуса, таким образом, заземляются. Поэтому специальных мер не планируется.

* + 1. Опасность возникновения пожара

Согласно существующей классификации НПБ-105-95 производственное помещение, в котором предполагается использовать систему, по пожарной опасности относится к категории «Д». Причиной возникновения пожара может стать неисправность электрооборудования.

Пожарная безопасность может быть обеспечена мерами пожарной профилактики и активной пожарной защитой. Понятие о пожарной профилактике включает комплекс мероприятий, необходимых для предупреждения возникновения пожара или уменьшения его последствий. Под активной пожарной защитой понимаются меры, обеспечивающие успешную борьбу с возникающими пожарами. На случай пожара необходимо предусмотреть безопасную эвакуацию людей.

Для тушения пожаров в производственном помещении необходимо применять углекислотные и порошковые огнетушители, которые обладают возможностью тушения электроустановок, высокой эффективностью борьбы с огнем.

Исходя из норм пожарной безопасности, в здании с ПЭВМ расположены внутренние средства пожаротушения, такие как пожарные краны, средства первичного пожаротушения, а именно: углекислотные огнетушители типа ОУ-2, с помощью которого можно тушить загорания различных материалов и установок под напряжением. Также в помещении установлена пожарная сигнализация, которая позволяет оповестить дежурный персонал о пожаре.

В качестве пожарных сигнальных датчиков в машинном зале устанавливаются извещатели ДТЛ.

* + 1. Негативное воздействие шума

Одним из наиболее распространенных факторов внешней среды, неблагоприятно воздействующих на организм человека, является шум. Шум вредно действует на организм и снижает производительность труда. Шум создает значительные нагрузки на нервную систему человека, оказывает на него психологическое воздействие. Источниками шума в помещении являются механические устройства и внутренние вентиляторы ЭВМ, а также шум от общеобменной вентиляционной установки, который также выбрасывается и в окружающую среду (на прилегающую к зданию территорию).

Человек, работая при шуме, привыкает к нему, но продолжительное действие сильного шума вызывает общее утомление, может привести к ухудшению слуха, а иногда и к глухоте, нарушается процесс пищеварения, происходит изменение объема внутренних органов. Эти вредные последствия шума тем больше, чем сильнее шум и продолжительнее его действие. Таким образом, шум на рабочем месте не должен превышать допустимых уровней. Значения допустимых уровней шума приведены в таблице 6.8.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Уровни звукового давления в дБ в октавных полосах со среднегеометрическими частотами, Гц** | | | | | | | |
| 63 | 125 | 250 | 500 | 1000 | 2000 | 4000 | 8000 |
| 71 | 61 | 54 | 49 | 45 | 42 | 40 | 38 |

* 1. Допустимые уровни шума

Для помещения ВЦ, в котором использование разработанной системы, уровень шума находится в пределах допустимых значений за счет применения звукопоглощающего материала для обработки стен рабочего помещения и использования активных глушителей шума в вентиляционных каналах.

* + 1. Негативное воздействие недостатка освещения

Правильно спроектированное и выполненное освещение обеспечивает возможность нормальной производственной деятельности. При освещении помещений используют естественное и искусственное освещение. Недостаток естественного света предусматривает применение системы общего освещения. Освещенность на рабочем месте должна соответствовать характеру зрительной работы, т.к. недостаточное освещение приводит к напряжению зрения, преждевременной усталости и ослабляет внимание, а чрезмерно яркое освещение вызывает ослепление, раздражение и резь в глазах. Неправильное направление света на рабочее место может создать резкие тени, блики и дезориентировать работающего. Сохранность зрения человека, состояние его центральной нервной системы и безопасность на производстве в значительной мере зависят от условий освещения.

Естественное освещение должно осуществляться через светопроемы, ориентированные преимущественно на север и северо-восток и обеспечивать коэффициент естественной освещенности (КЕО) не ниже 1.5 Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300-500 лк.

Для общего освещения помещения используются люминесцентные лампы ЛДЦ-40.

* + 1. Утомление и травматизм кистей рук

В качестве основного устройства ввода для ПЭВМ используется клавиатура. Длительная работа на клавиатуре может вызвать значительное утомление пальцев и кистей рук оператора.

Клавиатура должна удовлетворять следующим требованиям:

* исполнение в виде отдельного устройства с возможностью свободного перемещения;
* опорное приспособление, позволяющее изменять угол наклона поверхности клавиатуры в пределах 5-15 градусов:
* высоту среднего ряда клавиш не более 30 мм.
* расположение часто используемых клавиш в центре, внизу и справа, редко используемых - вверху и слева;
* выделение цветом, размером, формой и местом расположения функциональных групп клавиш;
* минимальный размер клавиш 13 мм, оптимальный - 15 мм;
* клавиши с углублением в центре и шагом 19 плюс-минус 1 мм;
* расстояние между клавишами не менее 3 мм;
* одинаковый ход для всех клавиш с минимальным сопротивлением нажатию 0.25 Н и максимальным - не более 1.5 Н;
* звуковую обратную связь от включения клавиш с регулировкой уровня звукового сигнала и возможности ее отключения.
  1. Расчет системы пожарной сигнализации

Системы электрической пожарной сигнализации предназначены для обнаружения пожара и сообщения о месте его возникновения. Ими рекомендуется оборудовать производственные здания и сооружения с производствами категорий А, Б, В и суммарной площадью таких зданий 1000 кв. м. и более, помещения с большими материальными ценностями и пр. Датчиками системы пожарной сигнализации являются тепловые, световые, дымовые, комбинированные извещатели.

* + 1. Выбор типа пожарного извещателя и расчет защищаемой площади

Тип пожарного извещателя для защиты конкретного объекта определяется с учетом двух основных требований:

* соответствие тактико-технических данных условиям применения на данном объекте. На извещатель не должны действовать такие факторы как температура, наличие агрессивных сред, наличие повышенных вибраций и т.д. На рассчитываемом объекте данные показатели соответствуют нормам;
* обеспечение требуемого времени обнаружения пожара на данном объекте.

Для ликвидации ложных срабатываний необходимо размещать пожарные извещатели таким образом, чтобы на них не влияли помехи. Обеспечение требуемого времени обнаружения пожара производится выбором инерционности пожарного извещателя, а также его размещением.

Предельно допустимый радиус действия пожарного извещателя можно определить по формуле:

,

где

* Z – высота установки извещателя, м;
* K – постоянная времени извещателя, мин;
* Qп – порог срабатывания извещателя, 0С
* Qо – максимальная рабочая температура в помещении до возникновения пожара, 0С;
* Ve – линейная скорость распространения пожара, м/мин;
* B – постоянная зависящая от условий горения и пожарной загрузки;
* - допустимое время срабатывания

Величину постоянной B для пожара в начальной стадии в закрытых помещениях можно определить по формуле:

,

где

*  - коэффициент недожога (принимается равным 0.9)
* М – весовая скорость выгорания кг/(м2\*с)
* Q – низшая теплота сгорания Дж/кг
* n – число направлений распространения пожара
* d – ширина очага пожара (в начальной стадии принимается 0.4 м)

Расчет будем проводить для извещателя ДТЛ и горючим материалом будем считать дерево.

Z = 2,5м; K = 0.4; Qп = 70 0С; Qо = 280С; Ve = 1.5 м/мин; = 1 мин.

М = 5 кг/(м2\*с); Q = 10000 кДж/кг; n = 3;

Таким образом:



для  = 1 мин.

Площадь



По расчетным данным построим график зависимости защищаемой площади от допустимого времени обнаружения пожара при горении определенного вещества (рис. 6.9).

* 1. График зависимости защищаемой площади от допустимого времени обнаружения пожара