**Знания-Онтологии-Теории (ЗОНТ-09**

# Метод кластеризации-классификации текстов на основе бинарных классифицирующих таксонов

Чанышев О.Г.

*[1]Институт Математики им. С.Л. Соболева СО РАН (Омский филиал,ул. Певцова, д. 13, г.Омск, 684090, Россия.*
*The Sobolev Institute of Mathematics of Siberian Branch of the RAS (Omsk branch, Omsk, Russia*
fedorov22@yandex.ru

**Аннотация.** *Автоматический анализ естественно-языковых текстов, кластеризация.* Основная *цель: определение пар текстов с максимальной тематической близостью (БКТ) из заданного множества. В качестве признаковых слов выбираются доминанты, являющиеся вершинами вербальных кластеров текста. При определении пересечения признаковых слов учитываются только вершины, имеющие непустые пересечения их кластеров. На приведенном примере кластеризации 160 текстов различных предметных областей показано, что все БКТ принадлежат своим предметным областям.*

**Ключевые слова: кластеризация текстов, бинарные классифицирующие таксоны, кластеры слов**

## 1 Введение

Базовые идеи, относящиеся к методам автоматического реферирования и кластеризации текстов, едва ли не исчерпывающе представлены в монографии Дж. Солтона [1] вплоть до определения кластеров путем коллапсирования пространства с помощью гравитационного притяжения. В современных автоматических методах кластеризации-классификации текстов выделяют два вида: методы категоризации и методы кластеризации [2]. Оба метода в качестве входных данных используют информационно-поисковые образы документов, представляемые в виде множества признаков, характеризующих содержание текста документа. Различие между ними заключается в том, что категоризационные методы распределяют документы по предопределенному набору рубрик, а методы кластеризации таксономируют документы на основе анализа тематической близости между ними и не требуют предварительного экспертного задания множеств документов определенной тематики. В работе [3], на основе обзора современных техник кластеризации и классификации документов, отмечается, что самой большой проблемой геометрических алгоритмов, основанных на введении мер близости между документами, является большая размерность пространства признаков. Для уменьшения размерности применяются различные методы, в частности, используются стоп-листы, а также лингво-статистические и чисто лингвистические методы. Последние включают: использование словарей и тезаурусов для группировки словоформ по нормальным формам и объединения нормальных форм в синонимические группы; использование не отдельных словоформ, а глагольных и именных групп, в том числе и устойчивых словосочетаний. Так или иначе, основной проблемой всех рассмотренных алгоритмов является автоматическое определение слов или словосочетаний, адекватных теме текста. Сами авторы пользуются критериями важности слова, в которых основную роль играет частота слова. Например, вычисление веса слова по методу IFIDF

[2,4]. Однако еще Солтон [1] указывал на недостаточность (при всей его важности) частотного критерия. Следующая по важности проблема (хотя и связанная с первой) – возможное наличие в множествах анализируемых текстов политематических текстов. В этом случае решений два: либо выявить, доминирующую (пусть и слабо) тему, и тогда мы возвращаемся к первой проблеме, либо рассматривать все множество текстов как сеть, вершины которой представлены подмножествами текстов узкой тематики (максимально близкими), а оставшиеся, при наличии какой-то корелляции с первыми, в качестве ребер, связывающих эти вершины. Конечно, это будет динамическая структура, изменяющаяся при поступлении нового множества текстов.

# *2 Метод кластеризации-классификации*

Нашей основной задачей являлась разработка автоматического, по возможности простого и точного, метода выделения из произвольного множества естественно-языковых текстов жанра деловой прозы (в дальнейшем – текстов) монотематических подмножеств. Затем, используя общее множество признаков этих подмножеств, попытаться классифицировать оставшиеся тексты по этим таксонам.

Рассматриваемый далее метод кластеризации-классификации занимает промежуточное положение между синтаксическими и геометрическими методами. Каждый текст представляется набором лексем, которым сопоставлена числовая характеристика их тематической важности (вес) в этом тексте. В дальнейшем определяется не геометрическое расстояние между текстами, а симметричная близость между i-ым j-ым документами, определяемая как сумма несимметричных сумм весов лексем пересечения: веса сначала суммируются отдельно по i-ому и j-ому документам. Таким образом, основная ставка делается на правильность определения тематической роли лексемы.

Подобный подход был использован ранее для кластеризации и классификации текстов по предметным областям на основе автоматически составленных словарей доминантных слов текстов из различных предметных областей [5,6,7]. В последнем случае отнесение текста к той или иной предметной области определялось по максимуму веса пересечения его доминант с доминантами предметной области. Выбор текстов предметной области и ее наименования производился экспертом. Эксперименты рассматривались скорее как доказательство правильности отбора признаковых лексем (ПЛ).

Современная модификация заключается, прежде всего, в еще более жестком отборе признаковых лексем. А именно: оставляются только доминантные лексемы – вершины кластеров лексем (см. ниже). При этом требуется, чтобы пересечение кластеров одинаковых доминант различных текстов было непустым. Состав кластеров контролируется при помощи стоп-словаря бинарных словосочетания (ПЛ - элемент кластера). Он приведен в Приложении. Далее определяются пересечения признаковых лексем предъявленных текстов и веса (близость) этих пересечений, служащих мерами близости. В качестве бинарных классифицирующих таксонов (БКТ) принимаются пары текстов с наибольшим значением близости. Им сопоставляются множества ПЛ пересечения. Каждой паре классифицирующих таксонов могут быть сопоставлены семантические метки, например, пересечение терминоподобных словосочетаний [9] или бинарных сочетаний ПЛ – элемент кластера, вошедших в пересечение (в приведенных примерах используется именно второй вариант). Остальные тексты классифицируются по принадлежности к той или иной паре по максимуму веса пересечения их доминант с доминантами БКТ.

## 2.1.Кластеры доминант.

Поскольку смысл слова определяется, прежде всего, его кластером, т.е. множеством слов, с которыми данное чаще всего встречается в текстах [8], построим кластеры лексем текста на основании анализа их областей существования (область существования лексемы – множество номеров предложений, в которых она встречается). Во избежание разночтений необходимо отметить, что лексема – это цепочка символов текста, ограниченная разделителями и лексемы равны, только если совпадают все символы в одинаковых позициях (лексема и лексемы - разные слова).

Положим:

$T=(t_1,t_2,..,t_i,..,t_N$ – исходное множество текстов,

$L=(l_{i,1},l_{i,2},...,l_{i,k},...l_{i,n}$ – множество лексем *i*-го документа с исключенными лексемами стоп-словаря и числом повторений в различных предложениях текста (частотой) > 1.

Обозначим через $Q_i$ область существования $l_i$ лексемы. Путем анализа пересечений областей существования подразделим лексемы на независимые лексемы связи (связи между предложениями) и атрибутивные. (Если $Q_i \supset Q_j$, то $l_i$–независимая лексема связи (НЛС), а $l_j$- атрибут $l_i$.

Лексемам, принадлежащим множеству НЛС, присваивается вес, равный числу других НЛС, входящих в предложения ее области существования, за исключением первого – ассоциативная мощность ($\psi$).

Частично упорядочим НЛС по убыванию $\psi$. Пронумеруем вложенные последовательности с одинаковыми значениями $\psi$ (группы) натуральным рядом чисел от 1 до R. Лексема будет иметь ранг, равный номеру группы. В случае анализа множества текстов в качестве веса лексемы в тексте принимается величина, обратная рангу. Лексемы со значениями $\psi$ >( $1/2$ )R названы доминантами. Их число не превышает 4% от размера множества $L$ .

В частично упорядоченном множестве НЛС кластеры лексем $(C_i)$ будем определять последовательно по критерию

$K_{i,j} = \dfrac{\rho(Q_i \cap Q_j)}{\rho(Q_i)}$, где ρ – функция, возвращающая размер множества, $i<j$.

Если $K_{ij}>0.5$, будем считать, что $l_j \in C_i$ и из дальнейшего процесса исключается. Атрибутивные лексемы включаются в состав кластеров по определению. Лексему $l_i$ назовем вершиной кластера. Отметим, что можно получить сеть кластеров текста, рассматривая общие неблизкие ($K_{i,j}<=0.5$) лексемы в качестве ребер графа.

## 2.2. Бинарные классифицирующие таксоны.

Положим:

$D_i=(d_{i,1},d_{i,2},..,d_{i,k},..,d_{i,n})$ – множество вершин кластеров i-го документа (ПЛ), частично упорядоченное по убыванию их весов – признаковое множество,

$s_{i,1},s_{i,2},..,s_{i,k},..,s_{i,n}$ – множества элементов кластеров признакового множества лексем,

$W_i=(w_{i,1},w_{i,2},..,w_{i,k},..,w_{i,n})$ – соответствующее множество весов признакового множества i-го текста (вес НЛС равен $1/R$, где $R$ – ее ранг).

$W_i^{sum} = \sum_{k=1}^{n} w_{i,k}$ - сумма весов элементов признакового множества текста, используемая для нормировки.

### 2.2.1.Определение близостей пар документов

Определяются множества пересечений элементов признакового множества документов и нормализованные веса пересечений :

$C_{i,j} = D_i \cap D_j$ ,

причем учитываются только $C_{i,j}$, для которых $\rho(C_{i,j})>1$ и

$((d_{i,k}, d_{j,l}) \in C_{i,j}, d_{i,k} = d_{j,l}) \Rightarrow \rho(s_{ik} \cap s_{j,l}) > 0$ .

$W_{i,j}^{Cross} = (\sum_{k}^{R(Ci,j)} w_{i,k}) / W_i^{sum}$,

$W_{j,i}^{Cross} = (\sum_{k}^{R(Ci,j)} w_{j,k}) / W_j^{sum}$,

$w_{i,k} \in W_i, w_{j,k} \in W_j w_{i,k} \in W_i, w_{j,k} \in W_j$.

$W_{i,j}^{Cross}$ назовем близостью документа $t_i$ к $t_j$, а $W_{j,i}^{Cross}$ - близостью документа $t_j$ к $t_i$.

Далее мы оперируем с симметричной близостью $W_{i \leftrightarrow j}^{cross} = W_{i,j}^{cross} + W_{j,i}^{cross}$ .

### 2.2.2. Определение бинарных классифицирующих таксонов

Пусть $\tau = (t_i, t_j, W_{i \leftrightarrow j}^{cross})$ и $L = (\tau_1, \ldots, \tau_i, \ldots, \tau_n)$ - список триад, упорядоченный по убыванию значения $W_{i \leftrightarrow j}^{cross}$. Список *L,* начиная с первой триады, фильтруем так, чтобы в итоговом списке не оказалось повторяющихся документов (представленных их номерами в порядке поступления на вход классификатора). Для ясности представим алгоритм на интуитивно понятном псевдокоде.

Пусть M – множество номеров документов, не допускающее дублирования элементов, вначале пустое, B – список, также пустой.

*every i:= 1 to ρ(L) do*
*{*
*if (not(member(M, $t_1 \in \tau_i$)) & (not(member(M, $t_2 \in \tau_i$)) then*
*{*
*insert(M,$t_1$);insert(M,$t_2$)*
*put(B,$\tau_i$)*
*}*
*}*

Получаем множество триад *B={τ},* которое и назовем множеством *бинарных классифицирующих таксонов* (БКТ) – бинарных – по числу документов в триаде. Каждому БКТ сопоставим множество пересечений элементов их признаковых множеств:

$D' = ((D_{i1} \cap D_{j1}),(D_{i2} \cap D_{j2}),\ldots,(D_{ik} \cap D_{jk}),\ldots,(D_{in} \cap D_{jn}))$.

#### 2.2.3. Классификация по БКТ

Из пар документов множества *B* и документов из *T*, не принадлежащих документам множества *B (*обозначим через *T'),* образуем новые кластеры. Каждую пара документов из БКТ рассматриваем как новый документ, представленный его признаковым множеством (пересечением признаковых множеств пары составляющих документов). Близости каждого $t' \in T'$ с каждым из B определяем вышеописанным способом (п. 2.2.1.), только вес пересечения не суммируется с весом БКТ. Документ относим к тому из БКТ, значение близости с которым максимально.

## 2.3. Реализация метода

Метод реализован в виде:

а) комплекса программ, анализирующих множества классифицируемых текстов и формирующих итоговую базу данных в виде фактов Prolog,

б) программы - кластеризатора-классификатора, оперирующей с этой базой данных. В настоящее время первый комплекс реализован на UnIcon (расширенная версия Icon), а кластеризатор-классификатор также и на PDC Visual Prolog v.5.2 Personal Edition (Prolog имеет значительно более мощные средства для организации интерфейса и не имеет проблем с кириллицей.

База данных представлена следующими типами фактов:

*dom(<номер_файла>,<доминанта_лексема>,<вес_доминанты>)***
*cluster(<вершина_кластера>,sostav(<тип_элемента>,<элемент_лексема>,<близость>***)*
*sochet(<номер_файла><терминоподобное_словосочетание>)*
*semnm(<номер_файла>,<наименование_текста>)*

Предупреждая возможные вопросы об эффективности реализации отметим, что это не промышленный программный комплекс, а прототип. В данном случае наглядность и простота важнее эффективности. В частности и поэтому предпочтительнее пользоваться фактами или просто данными в символьном виде, а не двоичными (бинарными) файлами.

**Исходя из содержания метода, этот тип фактов можно было бы отбросить, добавив параметр <вес_доминанты> в факт cluster, Но, возможно, параметр пригодится в дальнейшем.

***Напомним, что эта «близость» есть отношение размера области существования элемента кластера к размеру области существования вершины. Области существования представлены множеством предложений вхождения. Параметр пока не используется, но при необходимости может быть использован для коррекции веса ПЛ.

# 3 Эксперимент

## 3.1. Тематические группы текстов, используемые в эксперименте

Тексты (в разное время взяты из Интернета) для эксперимента подбирались так, чтобы, по крайней мере, в них можно было указать группы, пересечение элементов которых в одном кластере недопустимо, иначе метод признается не работоспособным. Такими группами являются: «Политология», тексты которой не могут пересекаться с текстами объединенной тематической группы «Computer Science»: «СУБД», «Искусственный интеллект», «Сетевые операционные системы», «Компьютерная лингвистика». Элементы тематической группы «СУБД» и «Сетевые операционные системы» не могут иметь пересечений с элементами группы «Философия». Очевидно, допустимы комбинации внутри группы «Computer Science». Другие комбинации также возможны, но крайне нежелательны.

Количества документов в предметных областях (ПО) (субъективная классификация)
Сетевые операционные системы 10
СУБД 13
Психология 18
Философия 54
Искусственный интеллект 18
Компьютерная лингвистика 15
Политология 32

Итого – 160.

Для описания результатов экспериментов используются следующие характеристики:

*адекватность* (А) – субъективная характеристика, выражается отношением числа текстов фиксированной тематики в кластере к общему числу текстов в том же кластере;

*суммарная адекватность* (СА) определяется отношением суммы А к числу соответствующих кластеров (*тематическая дробность)*

*полнота в целом* (П) – отношение числа текстов во всех кластерах к входному числу текстов;

*тематическая полнота* (ТП) – отношение числа текстов фиксированной тематики в кластерах к общему числу текстов этой же тематики во входном потоке.

## 3.2.Эталонный эксперимент.

**Иллюстрирует полное разделение текстов ПО «СУБД» и «Психология»**
БИНАРНЫЕ КЛАССИФИЦИРУЮЩИЕ ТАКСОНЫ (вместо БКТ в программе используется англоязычная аббревиатура BCT)
BCT 1
13 Г.М.Ладыженский.Раздел 2. Сервер базы данных
15 Г.М.Ладыженский. Раздел 4. Обработка транзакций.
Близость=0.348, Сочетания [СУБД (реляционных, современных, большинстве), сервера (активного, клиента), данных(базами, баз, базу, базы, база, базе), sql (языка),системы(операционной)]
BCT 2
6 С. Д. Кузнецов. Введение в СУБД часть 5
8 С. Д. Кузнецов. Введение в СУБД.Часть 7.
Близость=0.344, Сочетания [журнала(буфера, буфер), память(внешнюю), сбоя(мягкого, жесткого, моменту), транзакции (откат, отката, начала, конце), данных(базой, состояние, восстановления, баз, базу, базы, состояния, база, базе), страниц (буферов), памяти (оперативной), захваты(синхронизационные), транзакция(а)]
BCT 3
3 С. Д. Кузнецов. Введение в СУБД: ЧАСТЬ 3
5 С. Д. Кузнецов. Введение в СУБД: ЧАСТЬ 4.
Близость=0.332, Сочетания [отношения (схемы), отношений(схем), данных (модели, баз, реляционных, базы, модель)]
BCT 4
12 Г.М.Ладыженский. Введение.
14 Г.М.Ладыженский. Раздел 3. Обработка распределенных данных

Близость=0.320, Сочетания [СУБД(современных), систем(информационных), данных (базой, базами, баз, базу, базы, база, распределенных, базе, тиражирования, data, обработки)]
ВСТ 5
2 С. Д. Кузнецов. Введение в СУБД: Часть 2.
11 С. Д. Кузнецов. Введение в СУБД. Часть 9.
Близость=0.298, Сочетания [систем (информационных), данных(базами, баз, базы, база, бд, восстановления)]
ВСТ 6
7 С. Д. Кузнецов. Введение в СУБД. Часть 6.
10 С. Д. Кузнецов. Введение в СУБД. Часть 8.
Близость=0.178, сочетания [данных (базами, баз, базы, базе), памяти(оперативной, внешней)]
ВСТ 7
4 И.Смирнов, Е.Безносюк, А.Журавлёв. ПСИХОТЕХНОЛОГИИ.
9 Т. П. Пушкина. Медицинская психология.
Близость=0.133, Сочетания [мозга (головного), состояние(функциональное), больных(неврозами, шизофренией), деятельности(психической)
ИТОГОВЫЕ КЛАСТЕРЫ
---- Кластер 1 ----
13 Г.М.Ладыженский.Раздел 2. Сервер базы данных
15 Г.М.Ладыженский. Раздел 4. Обработка транзакций.
---- Кластер 2 ----
6 С. Д. Кузнецов. Введение в СУБД часть 5
8 С. Д. Кузнецов. Введение в СУБД.Часть 7.
---- Кластер 3 ----
5 С. Д. Кузнецов. Введение в СУБД: ЧАСТЬ 4.
3 С. Д. Кузнецов. Введение в СУБД: ЧАСТЬ 3
---- Кластер 4 ----
12 Г.М.Ладыженский. Введение.
14 Г.М.Ладыженский. Раздел 3. Обработка распределенных данных
---- Кластер 5 ----
2 С. Д. Кузнецов. Введение в СУБД: Часть 2.
11 С. Д. Кузнецов. Введение в СУБД. Часть 9.
---- Кластер 6 ----
10 С. Д. Кузнецов. Введение в СУБД. Часть 8.
7 С. Д. Кузнецов. Введение в СУБД. Часть 6.
1 С. Д. Кузнецов. Введение в системы управления базами данных
---- Кластер 7 ----
4 И.Смирнов, Е.Безносюк, А.Журавлёв. ПСИХОТЕХНОЛОГИИ.
9 Т. П. Пушкина. Медицинская психология.
Все характеристики равны 1 (за исключением тематической дробности).

## 3.3.Основной эксперимент

**Вход**: 160 текстов всех указанных предметных областей.
**Результат.**
*1. БКТ*. Значения близости и состав пересечения кластеров ПЛ.

1) 0.546 [сознание(схватывает, конституирует, формирует, подвергается, чистое, предметно), логических(томе, исследований, исследованиях), феноменологической(редукции, результате), сознания(определение, явления, смыслоформирование), гуссерля(эдмунда, идей), понятие(нуль, времени, восприятия), понятия(число),содержание(фазе, предметное, пережитое), мир(объективный, противоположен, жизненный), предмета(существование, схватывание)]
2) 0.359 [знания(новые),знаний(базы)]
3) 0.348 [субд(реляционных, современных, большинстве), сервера(активного, клиента), данных(базами, баз, базу, базы, база, базе), sql(языка), системы(операционной)]
4) 0.344 [журнала(буфера, буфер), память(внешнюю), сбоя(мягкого, жесткого, моменту), транзакции(откат, отката), начала(конце), данных(базой, состояние, восстановления, баз, базу, базы, состояния, база, базе), страниц(буферов), памяти(оперативной), захваты(синхронизационные),транзакция (а)]

5) 0.337 [система(операционная), систем(операционных), времени(процессорного), ос(выполнять), windows(nt), системы(операционные, операционной)]
6) 0.336 [режиме(защищенном), памяти(оперативной),сетевых(адаптеров), системы(операционной), управления(памятью), систем(файловых, операционных), данных(баз, база), netware(сервере), доступа(несанкционированного), windows(nt), систему(файловую)]
7) 0.332 [отношения(схемы), отношений(схем), данных(модели), баз(реляционных, базы, модель)]
8) 0.331 [знаний(представления), системы(экспертные)]
9) 0.327 [система(операционная, файловая), файла(открытии, имени), времени(разделения), памяти(оперативной, физической, виртуальной), системы(файловые, операционные, операционной, файловой), файлы(обычные, специальные), память(оперативную), процессов(планирования)]
10) 0.320 [субд(современных),систем(информационных), данных(базой, базами, баз, базу, базы, база, распределенных, базе, тиражирования, data, обработки)]
11) 0.302 [мира(объективного), сознание(общественное), сознания(общественного)]
12) 0.298 [систем(информационных), данных(базами, баз, базы, база), бд(восстановления)]
13) 0.284 [слов(словосочетаний), слова (словосочетания)]
14) 0.256 [философской(мысли), философии(древнегреческой), школы(милетской)]
15) 0.256 [общества(структуру), отношений(рыночных), условиях(современных), жизни( общественной, образ), общественного(богатства), производства(материального)]
16) 76 97 0.254 [познания(процесса), знания(истинного)]
17) 0.248 [философии(классической), созерцания(формами), мысли(философской)]
18) 0.245 [философии(диалектико-материалистической), человека(природы)]
19) 0.245 [система(файловая), данных(баз, базы), информации(хранении), системы(файловые, файловой)]
20) 0.243 [культуры(национальной), открытий(возможных), ценностей(культурных)]
21) 0.241 [государства(защиты), мер(полицейских)]
22) 0.219 [революции(перманентной), ленина(смерти)]
23) 0.213 [природы(естественной), человека(сущность)]
24) 0.195 [философской(мысли), жизни(смысл), человека(общества)]
25) 0.178 [данных(базами, баз, базы, базе), памяти(оперативной, внешней)]
26) 0.173 [философии(средневековой, античной), мир(создан)]
27) 0.169 [элиты(интеллектуальной), современного(мира)]
28) 0.159 [общества(природы), природы(общества)]
29) 0.152 [общества(развития), социальных(групп)]
30) 0.148 [рассудка(чувственности), чувственности(рассудка)]
31) 0.141 [интеллекта(искусственного), человека(мышления)]
32) 0.133 [мозга(головного), состояние(функциональное), больных(неврозами, шизофренией) деятельности(психической)]
33) 0.130 [переворота(государственного), брюмера(го)]
34) 0.086 [переворота(государственного), правительство(свергнуть)]
35) 0.060 [пользователей(удаленных), данных(базы), системы(операционные, операционной)]
36) 0.057 [сети(нейронные), интеллект(искусственный)]
37) 0.027 [система(операционная), систем(операционных)]

**2. Состав предметных областей кластеров**
1 (Философия, Философия);
2 (Искусственный интеллект, Компьютерная лингвистика, Компьютерная лингвистика);
3 (СУБД, СУБД);
4 (СУБД, СУБД);
5 (Сетевые операционные системы, Сетевые операционные системы); 6 (Сетевые операционные системы, Сетевые операционные системы);
7 (СУБД, СУБД );
8 (Искусственный интеллект, Искусственный интеллект, Искусственный интеллект);
9 (Сетевые операционные системы, Сетевые операционные системы);
10 (СУБД, СУБД);
11 (Психология, Философия, Философия);
12 (СУБД, СУБД);
13 (Компьютерная лингвистика, Компьютерная лингвистика, Компьютерная лингвистика);
14 (Философия, Философия);

15 (Философия, Философия, Философия);
16 (Философия, Философия);
17 (Философия, Философия, Философия, Философия, Философия, Философия);
18 (Философия, Философия, Философия, Философия, Философия, Философия );
19 (Сетевые операционные системы, СУБД, Искусственный интеллект, Компьютерная лингвистика)
20 (Философия, Политология, Политология, Политология);
21 (Политология, Политология);
22 (Политология, Политология, Политология);
23 (Философия, Философия);
24 (Психология, Философия, Философия, Философия, Философия);
25 (СУБД, СУБД);
26 (Философия, Философия, Философия, Философия);
27 (Политология, Политология);
28 (Философия, Философия);
29 (Философия, Философия);
30 (Философия, Философия);
31 (Искусственный интеллект, Искусственный интеллект, Искусственный интеллект)
32 (Психология, Психология);
33 (Политология, Политология, Политология)
34 (Политология, Политология, Политология);
35 (Сетевые операционные системы, Компьютерная лингвистика);
36 (Искусственный интеллект, Искусственный интеллект);
37 (Сетевые операционные системы, Сетевые операционные системы).

3. **Общее число документов ПО в кластерах, ТП и СА**

СУБД 13 (ТП=1, СА=0.89 (6.25/7)), Сетевые операционные системы 10 (ТП=1, СА=0.79(4.75/6)),
Философия 40 (ТП=0.74, СА=0.85(11.87/14)),
Искусственный интеллект 10 (ТП=0.56,СА=0.71(3.55/5)), Политология 16 (ТП=0.5, СА=0.96(5.75/6)),
Психология 4 (ТП=0.22, СА=0.5(1.5/3)), Компьютерная лингвистика 7 (ТП=0.47, СА=0.58(2.32/4))
П=0.625

### 3.3.1. Обсуждение результата.

1. 35 из 37 БКТ определены тематически верно. Во всяком случае, адекватность каждого БКТ равна 1, за исключением БКТ 19 и 35.

*ВСТ 19*
*С. Д. Кузнецов. Введение в системы управления базами данных*
*Н.А. Олифер, В.Г. Олифер Сетевые операционные системы. 3. Управление распределенными ресурсами.*
*ВСТ 35*
*Н.А. Олифер, В.Г. Олифер. Сетевые операционные системы. 10. Обзор сетевых операционных систем*
*Концепция построения системы управления документами*

2. Поскольку главной задачей было выявление максимально тематически близких документов, то полученные значения П и ТП (не всегда высокие) вполне объяснимы.

3. Если считать, что предлагаемый метод верен, то нет ничего неожиданного в высоких значениях ТП и СА для тематических групп «СУБД», «Сетевые операционные системы», «Философия»: тексты этих групп естественно близки. Например, в качестве текстов тематики «СУБД» используются, ставшие уже классическими, лекции С. Д. Кузнецова и Г.М. Ладыженского. «Сетевые операционные системы» представлены 10 главами монографии Н. А. Олифер, В. Г. Олифер того же наименования. Очень высокий показатель СА для текстов тематики «Политология» объясняется их резкой тематической отграниченностью от других текстов. Это - Роже Генон «Кризис современного мира», главы 1-10, Николай Трубецкой. «Европа и человечество», главы 1-6 и Курцио Малапарте. «Техника государственного переворота» части 1-16.

4. Неприятной неожиданностью стали низкие значения СА групп «Компьютерная лингвистика» и «Психология», несмотря на то, что соответствующие ВСТ определены правильно:

*ВСТ 13*

*Антонов А.В. Автоматическое определение тематики большого необработанного текстового массива.*
*Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. Глава 6. Семантико-синтаксический анализ и синтез текстов.*
*ВСТ 32*
*122 Т. П. Пушкина. Медицинская психология.*
*127 И.Смирнов, Е.Безносюк, А.Журавлёв. Психотехнологии.*
Хотя трудно назвать случайным (если судить по наименованиям) объединение документов в
*Кластер 11*
*ВВЕДЕНИЕ В ФИЛОСОФИЮ. Сознание, его сущность и генезис.*
*ВВЕДЕНИЕ В ФИЛОСОФИЮ. Духовная жизнь общества.*
*Первушина О. Н. Общая психология. Процессы психического регулирования.*

В работе [2] указывается на возможность кластеризации по вхождению наиболее высокочастотных слов текста в наименование. Но это рискованный способ, поскольку нельзя ожидать от автора следования неким, даже естественным, «правилам» образования наименований.

4. Ряд проведенных нами экспериментов показывает, что наиболее сильное влияние на адекватность (например, по сравнению с варьированием формулы расчета близости) оказывает состав стоп-словаря сочетаний признаковых лексем с элементами их кластера. Этот факт только подтверждает чрезвычайную важность правильности отбора ПЛ.

# 2 Выводы

Поскольку у подавляющего большинства БКТ (35 из 37) адекватность =1, то есть все основания считать, что основная цель достигнута.

Полагаем, что определяющим фактором в достижении цели явился метод отбора признаковых лексем как вершин вербальных кластеров текста и требование непустого пересечения кластеров одинаковых признаковых лексем в различных текстах при определении их тематической близости.

# ПРИЛОЖЕНИЕ

**Стоп-словарь бинарных словосочетаний.**
время(настоящее, долгое, последнее, длительное, ближайшее, хорошее, первое), времени(течение, настоящему, настоящего, недавнего), мир(окружающий), мира(окружающего), возможность(дает), точки(зрения), зрения(точки), зрения(точка), точка(зрения), место(занимает), операция(выполнена), времени(среднего, промежуток, промежутки), количество)(большое, максимальное), качестве(основы), развитие(дальнейшее), начало(положили), работы(коллективной), значений(убывания), предмет(сделать), процесса(нового, контекст), друг(друга, другу), идет(речь), речь(идет), степени)(высшей), высшей(степени), объект(соответствующий, измененный, изменяет), объектов(соответствующих), данных(явно), операции(выполняются, выполнением), уровня(разного), выполнения(совместного), существует(множество), определенных(взглядов), роль(играет, немалую, важную, роль, большую, сыграли), немалую(роль), выше(отмечено, сказали, сказанное), сознания(дает), свойствами)(различными), новые(ставит), ставит(новые), развития(дальнейшего), права(равные), равные(права), очередь(первую),первую(очередь), состоит(идея), идея(состоит), проблем(множество), проблемы(решаются), вопрос(главный), вопрос(основной), сказать(хотите), хотите(сказать), степени(большей), большей)(степени), правильно(определить), характер(носит), носит(характер), методы(направлены), значимости(максимальной), максимальной(значимости), уровне(имеющемся), имеющемся(уровне), результаты(сравнивать), сравнивать(результаты), являются(сложными), возможности(описали, давало), условиях(идентичных), идентичных(условиях), имеет(конечную), имеют(конечную), авторы(объясняют, отмечают, (ссылаются), наличие(постулировать),значительно(слабее), средства(совершенные), модели(вышеописанной), вышеописанной(модели), влияние)(оказывают, существенное, значительное), данных(литературных), литературных(данных), место(имевшим, имевшего),задач(поставленных, решаемых), редким(исключением), исключением(редким), использование(широкое), широкое(использование), выше(описанных), степени(умеренной), умеренной(степени), интерес(наибольший), внимание(уделено), вопрос(открытым), задачи(поставленные), понять(пытаются), пытаются(понять), часть(большая, значительная), значительная(часть), большая(часть), работы(тщательной),

имеет(дополнительные), речь(шла, идет), позволяет(увеличить), средства(появились, новые), новые(средства), работы(улучшению), набор(широкий), примеры(рассматриваются), имеет(вид), базы(являются), система(предназначена), день(сегодняшний), модели(полученной), номер(отд), этапы(основные), имеет(побочных), основе(анализа), модель(данная), точки(показанные), подобных(таблиц), основные(понятия), позволяет(эффективно, реализовывать), используются(слабо), сотр(поле), отдела(численность), сотрудников(списки), число(указанное), система(основана), операции(входят), отношение(данное), размер(превосходит), размер(используемой), отношение(указанное), отношений(указанных), число(одинаковое), реализации(способ), основная(характеристика), характеристика(основная), субд(большинства), название(получила), номер(руководителя), роль(пассивная), один(человек), пользователей(чрезвычайно), изменения,(вносит), человека(определяется, лице), решение(целого), философии(шло), представители(видные), имеет(дело), человека(представлял), целого(комплекса), людей(позволяет), имеет(сложную), влияние(оказывать), основой(являющиеся), роль(особая), дело(идет), общий(известный), сходства(меньшего), народа(данного), народ(перестает), мере(меньшей, крайней), проявления(дает), момент(определенный), современные(исследователи), принято(считать), людей(подобных), смысле)(широком), привести(конкретный), логически(вытекающее), философия(вынуждена), конца(логического), подобных(случаях), единственной(целью), день(следующий), борьбу(вели)

# 4. Литература


1. Солтон Дж. Динамические библиотечно-информационные системы. М.:Мир, 1979.
2. Пескова О.В. Автоматическое формирование рубрикатора полнотекстовых документов.\\ *Труды 10-й Всероссийской научной конференции Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2008, Дубна, Россия, 2008.*
3. Киселев М. В. Пивоваров В. С. Шмулевич М. М. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики. //http://company.yandex.ru/grant/2005/10_Kiselev_102930.pdf
4. Агеев М.С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов. Диссертация на соискание ученой степени кандидата физико-математических наук. Москва. 2004//*http://www.cir.ru/docs/ips/publications/2005_diss_ageev.pdf*
5. О.Г.Чанышев. Ассоциативная модель реального текста и ее применение в процессах автоиндексирования. //*Труды Седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. М.: Физико-математическая литература, 2000, с. 430-438.*
6. Чанышев О.Г. Критерий близости документов и кластеризация. // *Математические структуры и моделирование: Сб. научн. тр. /Под ред. А.К. Гуца, Омск: ОмГУ, 2001. - Вып. 8. с. 111-120.*
7. Чанышев. Автоматическая классификация текстов по доминантным лексемам. *VII Международная конф. по электронным публикациям EL-Pub2002. Новосибирск, 2002 г. //http://www.sbras.ru/ws/list_doc.dhtml?ru+45+0+40*
8. Bookstein, A., Klein,S.T. Clumping Properties of Content-Bearing Words. *JASIS,2,1988*
9. Чанышев О.Г. Автоматическое извлечение доминантных словосочетаний// *Материалы Всероссийской конференции с международным участием "Знания-Онтологии-Теории" (ЗОНТ-07), 14-11 сентября 2007 г.,Т.1, стр.236-246, Новосибирск 2007.*