Национальный Исследовательский Университет Высшая Школа Экономики

Курс “Практические задачи анализа данных”
Раздел курса: *Поиск закономерностей в данных (Pattern Mining)*.
Тема: *Частые множества признаков и ассоциативные правила*.
Домашнее задание №4

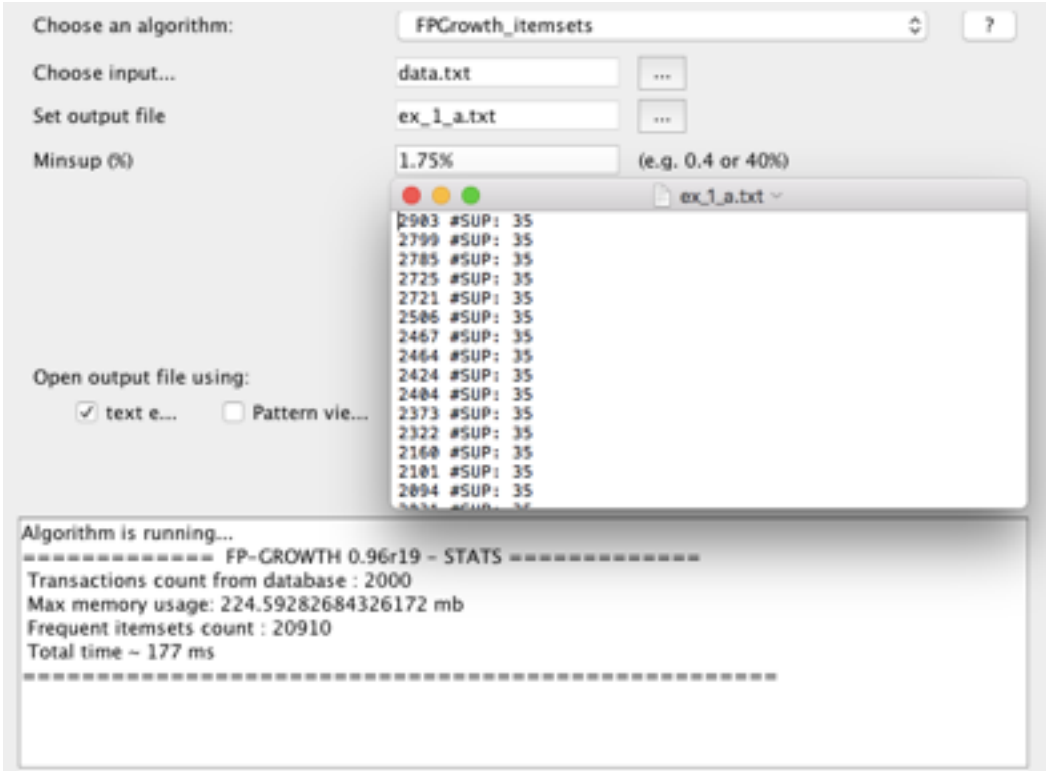
Выполнили:
Булгаков Д. М.
Тефикова А. Р.
Группа: ИАД-2
Преподаватель:
Черняк Е. Л.

Москва 2017

Задание 1 (3 балла). Поиск частых множеств

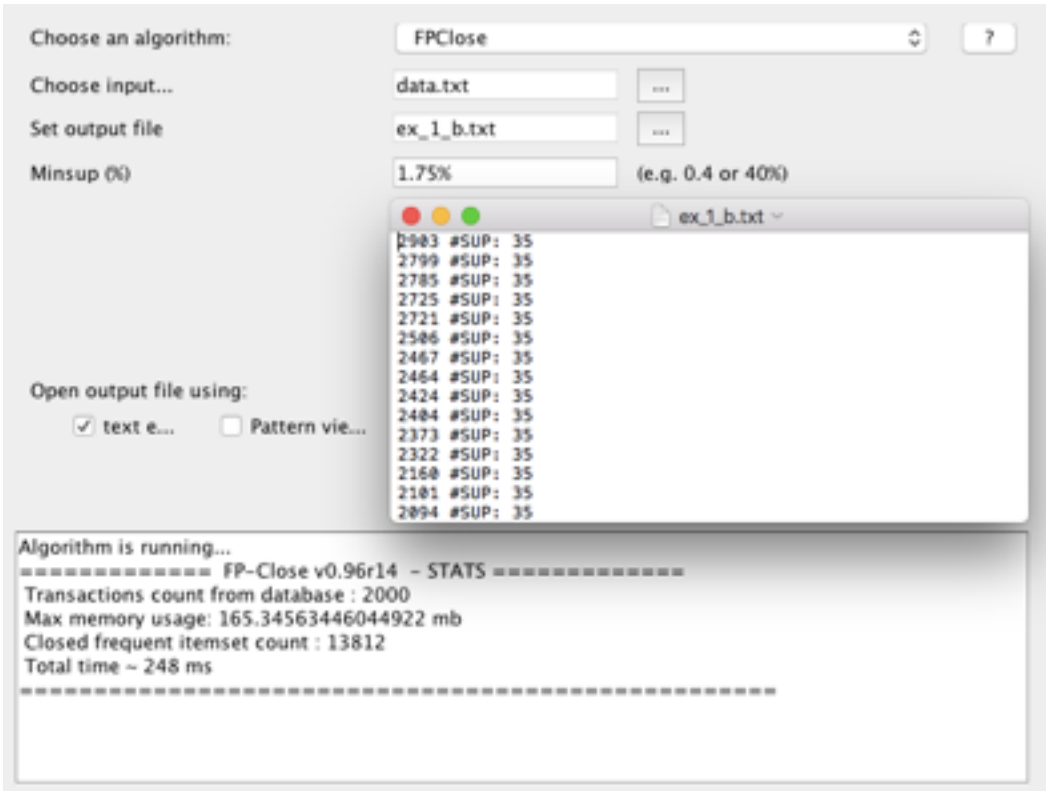
а) Для массива данных о контекстной рекламе размером 2000 компаний × 3000 словосочетаний найти частые множества для минимальной поддержки minsupp = 35. Необходимо указать число таких множеств.

FPGrowth_itemsets algorithm. Frequent itemsets count: **20910**



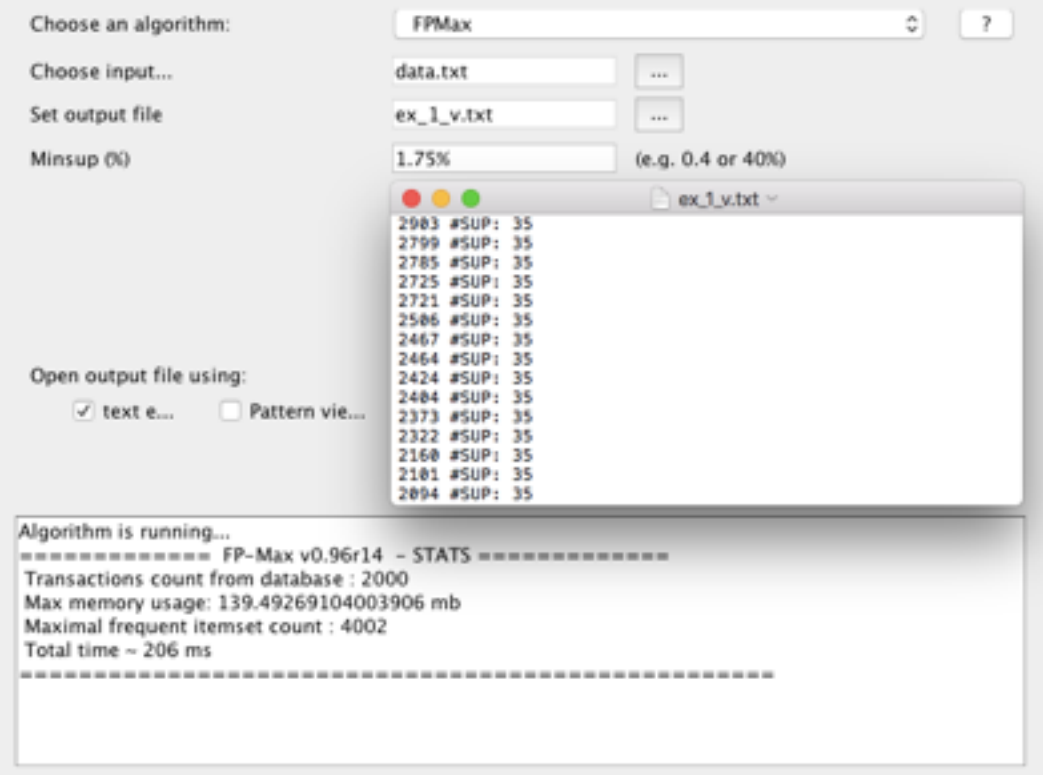
б) Повторить подзадание а) для частых замкнутых множеств.

FPClose algorithm. Closed frequent itemset count: **13812**



в) Повторить подзадание а) для частых максимальных множеств.

FPMax algorithm. Maximal frequent itemset count : **4002**



г) Среди множеств, найденных в заданиях а), б), в) указать примерно 10 множеств размером около 10 словосочетаний и провести их интерпретацию как рынков.

1. 969 989 995 1001 1011 1021 1882 1888 1906 #SUP: 37 (**а) FPGrowth_itemsets**)

   "casino", "casino gambling", "casino gambling online", "casino game online", "casino internet", "casino online", "gambling", "gambling internet", "gambling online"

2. 989 995 1001 1011 1021 1888 1906 1913 #SUP: 35 (**а) FPGrowth_itemsets**)

   "casino gambling", "casino gambling online", "casino game online", "casino internet", "casino online", "gambling internet", "gambling online", "gambling site»

3. 1021 1906 995 1888 1011 989 1001 1019 1016 #SUP: 37 (**в) FPMax**)

Рынок онлайн-казино. В контексте наших данных множеств, которые содержат ~10 словосочетаний, оказалось в разы больше для словосочетаний, связанных с рынком интернет-казино, чем с какой то другой областью. Можно предположить, что компаний, предоставляющих услуги онлайн-гемблинга становятся все больше в связи с тем, что этот рынок набирает с каждым годом огромную популярность. Причиной тому, может служить сразу несколько факторов:

- Увеличение общего объёма рынка;
- Увеличение мобильного сегмента;
- Новые игровые решения;
- Сегментирование целевой аудитории;
- Появление биткоин-казино;

4. 2501 2330 2505 2511 2086 1852 2503 #SUP: 35 (**в) FPMax**)

"make money", "internet make money", "make money online", "make money web", "home make money", "from home make money", "make money net"

5. 1852 2086 2330 2501 2505 2511 #SUP: 38 (**а) FPGrowth_itemsets**)

"from home make money", "home make money", "internet make money", "make money", "make money online", "make money web"

Рынок удаленной работы, фриланса. Клиентами рынка зачастую являются как сами «фрилансеры», люди ищущие работу, так и их потенциальные работодатели: малый и средний бизнес. Причина появления на рынке первых понятна, а вот с последними все не так очевидно. В действительности, малый и средний бизнес редко может набрать целый IT-отдел. В случае, если человек в штате один, да еще и узкоспециализированный, он не может выполнять разнородные задачи, а они в бизнесе возникают очень часто. Также стоит заметить, что фрилансерами чаще всего являются дизайнеры, программисты, переводчики и копирайтеры.

6. 663 674 681 667 665 346 332 #SUP: 35 (**б) FPClose**)

"business home", "business home opportunity", "business home work", "business home internet", "business home idea", "based business home idea", "base business home»

7. 663 345 674 355 2991 681 665 #SUP: 36 (**в) FPMax**)

"business home", "based business home", "business home opportunity", "based business home opportunity", "work home", "business home work", "business home idea"

Рынок домашнего бизнеса либо услуг по предоставлению идей для бизнеса на дому.

Клиентами рынка могут являться люди, ограниченные в передвижении. Например, женщины, находящиеся в декретном отпуске. Также, это могут быть частные предприниматели, предоставляющие услуги на дому. Например, портные или мастера по маникюру и прочим областям бьюти-индустрии.

8. 1170 2130 2131 2136 2155 2156 2159 2166 #SUP: 35 (**а) FPGrowth_itemsets**)

"company hosting web", "hosting internet site web", "hosting internet web", "hosting page web", "hosting services site web", "hosting services web", "hosting site web", "hosting web»

9. 120 1074 1075 1170 1233 1471 2156 2159 2166 #SUP: 35 (**а) FPGrowth_itemsets**)

"affordable hosting web", "cheap hosting site web", "cheap hosting web", "company hosting web", "cost hosting low web", "discount hosting web", "hosting services web", "hosting site web", "hosting web"

10. 2166 2159 2156 120 1233 1075 1170 1074 1471 #SUP: 35 (**в) FPMax**)

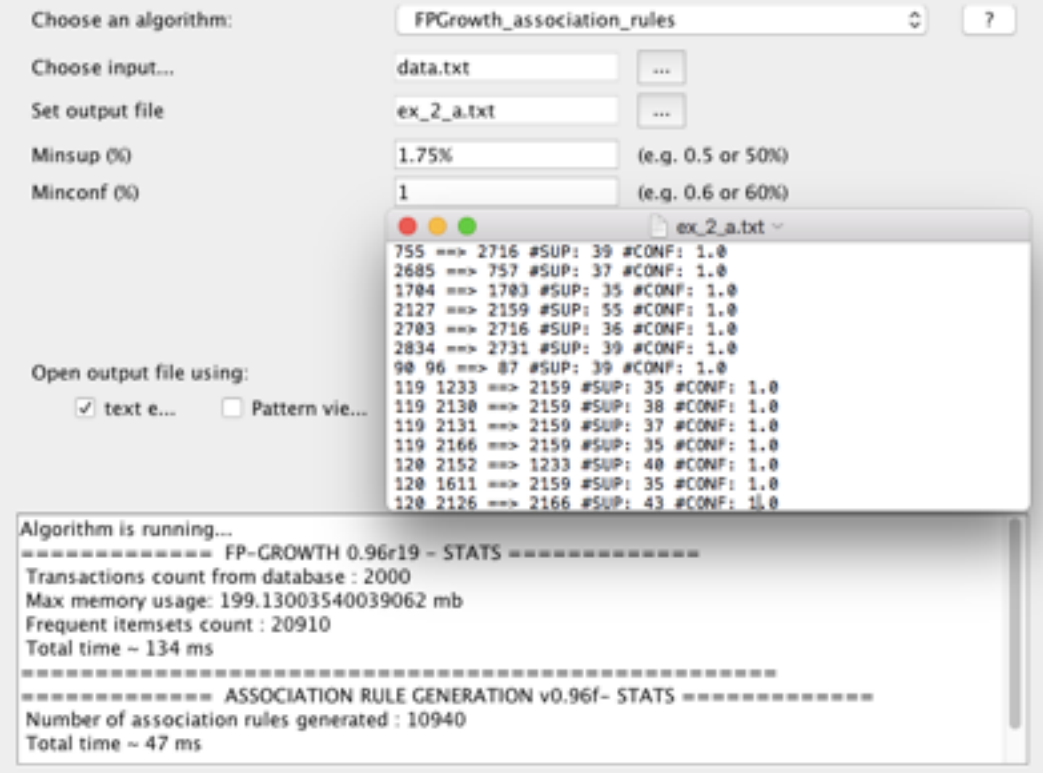
"hosting web", "hosting site web", "hosting services web", "affordable hosting web", "cost hosting low web", "cheap hosting web", "company hosting web", "cheap hosting site web", "discount hosting web"

Рынок услуг по размещению web-сайтов. Важно заметить, что наряду со словами web site, присутствуют также слова cheap, low cost, affordable, discount, что возможно говорит о более узкой ниши рынка - предоставление именно недорогих услуг. Такими услугами в большинстве своем может пользоваться как малый бизнес, так и частный предприниматель.

Задание 2 (3 балла). Поиск ассоциативных правил

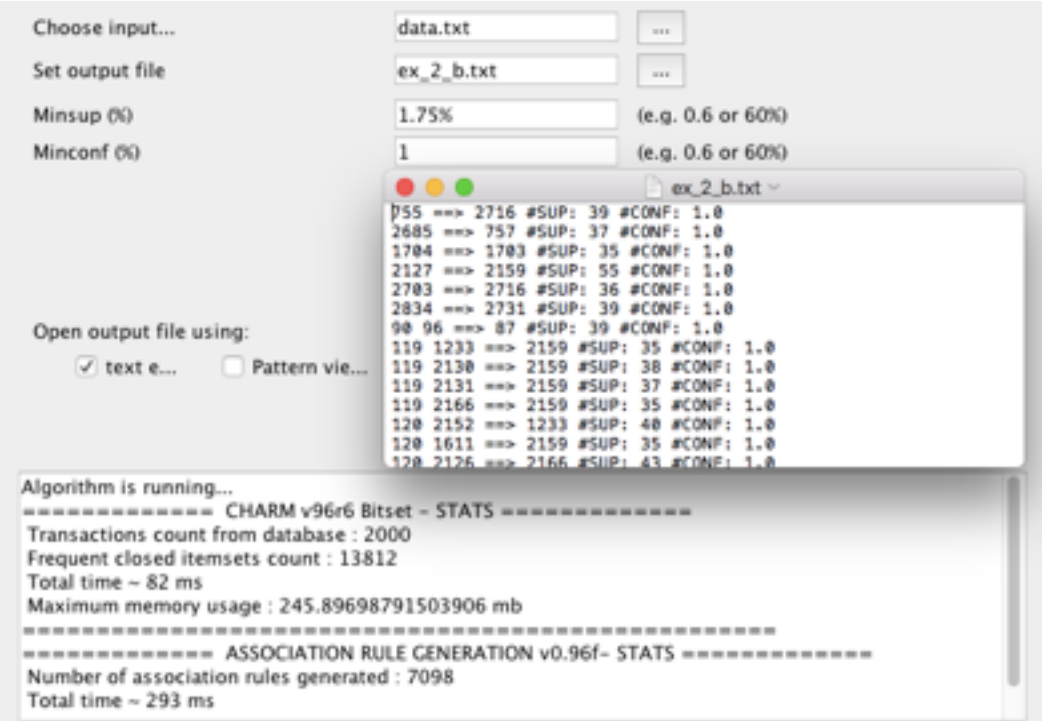
а) Для массива данных о контекстной рекламе 2000 компаний × 3000 словосочетаний найти ассоциативные правила для минимальной поддержки minsupp = 35 и minconf = 1. Необходимо указать число таких правил.

FPGrowth_association_rules algorithm. Number of association rules generated: **10940**



б) Для исходного массива данных найти замкнутые ассоциативные правила для минимальной поддержки minsupp = 35 и minconf = 1. Необходимо указать число таких правил.

Closed_association_rules algorithm. Number of association rules generated: **7098**



в) Для исходного массива данных найти 5 самых частых правил при минимальной достоверности minconf = 0,8. Необходимо указать эти правила и дать интерпретацию.

TopKRules algorithm

1. 345 674 ==> 663 #SUP: 90 #CONF: 0.8490566037735849

   "based business home", "business home opportunity" ==> "business home"

   Для компаний, предоставляющих услуги в сфере домашнего бизнеса, удаленной работы, фриланса есть предложение использовать в контекстной рекламе следующие словосочетания: "based business home", "business home opportunity", "business home", "based business home opportunity", потому как потенциальные клиенты данного рынка в поисковых запросах указывая хотя бы одно словосочетания из предложенного набора, будут указывать и другое (с учетом достаточно высокой поддержки и достоверности).
2. 663 674 ==> 345 #SUP: 90 #CONF: 0.8256880733944955

   "business home", "business home opportunityc ==> "based business home"

   см. первый пункт
3. 2536 ==> 2336 #SUP: 91 #CONF: 0.8666666666666667

   "marketing online" ==> "internet marketing"

   В данном случае, компании по предоставлению маркетинговых онлайн-услуг можно предложить помимо покупки словосочетания marketing online, купить также словосочетание internet marketing, потому что с поддержкой 91 транзакции человек в поисках информации о данной услуге, в запросе будет использовать оба словосочетания. Также можно заметить, что из всех остальных правил, данное обладает большей достоверностью.
4. 355 ==> 345 #SUP: 102 #CONF: 0.8292682926829268

   "based business home opportunity" ==> "based business home"

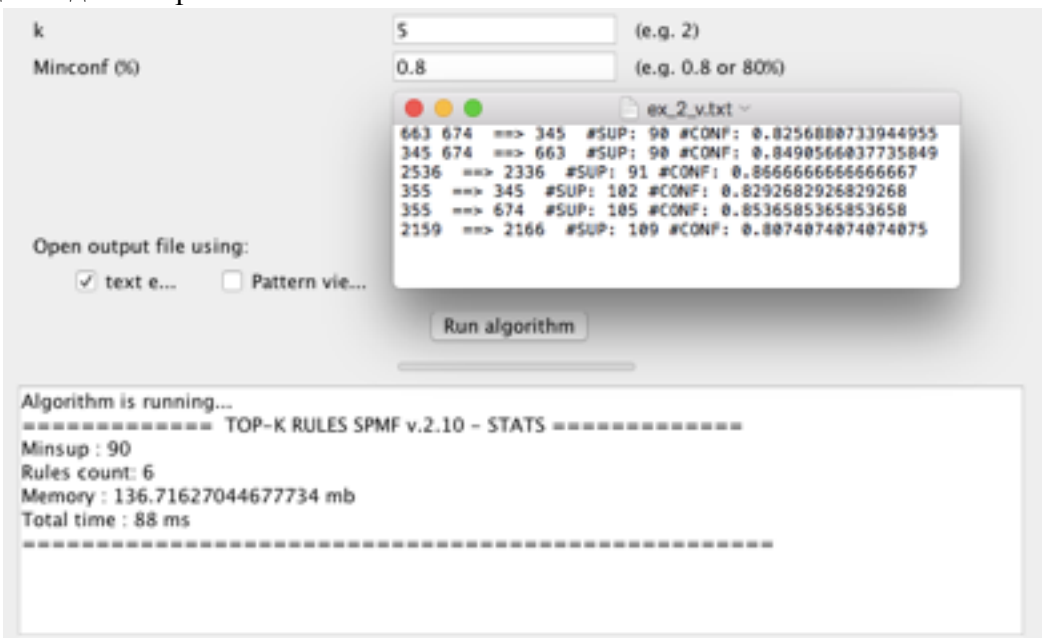
   см. первый пункт
5. 355 ==> 674 #SUP: 105 #CONF: 0.8536585365853658

   "based business home opportunity"" ==> "business home opportunity"

   см. первый пункт
6. 2159 ==> 2166 #SUP: 109 #CONF: 0.8074074074074075

   "hosting site web" ==> "hosting web"

   Рекомендацией фирмам рынка услуг по размещению сайтов может являться использование данных словосочетаний в контекстной рекламе, так как с поддержкой в 109 транзакций и достоверностью ~81%



Задание 3 (4 балла). Анализ посещаемости сайтов на основе решеток формальных понятий

Для трех контекстов о посещаемости сайта Высшей школы экономики в терминах посещений сайтов новостной, образовательной и финансовой тематики необходимо выполнить пункты задания ниже.

а) Удалением некоторого числа сайтов (признаков) или пользователей (объектов) добиться числа формальных понятий не менее 100, но не сильно превышающего это значение.

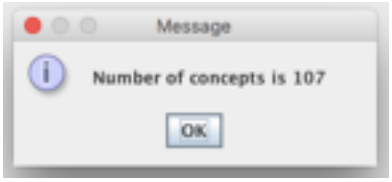
Контекст №1

новостная тематика

Контекст №2

финансовая тематика

Контекст №3



образовательная тематика

б) Для контекстов, полученных удалением объектов или признаков в пункте а), построить диаграммы решеток понятий.

Диаграмма решеток понятий для контекста №1

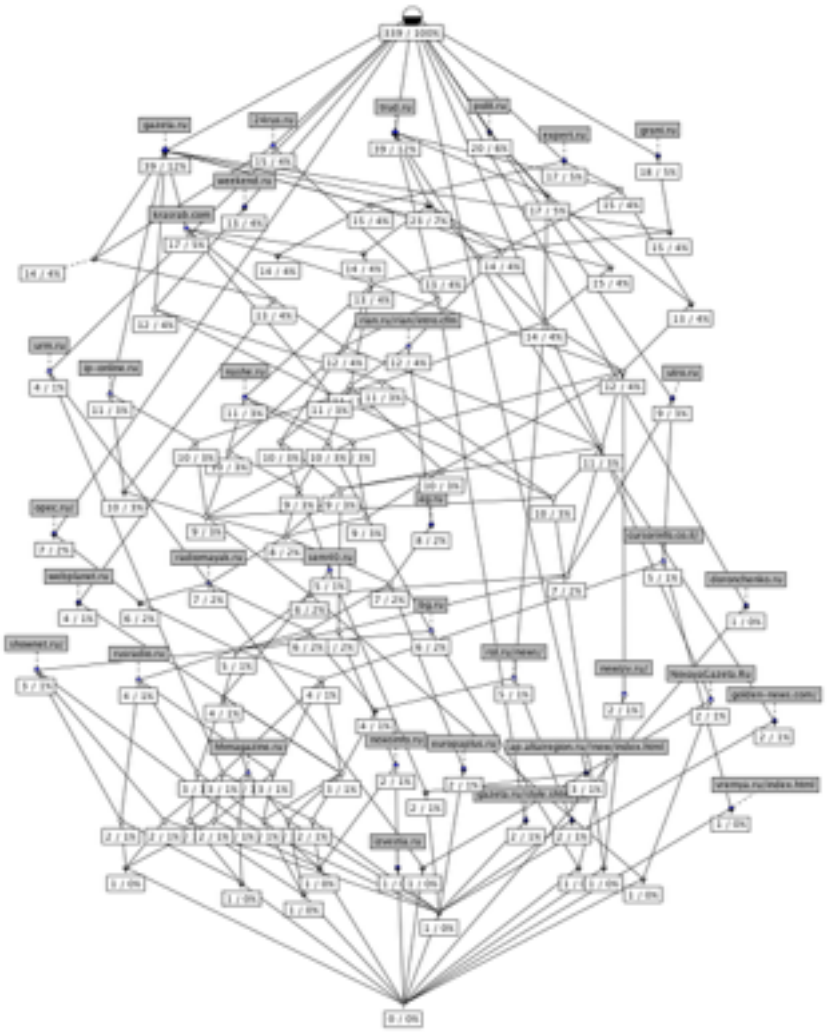


Диаграмма решеток понятий для контекста №2

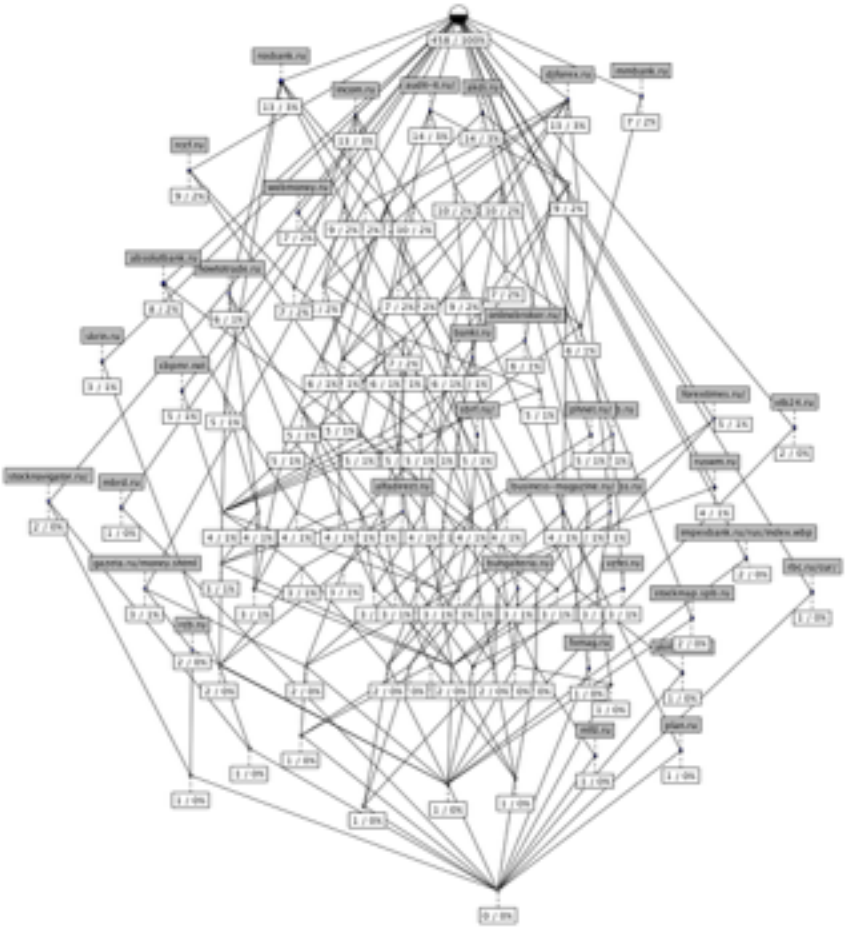
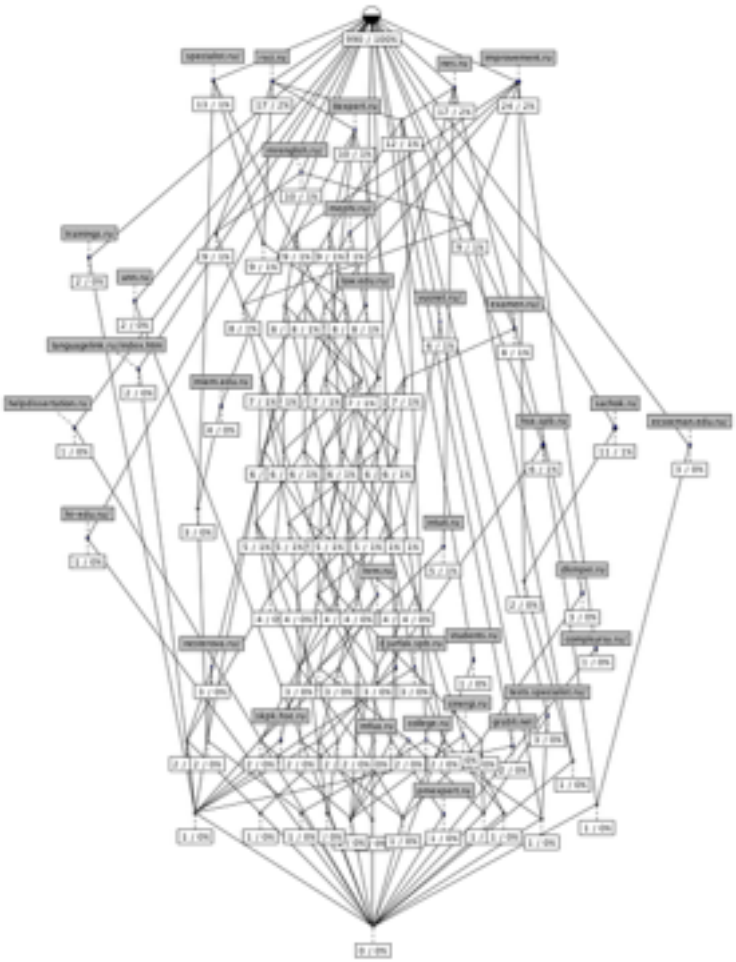
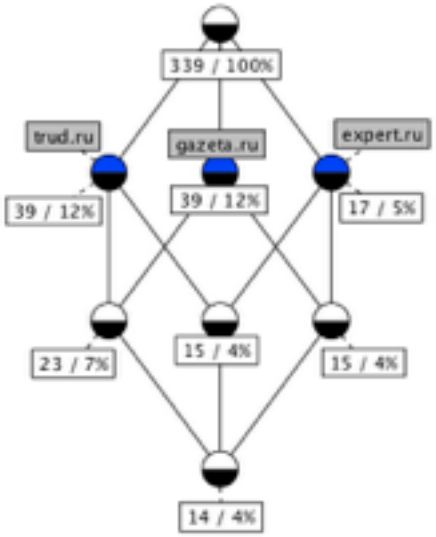


Диаграмма решеток понятий для контекста №3



в) Привести 3–5 примеров понятий в виде пары ⟨размер объема понятия, содержание понятия⟩ для размера содержания 2 и более сайта. Дать содержательную интерпретацию найденных понятий.

Контекст №1

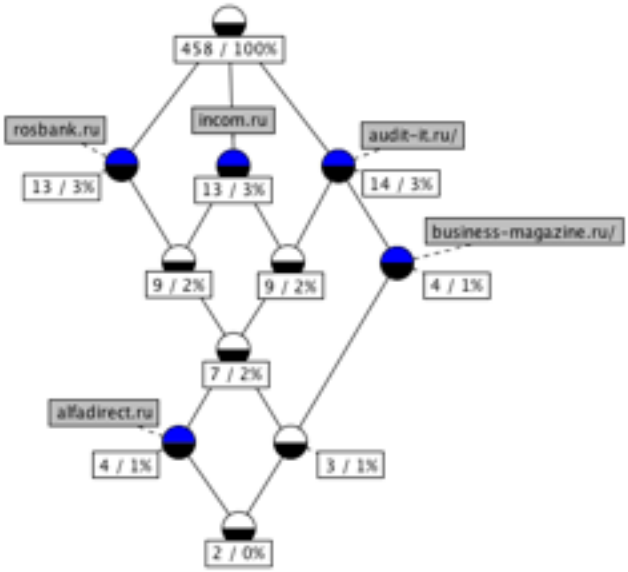


новостная тематика

больший интерес пользователей, которые посетили сайт ВШЭ, проявляется к сайтам trud.ru и gazeta.ru

1. <23, trud.ru gazeta.ru > 23 пользователя, посетившие сайт Высшей Школы Экономики, посетили также сайт общественной политической газеты trud.ru и информационную ленту gazeta.ru;
2. <15, trud.ru expert.ru> 15 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт общественной политической газеты trud.ru и сайт российского делового еженедельника expert.ru;
3. <15, gazeta.ru expert.ru> 15 пользователей, посетивших сайт Высшей Школы Экономики, посетили также информационную ленту gazeta.ru и сайт российского делового еженедельника expert.ru;
4. <14, trud.ru gazeta.ru expert.ru> 14 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт общественной политической газеты trud.ru, информационную ленту gazeta.ru и сайт российского делового еженедельника expert.ru;

Контекст №2

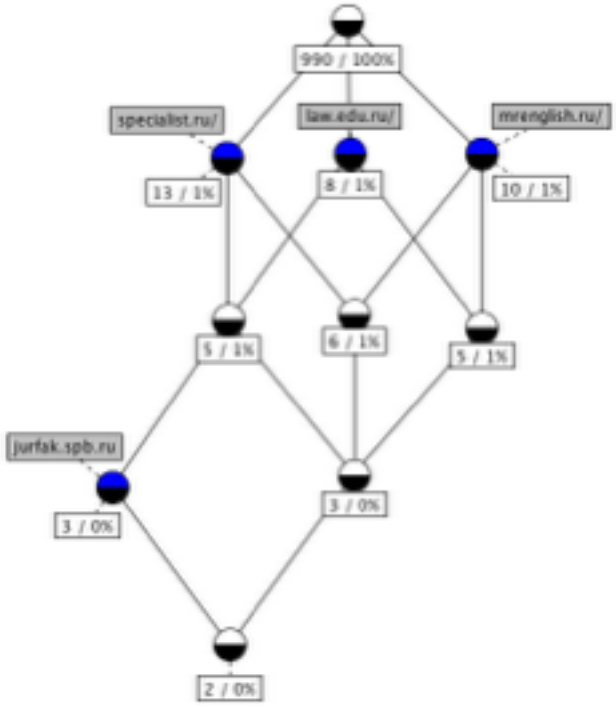


финансовая тематика

Среди посетителей сайта ВШЭ, интерес в финансовой тематике проявляется к сайту финансового аудита

1. <9, rosbank.ru incom.ru> 9 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт российского универсального банка rosbank.ru и сайт агентства недвижимости Инком incom.ru;
2. <9, incom.ru audit-it.ru/> 9 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт агентства недвижимости Инком incom.ru и интернет-портал, посвященный бухгалтерскому учету, налогам и аудиторской деятельности audit-it.ru/;
3. <7, rosbank.ru incom.ru audit-it.ru/> 7 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт российского универсального банка rosbank.ru, сайт агентства недвижимости Инком incom.ru и интернет-портал, который посвящен бухгалтерскому учету, налогам и аудиторской деятельности audit-it.ru/;
4. <3, rosbank.ru incom.ru audit-it.ru/ business-magazine.ru/> 3 пользователя, посетившие сайт Высшей Школы Экономики, посетили также сайт российского универсального банка rosbank.ru, сайт агентства недвижимости Инком incom.ru, интернет-портал, который посвящен бухгалтерскому учету, налогам и аудиторской деятельности audit-it.ru/, а также сайт бизнес-журнала business-magazine.ru/;
5. <2, rosbank.ru incom.ru audit-it.ru/ business-magazine.ru/ alfadirect.ru> 2 пользователя, посетивших сайт Высшей Школы Экономики, посетили также сайт российского универсального банка rosbank.ru, сайт агентства недвижимости Инком incom.ru и интернет-портал, который посвящен бухгалтерскому учету, налогам и аудиторской деятельности audit-it.ru/, сайт бизнес-журнала business-magazine.ru/ и сайт системы интернет-трейдинга Альфа-Банка alfadirect.ru.

Контекст №3



образовательная тематика

Пользователи, посетившие сайт ВШЭ, проявили интерес к сайту учебного центра «Специалист» при МГТУ имени Н.Э.Баумана

1. <5, specialist.ru/ law.edu.ru/> 5 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт учебного центра «Специалист» при МГТУ имени Н.Э.Баумана и образовательный правовой портал law.edu.ru/;

2. <6, specialist.ru/ mrenglish.ru/> 6 пользователей, посетивших сайт Высшей Школы Экономики, посетили также сайт учебного центра «Специалист» при МГТУ имени Н.Э.Баумана и сайт курсов изучения иностранных языков в Москве Mr. English;
3. <5, law.edu.ru/ mrenglish.ru/> 5 пользователей, посетивших сайт Высшей Школы Экономики, посетили также образовательный правовой портал law.edu.ru/ и сайт курсов изучения иностранных языков в Москве Mr. English;
4. <3, specialist.ru/ law.edu.ru/ mrenglish.ru/> 3 пользователя, посетившие сайт Высшей Школы Экономики, посетили также сайт учебного центра «Специалист» при МГТУ имени Н.Э.Баумана, образовательный правовой портал law.edu.ru/ и сайт курсов изучения иностранных языков в Москве Mr. English;
5. <2, specialist.ru/ law.edu.ru/ mrenglish.ru/ jurfak.spb.ru> 3 пользователя, посетившие сайт Высшей Школы Экономики, посетили также сайт учебного центра «Специалист» при МГТУ имени Н.Э.Баумана, образовательный правовой портал law.edu.ru/, сайт курсов изучения иностранных языков в Москве Mr. English и сайт юридического факультета Санкт-Петербургского государственного университета jurfak.spb.ru;

г) Привести пример импликации вида A→B, найденной по диаграмме решетки понятий с указанием ее поддержки.

Примеры импликаций для контекста №1

```
1 < 17 > krasrab.com ==> gazeta.ru;
2 < 14 > trud.ru gazeta.ru expert.ru ==> grani.ru;
3 < 14 > trud.ru grani.ru ==> gazeta.ru expert.ru;
4 < 14 > expert.ru grani.ru ==> trud.ru gazeta.ru;
5 < 13 > expert.ru polit.ru ==> trud.ru;
6 < 13 > gazeta.ru krasrab.com grani.ru ==> trud.ru expert.ru;
7 < 12 > gazeta.ru polit.ru krasrab.com ==> trud.ru expert.ru grani.ru;
8 < 12 > rian.ru/rian/intro.cfm ==> trud.ru gazeta.ru expert.ru grani.ru;
9 < 12 > polit.ru grani.ru ==> trud.ru gazeta.ru expert.ru krasrab.com;
10 < 12 > trud.ru gazeta.ru 24rus.ru ==> expert.ru krasrab.com grani.ru;
```

Рассмотрим первый пример импликации:

1. <17> krasrab.com ==> gazeta.ru

Здесь, число 17 отвечает за значение поддержки того, что каждый объект, обладающий признаками krasrab.com, обладает также и признаками gazeta.ru, то есть пользователь, посетивший сайт красноярского рабочего, посетит также и информационную ленту новостей gazeta.ru.

Примеры импликаций для контекста №2

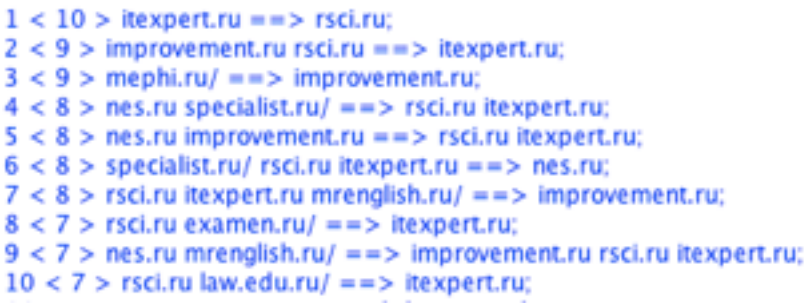
```
1 < 9 > incom.ru audit-it.ru/ ==> djforex.ru;
2 < 7 > akdi.ru incom.ru ==> djforex.ru;
3 < 7 > rosbank.ru audit-it.ru/ ==> djforex.ru incom.ru;
4 < 6 > onlinebroker.ru/ ==> djforex.ru;
5 < 6 > banki.ru ==> akdi.ru djforex.ru;
6 < 6 > akdi.ru mmbank.ru ==> djforex.ru audit-it.ru/;
7 < 6 > djforex.ru mmbank.ru ==> akdi.ru audit-it.ru/;
8 < 6 > mmbank.ru audit-it.ru/ ==> akdi.ru djforex.ru;
9 < 6 > rosbank.ru rccf.ru ==> incom.ru;
10 < 6 > howtotrade.ru ==> rosbank.ru;
```

Один из примеров интерпретации импликации:

1. <9> incom.ru audit-it.ru/ ==> djforex.ru

Каждый объект, обладающий признаками incom.ru audit-it.ru/, обладает также и признаками djforex.ru, то есть пользователь, посетивший первые два сайта (агенство недвижимости и портал, посвященный бухгалтерскому учету, налогам и аудиторской деятельности в РФ), посетит также и ленту новостей DJ Forex с поддержкой в 9 транзакций.

Примеры импликаций для контекста №3



Пример интерпретации:

1. <9> mephi.ru/ ==> improvement.ru

При поддержке в 9 транзакций, пользователь, посетивший сайт Национального Исследовательского Ядерного Университета «МИФИ», посетит также и сайт организации времени. Миссия компании, владеющей сайтом заключается в следующем: «помогать людям, обществу, бизнесу и государству в настройке отношений с самым ценным и трудноуправляемым ресурсом – Временем». Таким образом данная импликация является весьма достоверной, так как у студентов зачастую возникают проблемы с распределением времени.