# Task 1 (30 points). Frequent Itemset Mining

**a) Given a contextual advertising dataset of 2000 companies × 3000 terms, find all frequent itemsets with minsupp=35. Report the number of such itemsets.**

All additional computations and data preprocessing was performed in google collab - HW3_MMDA_Taran.ipynb . Also notebook and additional files are in the zip folder.

For the Mining Frequent Itemsets with minsupp = 35 the FP-Growth Algorithm was used. Firstly data was preprocessed in python with the help of pandas library as an input file for SPMF FP-Growth must look like:

…

*2503 2504 2505 2506 2507 2508 2509 2510 2511 2515 2517 2519 2520 2615 2616 2991*

*281 285 286 287 740 741 766 767 830 831 832 833 835 1719 1721 1808 1809 1810 1811 2842 2963*

…

Input file name is `'list_of_companies_terms.txt'` Minsupp=35, which means that for SPMF we must pass minsupp = 35/2000 = 1.75%, because SPMF requires a minsupp parameter as a percentage value. Output file name: FIM_FPG.txt

```
Algorithm is running... (01:19:00 PM)
============= FP-GROWTH 2.42 - STATS =============
 Transactions count from database : 2000
 Max memory usage: 153.97693634033203 mb
 Frequent itemsets count : 20910
 Total time ~ 99 ms
================================================
```

The number of itemsets with these parameters is 20910.

**b) Repeat subtask a) for frequent closed itemsets.**

For frequent closed itemsets I'll use an FPClose algorithm in SPMF. Input file and other parameters are the same. Output file name: FIM_FPClose.txt

```
Algorithm is running... (01:25:53 PM)
============= FP-Close v0.96r14 - STATS =============
 Transactions count from database : 2000
 Max memory usage: 158.77972412109375 mb
 Closed frequent itemset count : 13812
 Total time ~ 298 ms
================================================
```

**c) Repeat subtask a) for maximal frequent itemsets.**

For maximal frequent itemsets FPMax Algorithm was used. Input file and other parameters are the same. Output file name: FIM_FPMax.txt

```
Algorithm is running... (01:28:20 PM)
============  FP-Max v0.96r14  - STATS ============
 Transactions count from database : 2000
 Max memory usage: 157.48270416259766 mb
 Maximal frequent itemset count : 4002
 Total time ~ 238 ms
=================================================
```

**d) Among the resulting itemsets for a), b), and c), indicate 10 itemsets composed of 10 terms or greater and interpret them as “markets”.**

For every output file there was a search for itemsets which contained 10 or more terms. Unfortunately the maximum number of items in itemsets for every algorithm was 9, so the analysis was performed for all itemsets consisting of 9 items. Terms were returned for string term names from indexes using python. There is the list of them:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| FP - Growth | 'casino gambling', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino net', "casino online', 'gambling internet', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino net', 'casino online', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', 'casino line', 'casino net', 'casino online', 'gambling internet', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino online', 'gambling internet', 'gambling online' | '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' | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino internet', 'casino online', 'gambling', 'gambling internet', 'gambling online' | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino online', 'gambling internet', 'gambling online' | 'casino', 'casino gambling' , 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino online', 'gambling ', 'gambling online' | 'casino', 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', 'casino online', 'gambling', 'gambling internet', 'gambling online' |
| FPClose | 'casino gambling', 'casino game', 'casino game online', 'casino internet', 'casino line', | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino | 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', 'casino | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino | 'affordable hosting web', 'cheap hosting site web', 'cheap hosting web', 'company hosting | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', | 'casino', 'casino gambling' , 'casino gambling online', 'casino game', 'casino internet', 'casino | 'casino', 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 'casino net', "casino online', 'gambling internet', 'gambling online' | internet', 'casino line', 'casino net', 'casino online', 'gambling online' | line', 'casino net', 'casino online', 'gambling internet', 'gambling online' | internet', 'casino line', 'casino online', 'gambling internet', 'gambling online' | web', 'cost hosting low web', 'discount hosting web', 'hosting services web', 'hosting site web', 'hosting web' | 'casino internet', 'casino online', 'gambling internet', 'gambling online' | 'casino internet', 'casino online', 'gambling', 'gambling online' | online', 'gambling', 'gambling internet', 'gambling online' | 'casino online', 'gambling', 'gambling internet', 'gambling online' |
| FPMax | 'casino gambling', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino net', "casino online', 'gambling internet', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino net', 'casino online', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', 'casino line', 'casino net', 'casino online', 'gambling internet', 'gambling online' | 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino line', 'casino online', 'gambling internet', 'gambling online' | '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' | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino online', 'gambling internet', 'gambling online' | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino game online', 'casino internet', 'casino online', 'gambling', 'gambling online' | 'casino', 'casino gambling', 'casino gambling online', 'casino game', 'casino internet', 'casino online', 'gambling', 'gambling internet', 'gambling online' | 'casino', 'casino gambling', 'casino gambling online', 'casino game online', 'casino internet', 'casino online', 'gambling', 'gambling internet', 'gambling online' |
| **market** | **online casino** | **online casino** | **online casino** | **online casino** | **web hosting** | **online casino** | **online casino** | **online casino** | **online casino** |

Interesting that for FPMax and FPClose algorithms the composition and order of itemsets are the same. And also there are large intersections between them and the FP-Growth algorithm.

# Task 2 (30 points). Association Rules Mining

**a) For advertising dataset with 2000 firms × 3000 terms find association rules with minsupp = 35 and minconf = 1. Indicate the number of such rules.**

```
Algorithm is running... (02:09:26 PM)
============= FP-GROWTH 2.42 - STATS =============
 Transactions count from database : 2000
 Max memory usage: 154.04537200927734 mb
 Frequent itemsets count : 20910
 Total time ~ 82 ms
=================================================
============= ASSOCIATION RULE GENERATION v2.19- STATS =============
 Number of association rules generated : 10940
 Total time ~ 49 ms
=================================================
```

The number of rules when mining all association rules is 10940

**b) For the input dataset find closed association rules with minsupp=35 and minconf=1. Indicate the number of such rules**.

```
Algorithm is running... (02:13:50 PM)
============= CHARM v96r6 Bitset - STATS =============
 Transactions count from database : 2000
 Frequent closed itemsets count : 13812
 Total time ~ 71 ms
 Maximum memory usage : 98.89827728271484 mb
=================================================
============= ASSOCIATION RULE GENERATION v2.19- STATS =============
 Number of association rules generated : 7098
 Total time ~ 112 ms
=================================================
```

The number of rules when mining closest association rules is 7098

**c) For the input dataset find the top-5 frequent rules with minconf = 0, 8. Provide all the found rules and their interpretation (at least for a couple of them) in the report.**

Because I've encountered some problems with running the TopKRules algorithm, I've run an Apriori association rules algorithm with parameters of minsupp = 35 (1.75%) and minconf = 0.8. The output file (apriori_80_ARM.txt) was then analyzed using python.

```
Algorithm is running... (03:38:15 PM)
============= APRIORI - STATS =============
 Candidates count : 288295
 The algorithm stopped at size 10
 Frequent itemsets count : 20910
 Maximum memory usage : 76.06134033203125 mb
 Total time ~ 10063 ms
=================================================
============= ASSOCIATION RULE GENERATION v2.19- STATS =============
 Number of association rules generated : 197750
 Total time ~ 150 ms
=================================================
```

So the top 5 most frequent association rules are:

- hosting site web ==> hosting web #SUP: 109 #CONF: 0.8074074074074075

For the rule `hosting site web ==> hosting web` with #SUP: 109, it means this combination of items appears in 109 transactions within the dataset. Denoted as #CONF, confidence measures the reliability of the rule. The rule `hosting site web ==> hosting web` has #CONF: 0.8074. This implies that in 80.74% of transactions containing "hosting site web", "hosting web" is also present.

- based business home opportunity ==> business home opportunity #SUP: 105 #CONF: 0.8536585365853658

—————//—————

- based business home opportunity ==> based business home #SUP: 102 #CONF: 0.8292682926829268

This rule suggests that when the phrase "based business home opportunity" appears, there's an 82.93% chance that "based business home" will also appear. This combination occurs in 102 transactions within the dataset.

- marketing online ==> internet marketing #SUP: 91 #CONF: 0.8666666666666667

This rule indicates that when "marketing online" is present, there's a high probability (86.67%) of finding "internet marketing" as well. This association is found in 91 transactions.

- business home business home opportunity ==> based business home #SUP: 90 #CONF: 0.8256880733944955

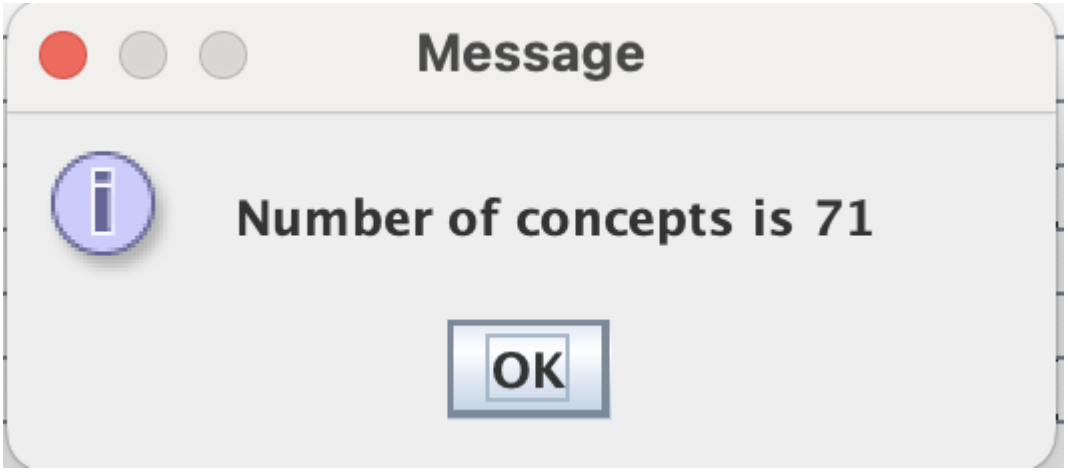
—————//—————

# Task 3 (40 points). Analysis of website visitors' behaviour based on concept lattices

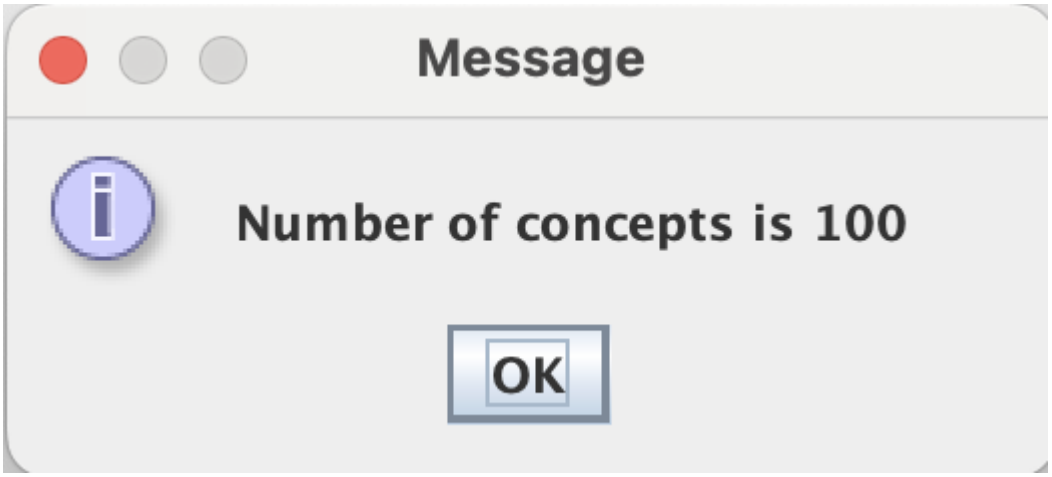
**For three input context about visitors of Higher School of Economics in terms of their visits of news websites, education-related and finance-related websites perform the sub- tasks below.**

**a) By removal of certain websites (attributes) or visitors (objects) in the input dataset make sure that the number of formal concepts is about 100.**

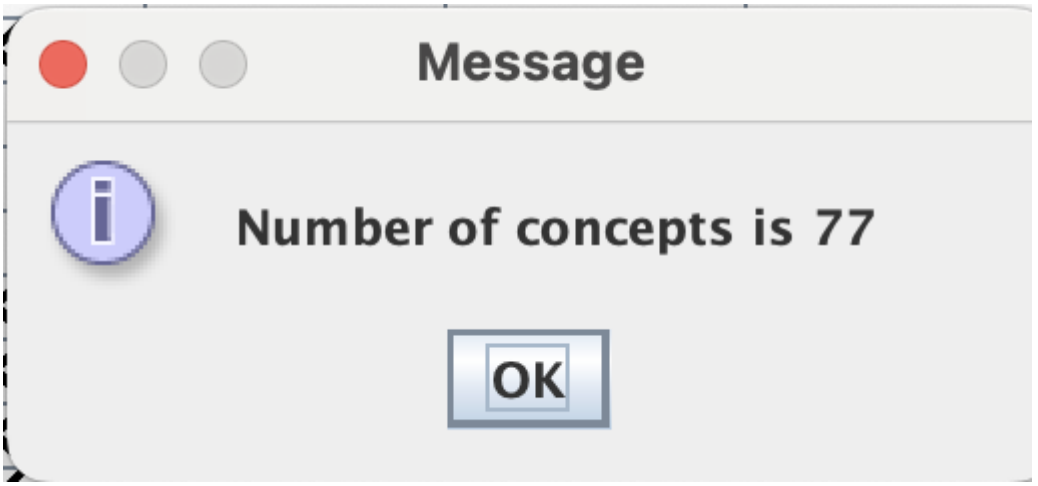
For hse_5516_20_sep_u (nothing was removed):




For hse_5289_20_sep_u (removed 6 objects):

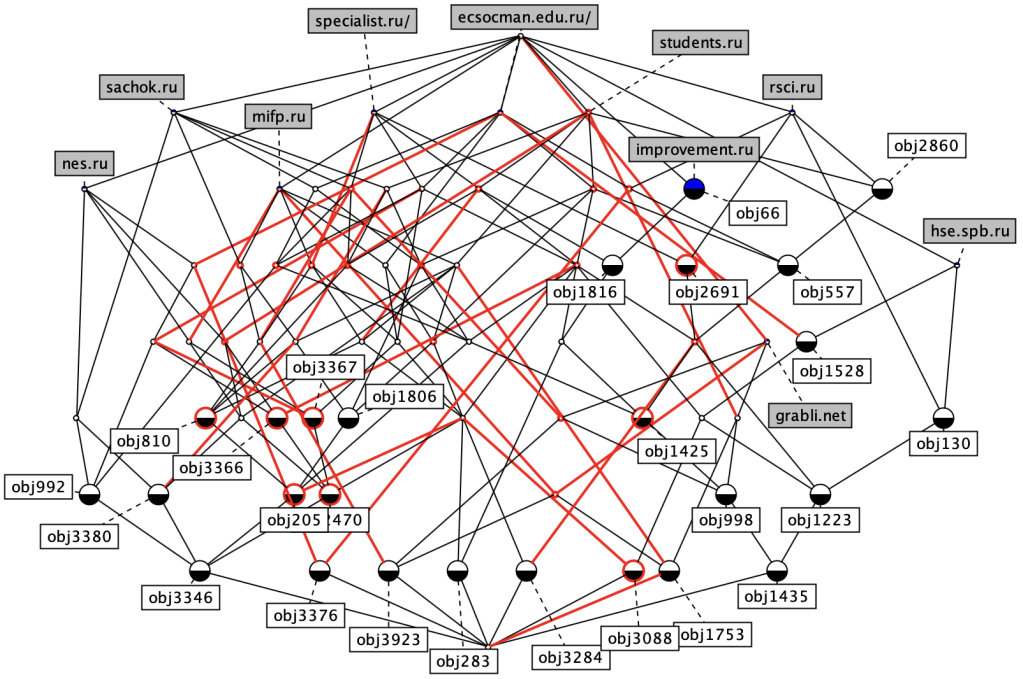



For hse_5282_20_sep_u (many objects and attributes were removed):

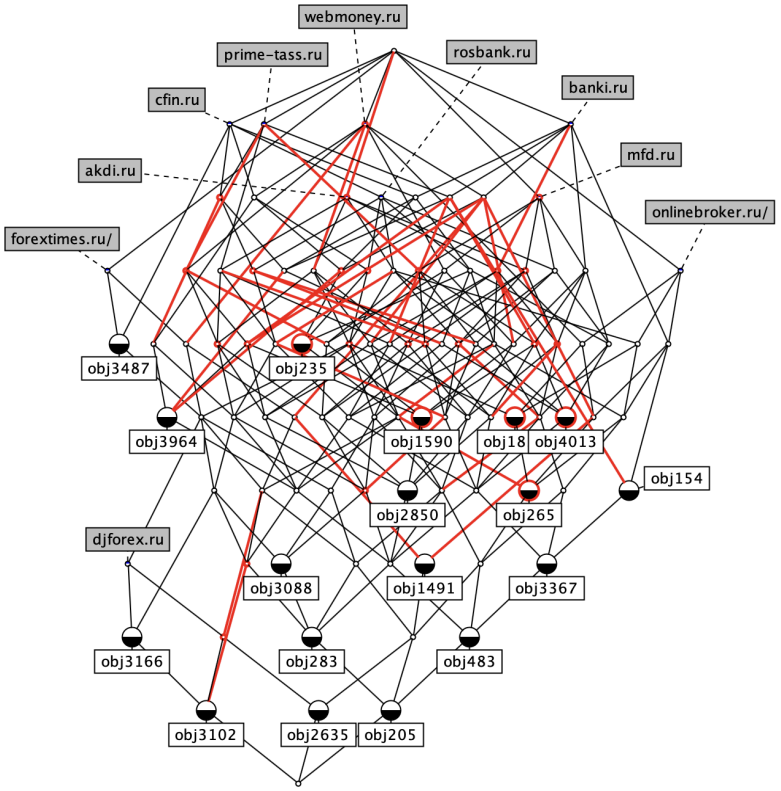



**b) For the context from subtask a) that obtained by object/attribute removal build the corresponding lattice diagrams.**

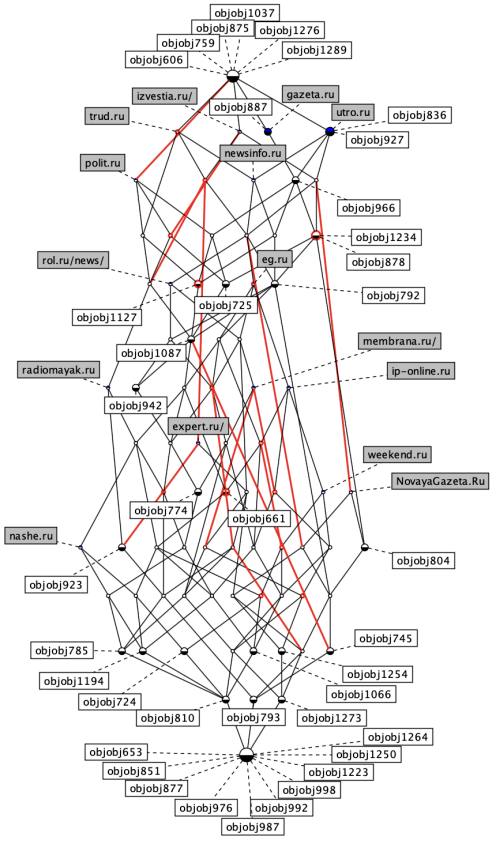
For hse_5516_20_sep_u:




For hse_5289_20_sep_u:

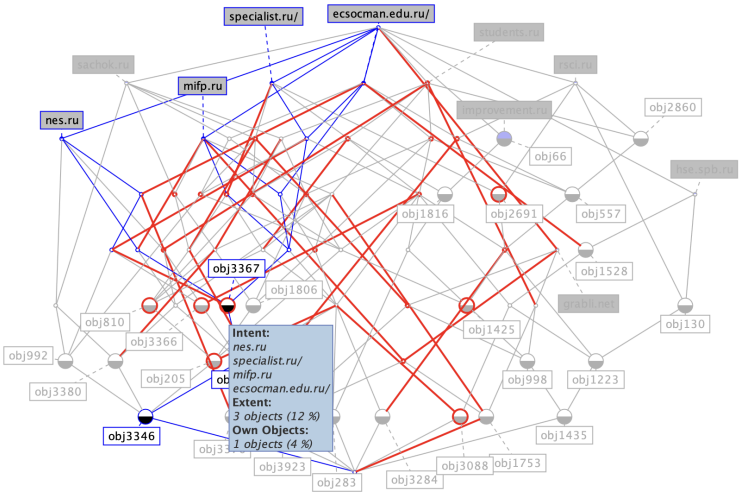


For hse_5282_20_sep_u:



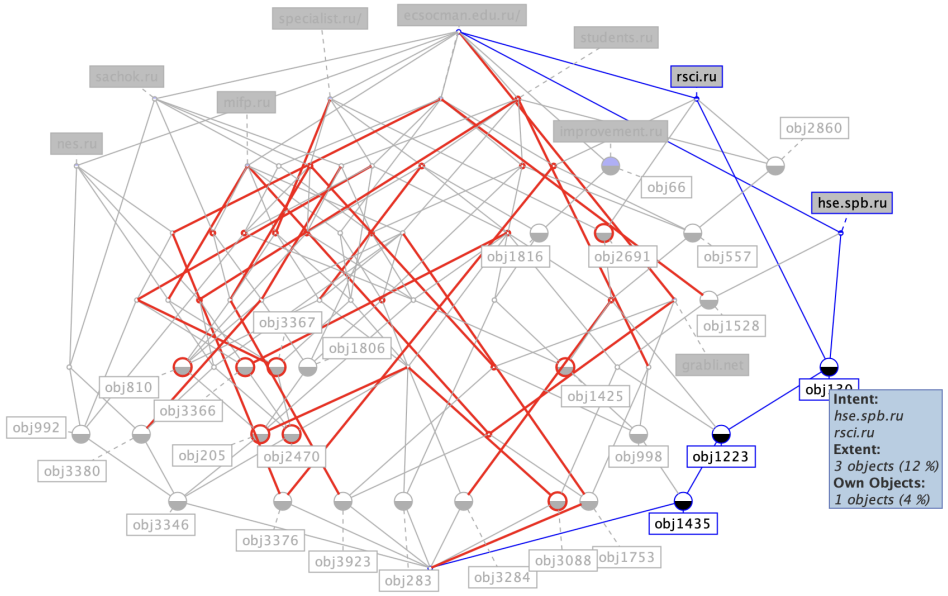

**c) Provide 3–5 examples of concepts as pairs ⟨ extent size, intent ⟩ for the intent size greater than 2. Give the interpretation of the found concepts.**

- For hse_5516_20_sep_u:



<{obj3367, obj2470, obj3346}, {nes.ru, mifp.ru, specialist.ru, ecsocman.edu.ru}>

This concept shows that 3 objects are reading all of the nes.ru, mifp.ru, specialist.ru, ecsocman.edu.ru sites. These sites have the same theme: education and professional specialization, maybe the aim of these objects is further of additional education.



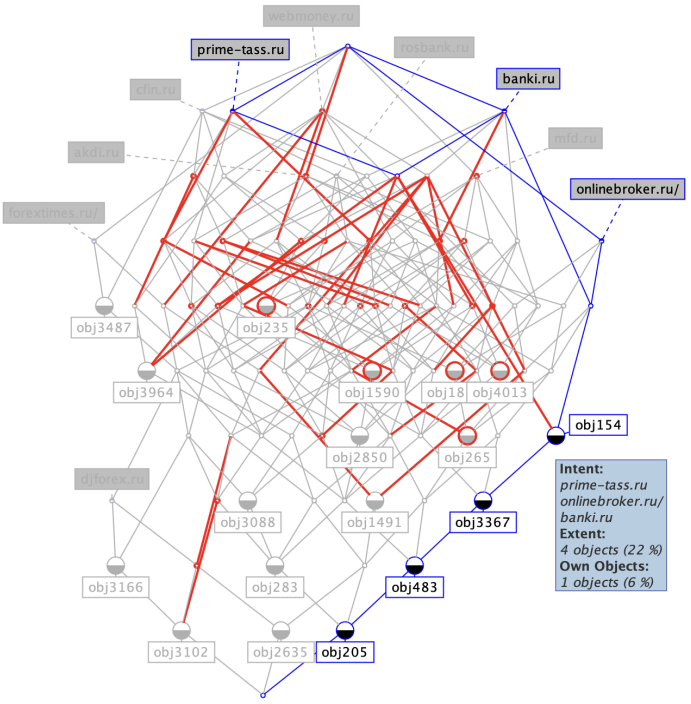

<{obj130, obj1223, obj1435}, {hse.spb.ru, rsci.ru}>

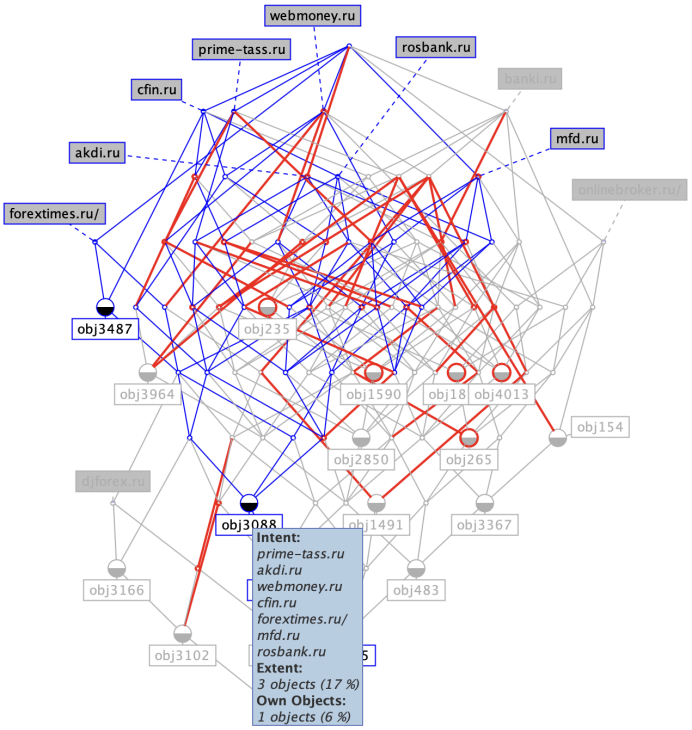
There is no common theme for these sites.

- For hse_5289_20_sep_u:

<{obj154, obj3367, obj483, obj205}, {prime-tass.ru, onlinebroker.ru, banki.ru}>

What this set of sites has in common is an ecommerce focus with the goal of saving money. It is likely that the objects that fall into this concept follow stocks.

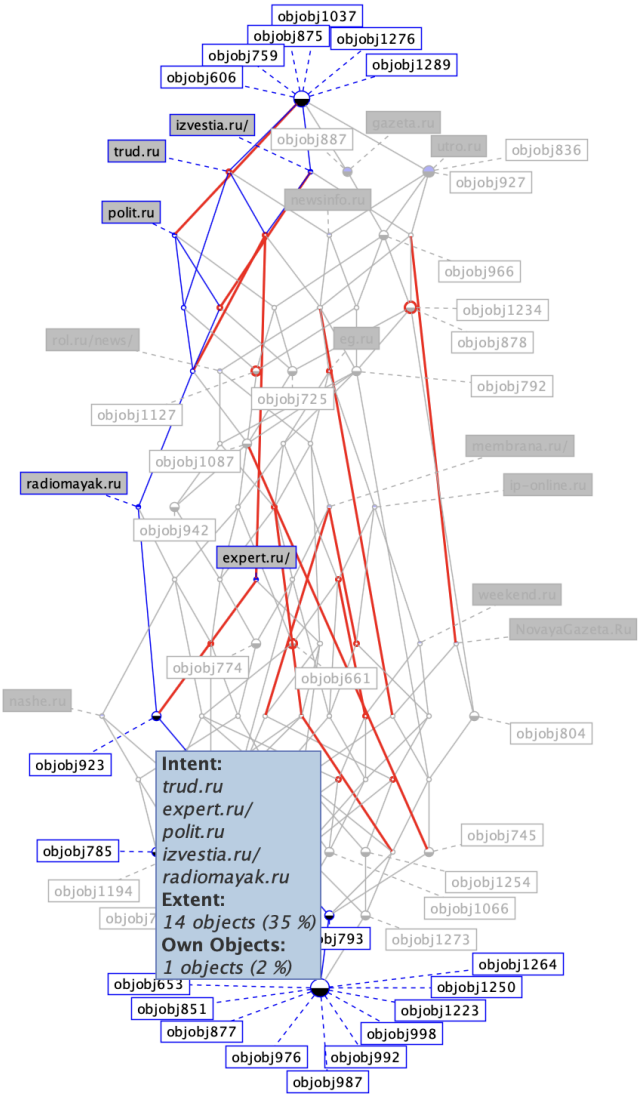



<{obj3487, obj283, obj205}, {prime-tass.ru, akdi.ru, webmoney.ru, cfin.ru, forextimes.ru, mfd.ru, rosbank.ru}>

There is a larger set of intent items but with a lower amount of objects. But the theme is similar to the previous analyzed concept.

- For hse_5282_20_sep_u:




This lattice turned out to have a very large number of objects with a large number of overlapping attributes. This concept connects the objects with interest in the news sites.

<{obj976, …, obj785}, {trud.ru, expert.ru, polit.ru, izvestia.ru, radiomayak.ru}>

**d) Provide the examples of implications A → B found by lattice diagram with the indication of their support and confidence.**

- For hse_5516_20_sep_u:

```
28 < 4 > improvement.ru ecsocman.edu.ru/ =[100 %]=> < 4 > specialist.ru/ students.ru;
```

28: Conf = 100%, Sup = 4 (15.38%)

```
42 < 7 > specialist.ru/ mifp.ru students.ru =[86 %]=> < 6 > ecsocman.edu.ru/;
```

42: Conf = 86%, Sup = 6 (23.08%)

- For hse_5289_20_sep_u:

```
< 10 > webmoney.ru cfin.ru rosbank.ru =[90 %]=> < 9 > akdi.ru;
```

10: Conf = 90%, Sup = 9 (50%)

```
< 7 > webmoney.ru cfin.ru banki.ru rosbank.ru =[86 %]=> < 6 > akdi.ru;
```

7: Conf = 86%, Sup = 6 (33.33%)

```
12 < 10 > akdi.ru cfin.ru =[100 %]=> < 10 > rosbank.ru;
13 < 6 > akdi.ru forextimes.ru/ =[100 %]=> < 6 > webmoney.ru cfin.ru rosbank.ru;
```

12: Conf = 100%, Sup = 10 (55.55%)

13: Conf = 100%, Sup = 6 (33.33%)

- For hse_5282_20_sep_u:

```
17 < 23 > gazeta.ru polit.ru =[100 %]=> < 23 > utro.ru;
18 < 27 > gazeta.ru izvestia.ru/ =[100 %]=> < 27 > utro.ru;
```

17: Conf = 100%, Sup = 23 (57.5%)

18: Conf = 100%, Sup = 27 (67.5%)

```
71 < 29 > gazeta.ru utro.ru =[93 %]=> < 27 > izvestia.ru/;
```

71: Conf = 93%, Sup = 27 (67.5%)