

# Исследование значений многозначных слов: статистический подход.

Группа ППКЛ-2024  
Компьютерная лингвистика ВШЭ 2024-2025  
Слушатель: Головченко Валерия

Факультет гуманитарных наук

# Цель:

Исследовать многозначные слова в русском языке через их примеры употребления, выявить частотные характеристики и значимость связей между ними.

Визуализировать результаты анализа для упрощения понимания частотности и взаимосвязей многозначных слов.

# Задачи:

1. Собрать данные о многозначных словах из Национального корпуса русского языка с использованием инструментов парсинга, таких как Selenium и BeautifulSoup.
2. Обработать собранные данные с помощью предобработки текста, включая токенизацию, лемматизацию и удаление стоп-слов.
3. Сформировать биграммы и пары слов для дальнейшего анализа частотности.
4. Произвести подсчёт частоты отдельных слов, пар слов и биграмм, а также визуализировать эти данные с помощью столбчатых диаграмм.
5. Генерировать графы связей между словами с использованием Graphviz, а также создавать облака слов, чтобы представить ключевые термины.
6. Провести статистический анализ, вычислить PMI и  $\chi^2$  для оценки значимости связей между многозначными словами.
7. Провести анализ размеченных данных, группируя их по значениям слов и создавая визуализации для разных значений одного и того же слова.

## Сбор данных

Использование Selenium и BeautifulSoup для парсинга примеров употребления слов из Национального корпуса русского языка.

Обработка и сохранение данных в формате CSV.

	игла	пара	каток	лист	сетка	кора	точка	кость	свет	чаща
0	и клали в них сломанные иглы, пустые тюрички (...)	Могу дать пару советов.	принято, поэтому не улица, а каток!	Порционно выкладываем на лист салата.	перпендикулярной осям цепочек плоскости регуля...	У нас в коре головного мозга есть три независимых	когда нужно было оставаться только «точкой» ум...	Вам этот Прохоров, прямо как кость в горле!	пене морей, / Всех чудес на свете милей / Ты — ...	круг постепенно начнет приобретать форму чаши.
1	голове, будто подушечка, утыканная швейными иг...	На пары ходить, лекции перечитывать, готовиться к	Тяжелый ящик, установленный на одинаковых катк...	Перечислить всех просто не хватит листа.	График бережно приводит в порядок сетки и пакеты.	родингитизированных даек, собственно родингито...	пространстве, я стала именно такой «точкой умн...	не будет легко отделяться от костей.	тем, кто уже ездил по свету, общался с людьми,...	разногласия и мелкие стычки переполняют чашу т...
2	этих несчастных заразилось через героиновую иглу.	А еще ходить на пары все, и если можно понять	В воскресенье на олимпийском катке в Хамаре (Н...	Сверху положить листья эстрагона и зеленый лук.	и армирование газона, уложив пластиковую сетку...	не облезлые бревна с содранной корой, — заботл...	нужно понять и принять свою точку зрения.	На дно кастрюли положить рубленные кости и вин...	понятно что [в] не правильном свете видят ее о...	чтобы жизнь новобрачных была полной чашей; под...

10 слов, 600 фраз.

# Предобработка текста

## Токенизация, лемматизация и удаление стоп-слов.

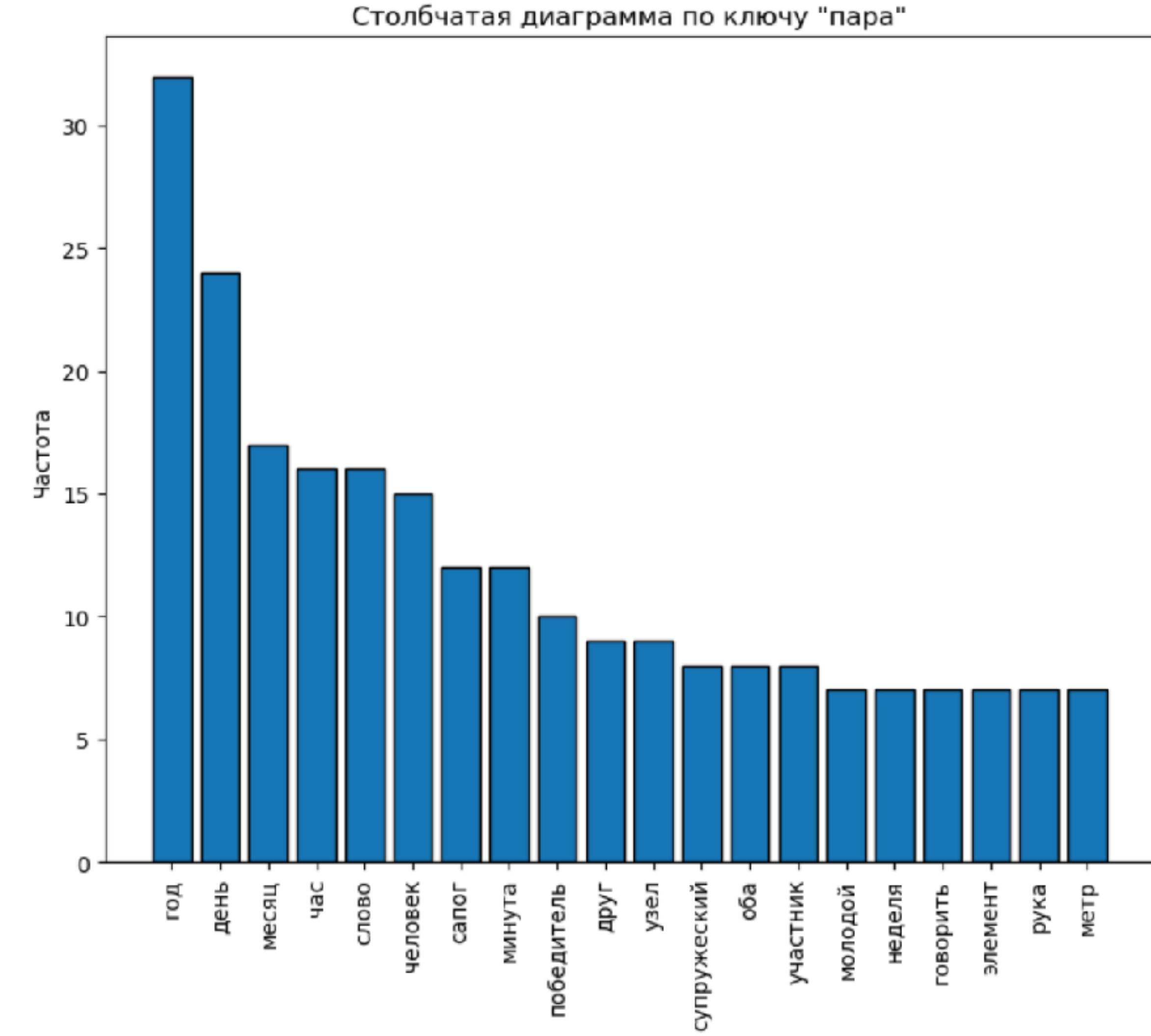
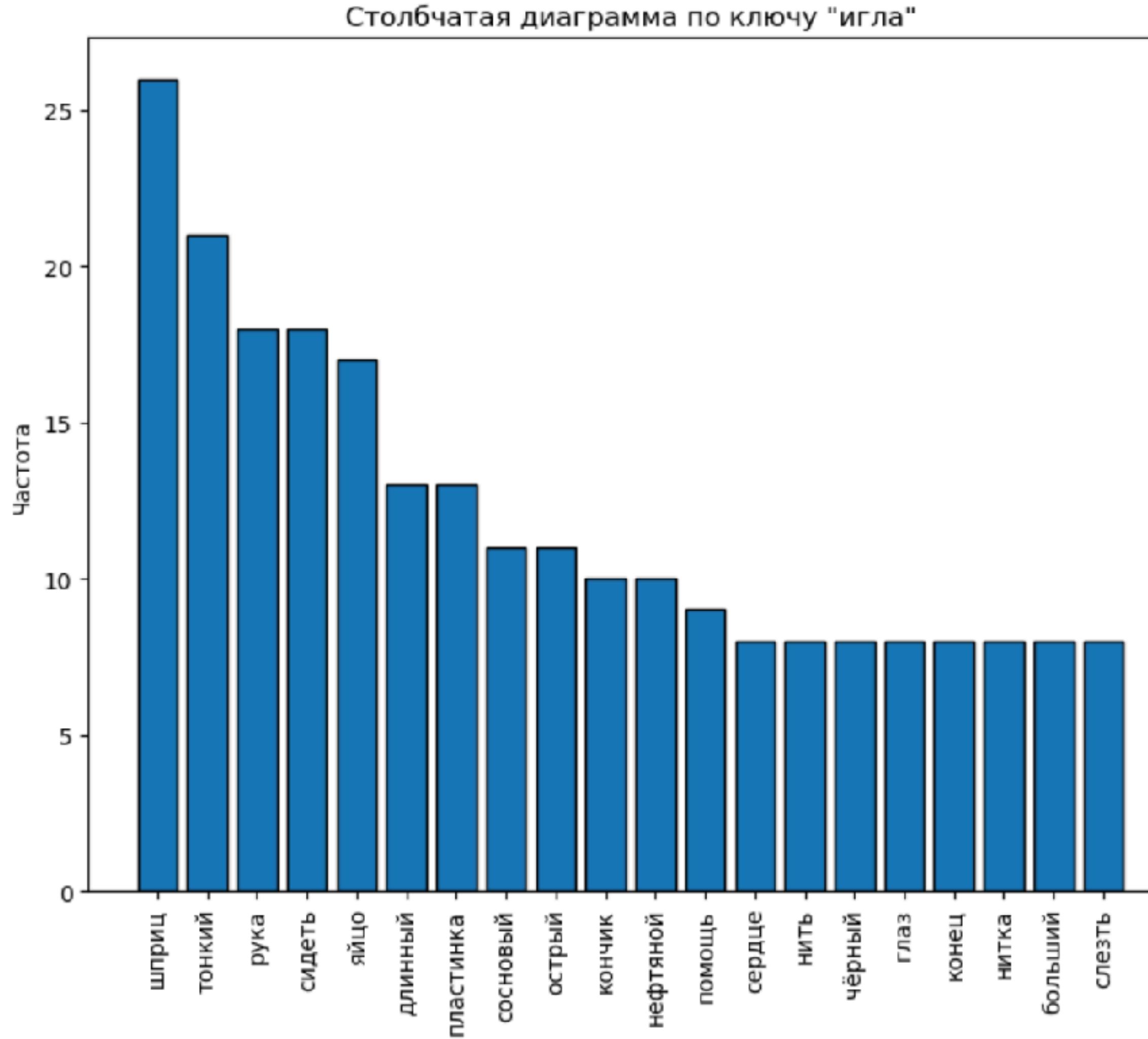
игла	пара	каток	лист	сетка	кора	точка	кость	свет	чаша
[класть, сломать, игла, пустой, тюричка, выпра...]	[дать, пара, совет]	[принять, поэтому, улица, каток]	[порционный, выкладывать, лист, салат]	[перпендикулярный, ось, цепочка, плоскость, ре...]	[кора, головной, мозг, независимый]	[нужно, оставаться, точка, умный, присутствие,...]	[прохоров, кость, горло]	[пена, море, чудо, свет, миля, убежище, мука]	[круг, постепенно, начать, приобретать, форма,...]
[голова, подушечка, утыкать, швейный, игла]	[пара, лекция, перечитывать, готовиться]	[тяжёлый, ящик, установленный, одинаковый, кат...]	[перечислить, хватить, лист]	[график, бережно, приводить, порядок, сетка, п...]	[родингитизировать, дайка, собственно, родинги...]	[пространство, точка, умный, присутствие]	[легко, отделяться, кость]	[ездить, свет, общаться, человек]	[разногласие, мелкий, стычка, чаша, терпение]
[несчастный, заразиться, героиновый, игла]	[пара, понять]	[воскресение, олимпийский, каток, хамар, зелёный, лук]	[сверху, лист, эстрагон, зелёный, лук]	[армированное, газон, уложить, пластиковый, сет...]	[облезлый, бревно, садрать, кора, заботливо, у...]	[нужно, понять, принять, точка, зрение]	[дно, кастрюля, рубить, кость, виноградный, лист]	[жизнь, новобрачный, правильный, свет]	[жизнь, полный, чаша, подруга, не...]

# Формирование биграмм и пар слов для дальнейшего анализа.

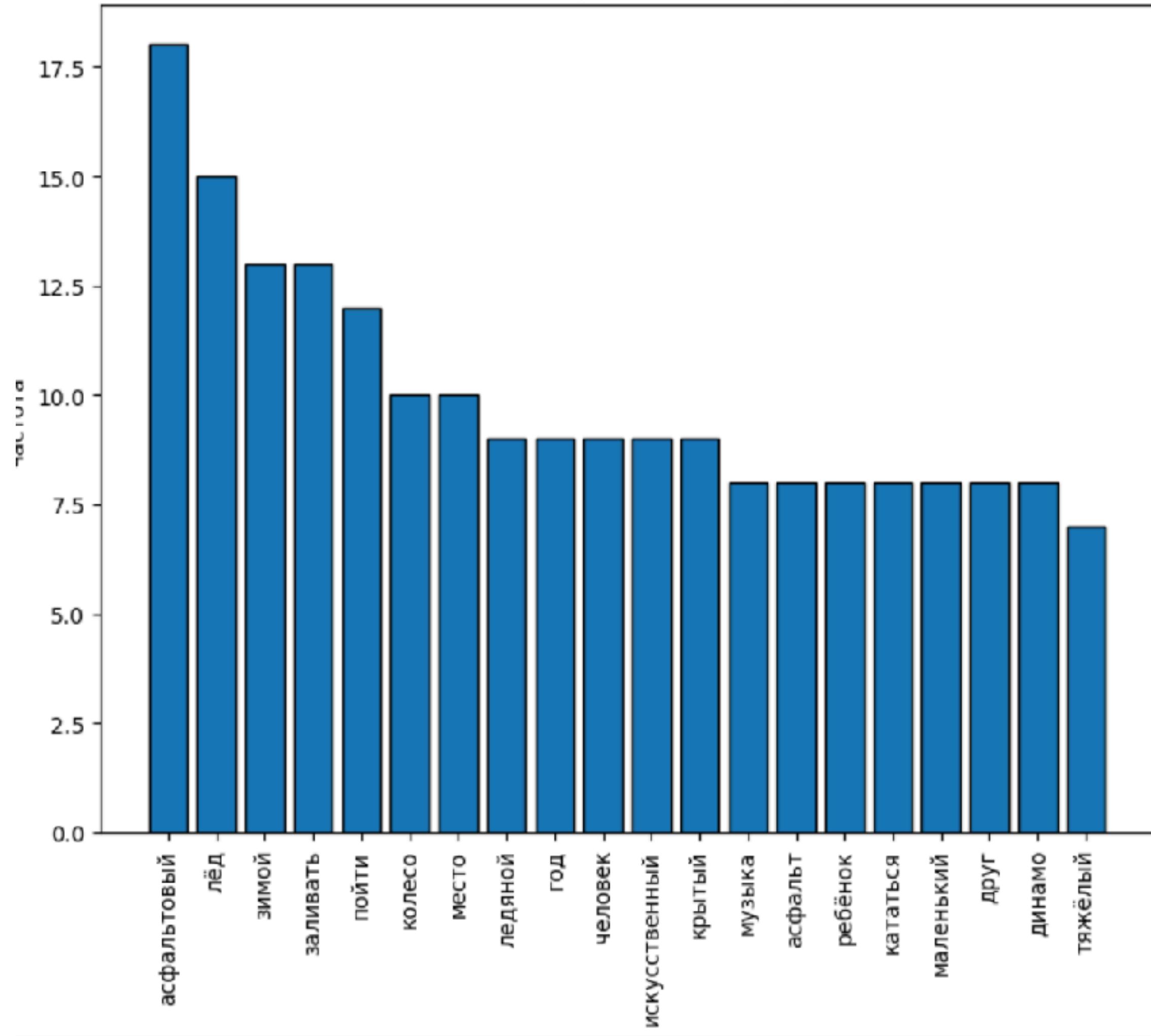
# Анализ частотности.

# Подсчёт частоты отдельных слов, пар слов и биграмм.

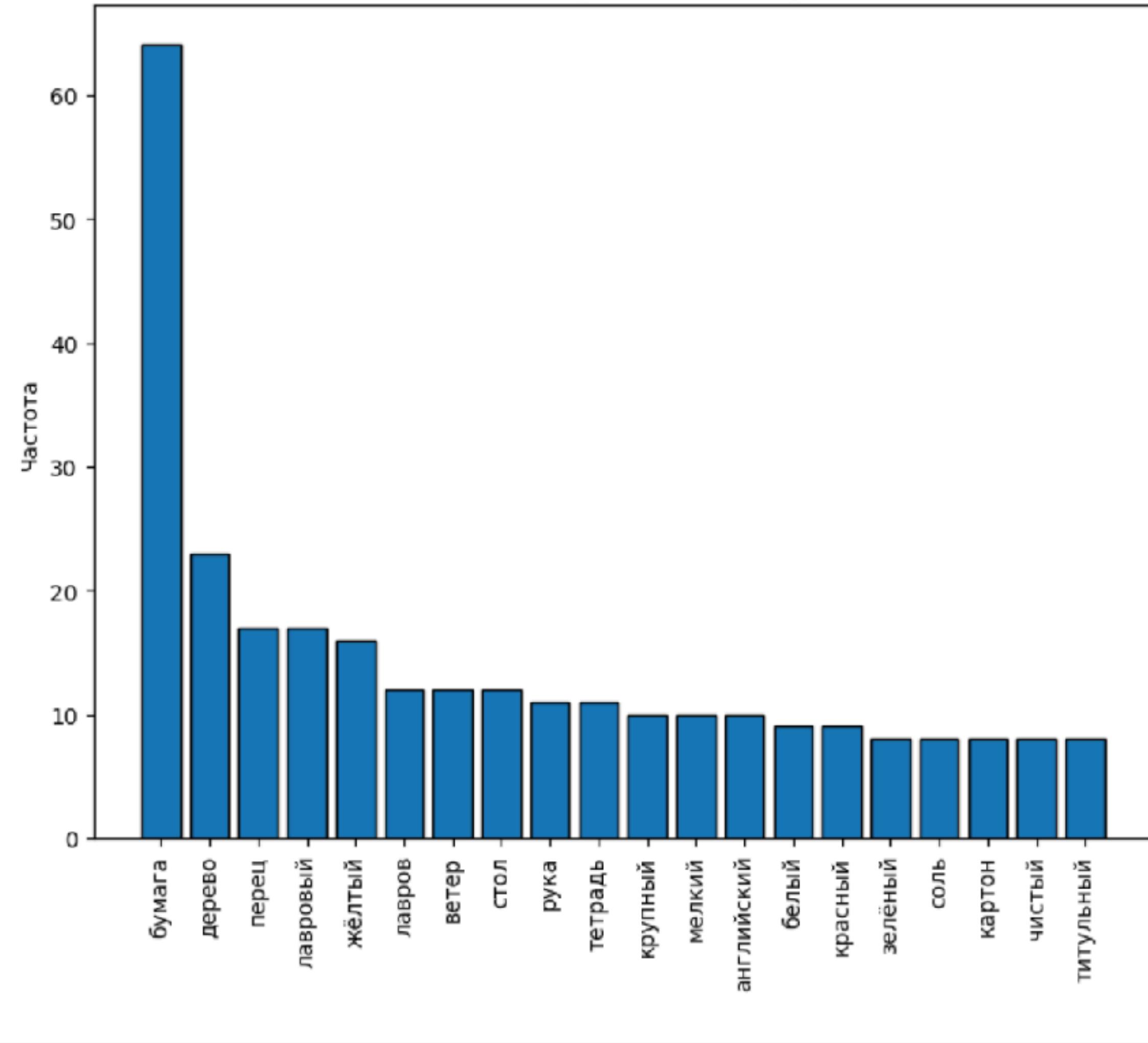
# Создание столбчатых диаграмм для визуализации частот.



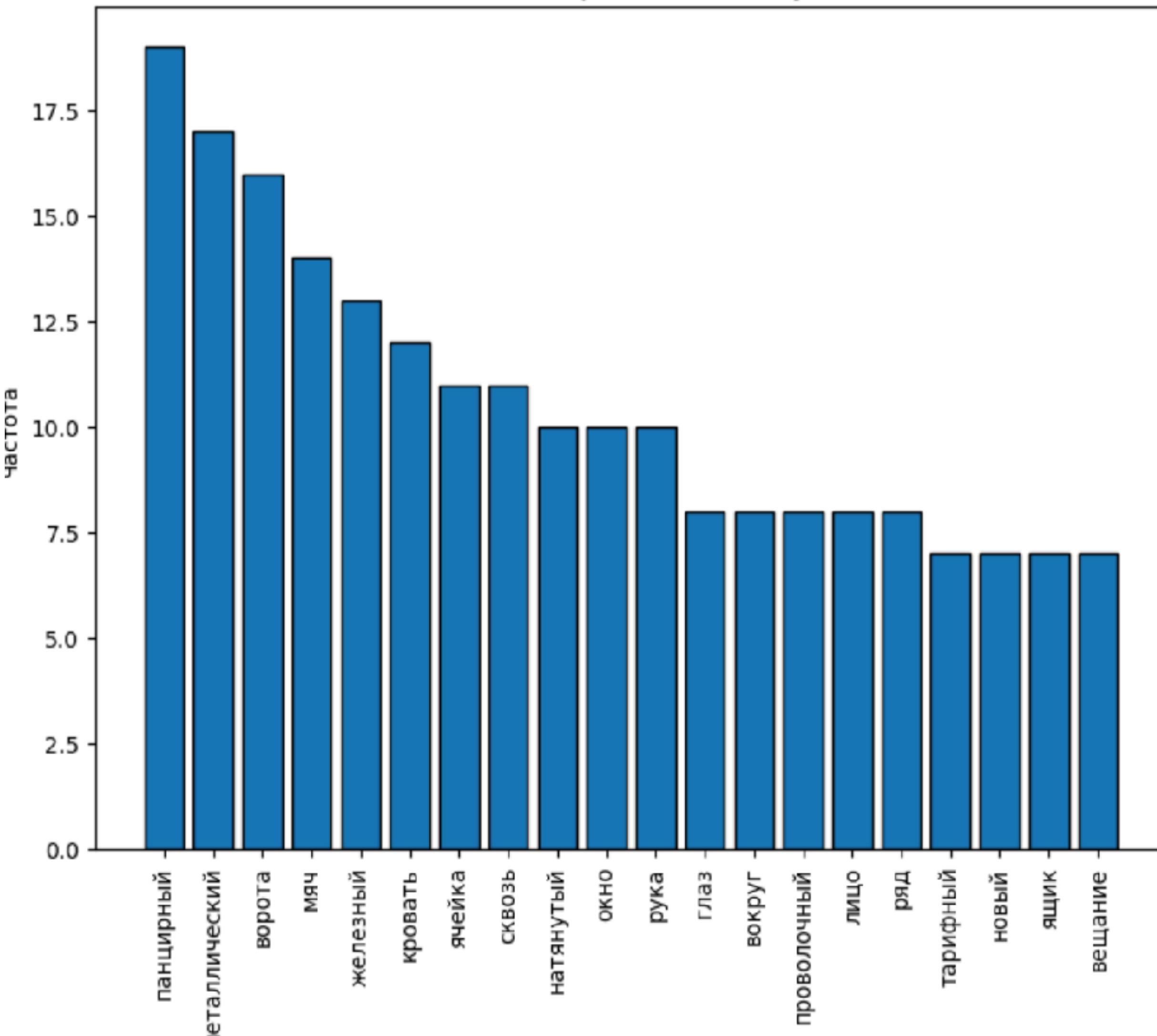
Столбчатая диаграмма по ключу "каток"



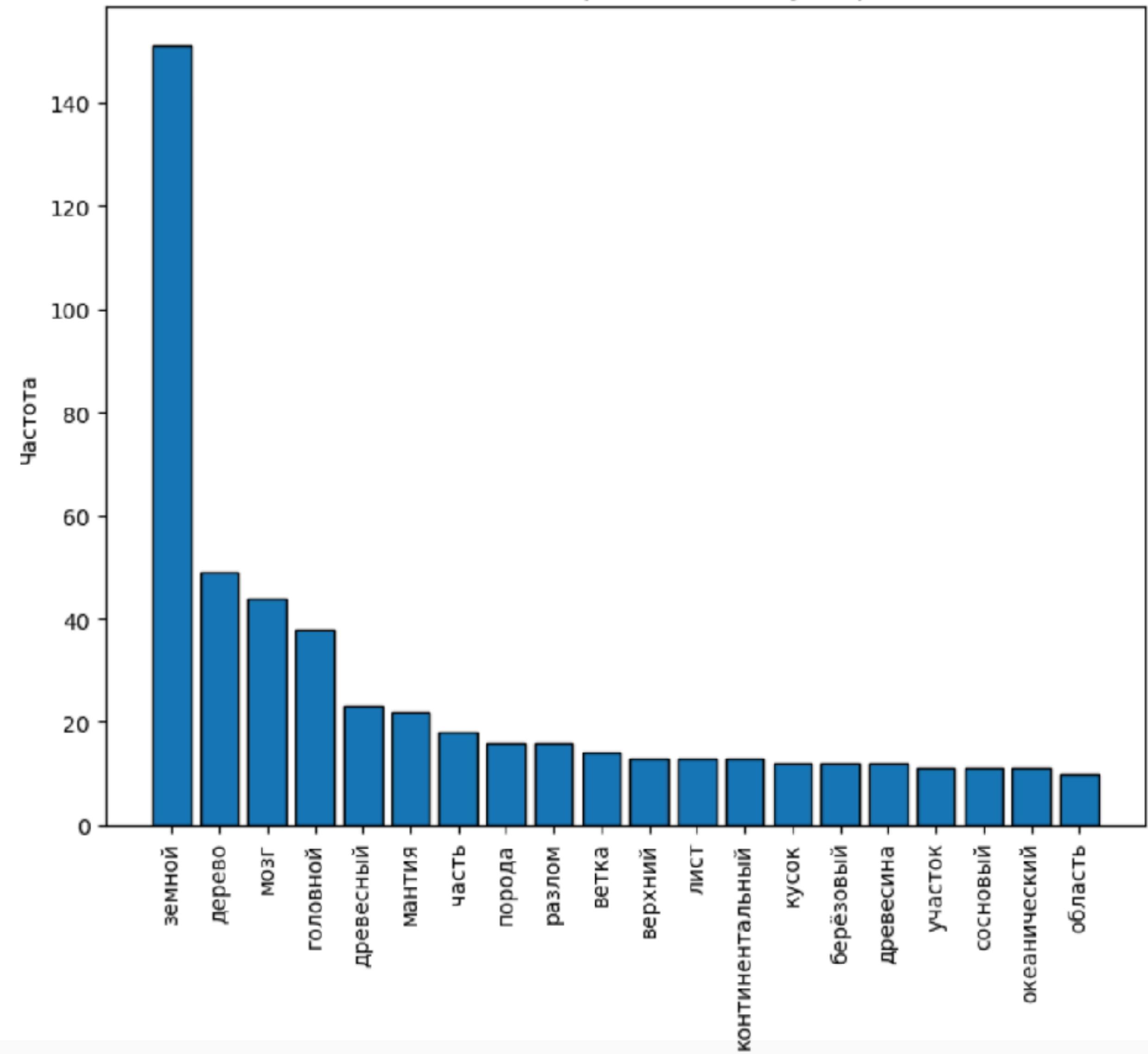
Столбчатая диаграмма по ключу "лист"



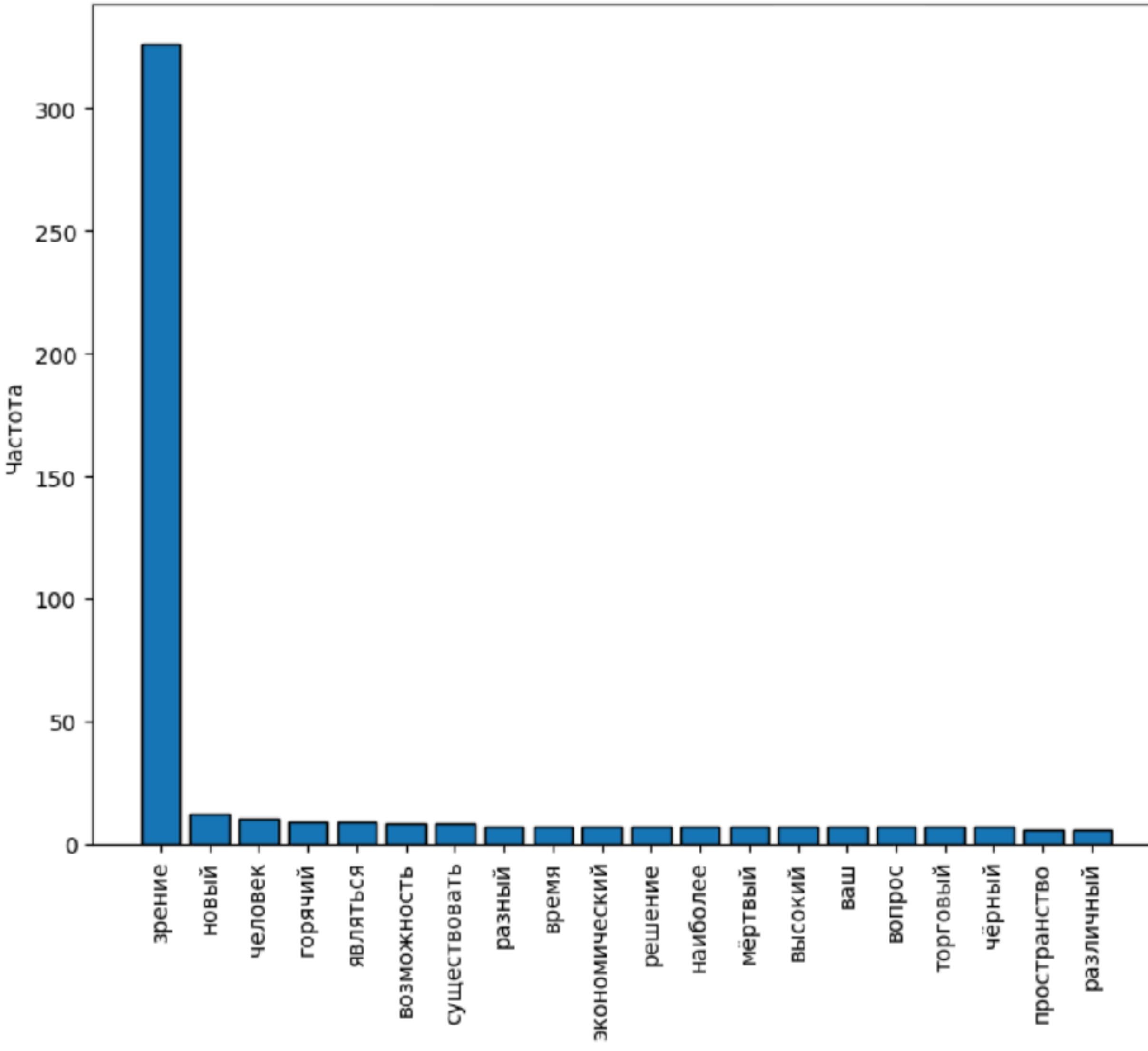
Столбчатая диаграмма по ключу "сетка"



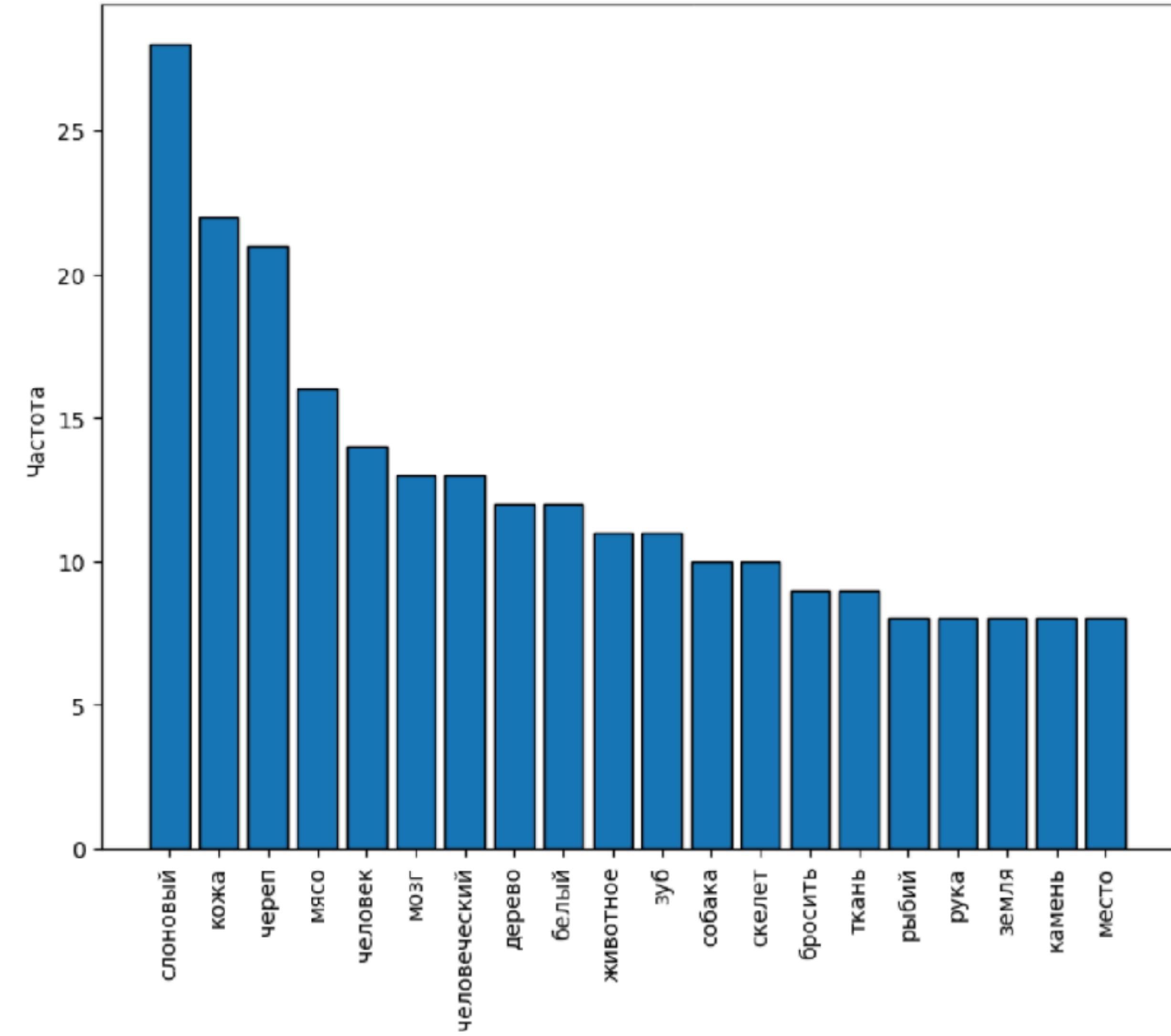
Столбчатая диаграмма по ключу "кора"



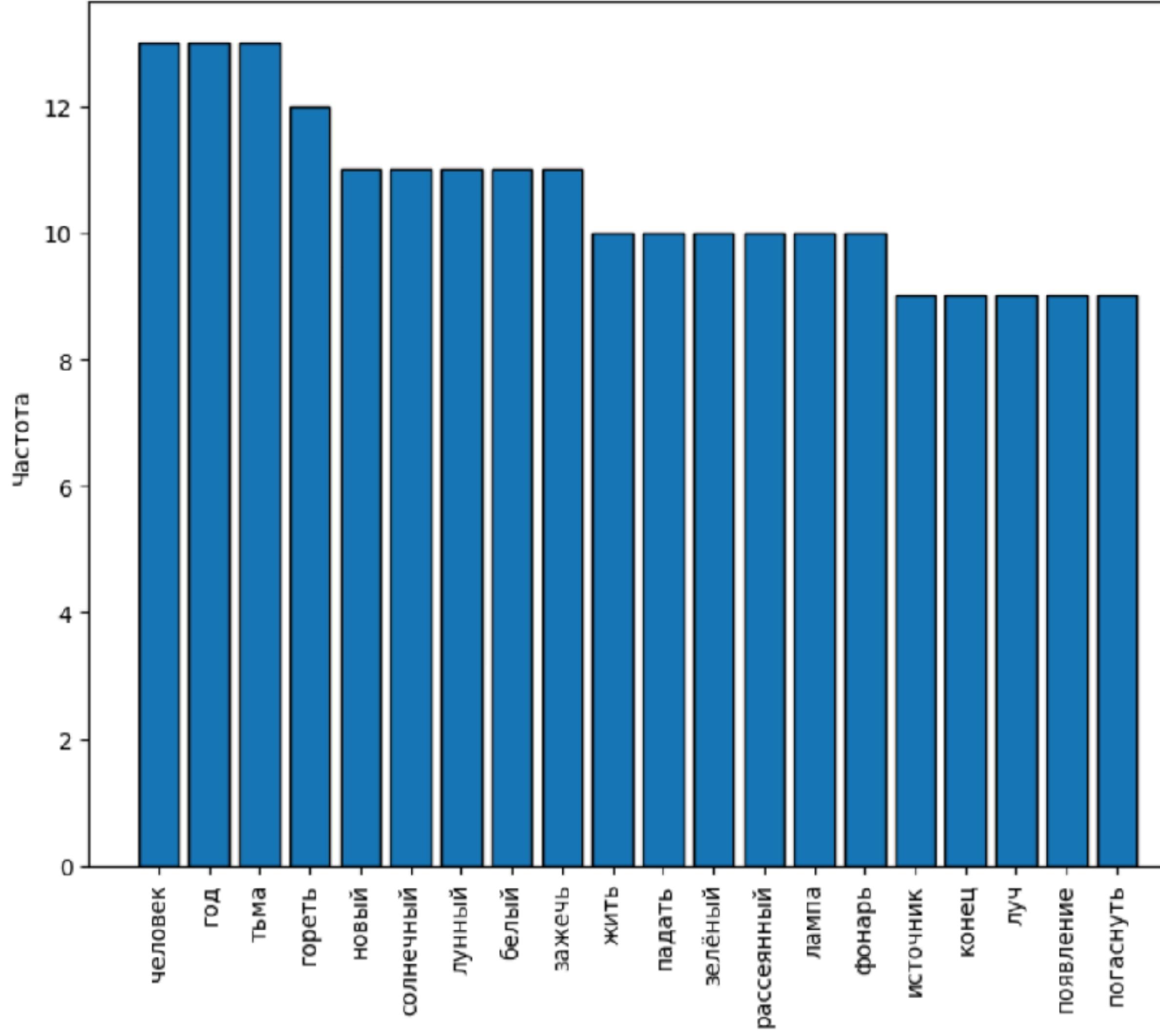
Столбчатая диаграмма по ключу "точка"



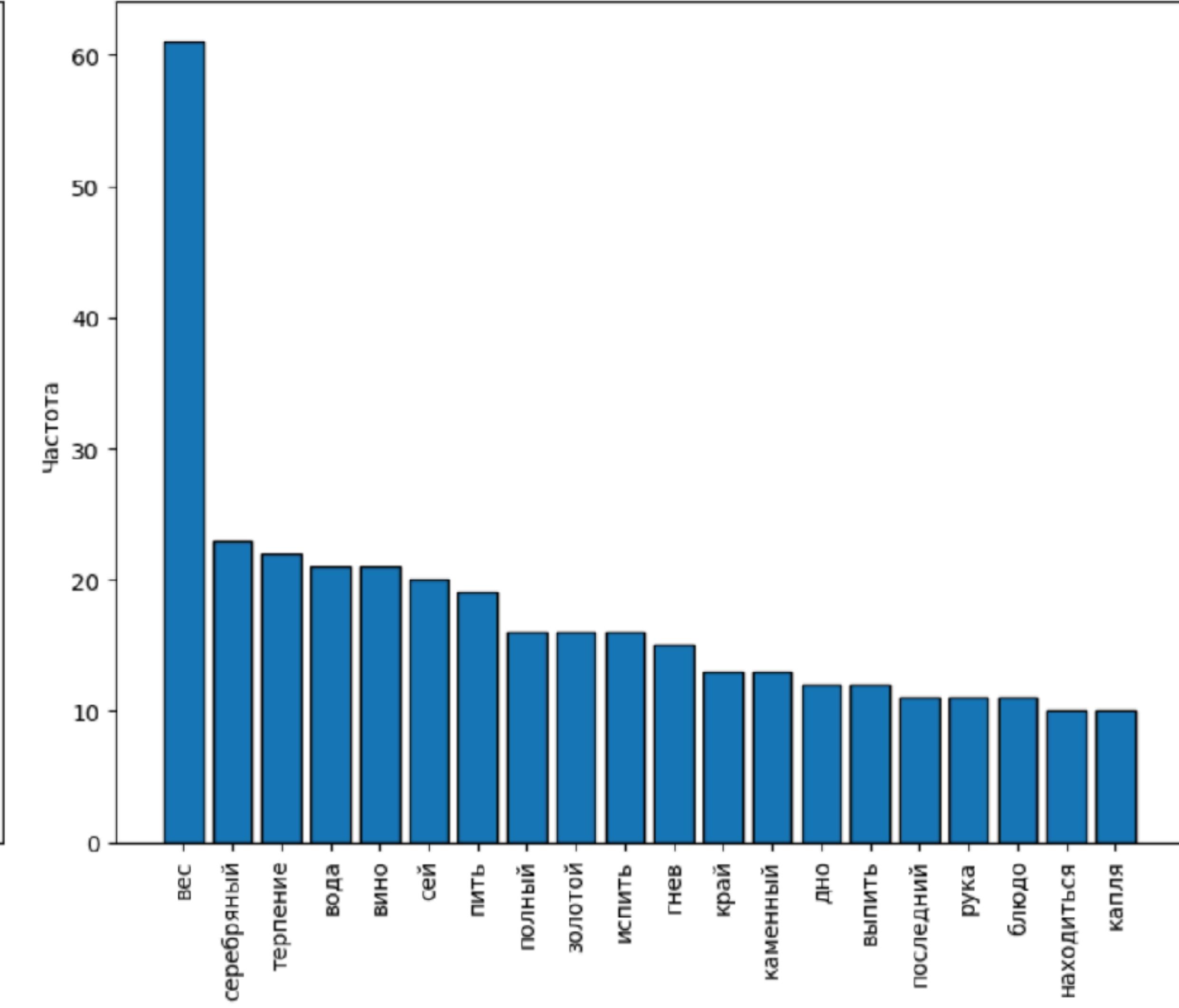
Столбчатая диаграмма по ключу "кость"



Столбчатая диаграмма по ключу "свет"



Столбчатая диаграмма по ключу "чаша"



# Создание облаков слов для ключевых терминов

Облако слов для ключа "игла"

ножницы слезть ёж  
острый сидеть глаз  
морской крючок золотой лист  
ледяной кривая шило  
сосновый крючок ушко память ткань  
войти пластинка  
нефтяной сердце конец рука  
штукатурка смерть нить  
тонкий длинный использовать

Облако слов для ключа "кора"

больший слой земля континентальный земной ветка берёзовый разлом МОЗГ место кожа Головной мантния континентальный земной ветка берёзовый разлом Мозг древесина результат часть

кусок время область порода верхний участок процесс мх сосновый древесный

Дерево институт вид океанический

Облако слов для ключа "каток"

динамо искусственный.. колесо  
пойти асфальт лёд друг<sup>г</sup>од  
прокатиться опорный конёк  
ледяной заливать кататься  
хороший пруд танк гусеница тяжёлый  
музыка вместе асфальтовый  
маленький улица человек место  
крытый скрипка  
зимой каток огромный  
время ребёнок вечер кино снег скрипка  
дорога

Образ слов для ключа "лист"  
поле газетный картон рука лавровый зелёный крупный карандаш белый тетрадь нижний  
**лиСТ** бумага лук чистый ветка лавров мелкий титульный жёлтый ветер глаза молодой  
стол перец соль дерево маршрутный сложить плотный маршрутный английский глаз последний  
перец лежать дерево маршрутный маршрутный английский глаз последний

Образ слов для ключа "свет"  
рассеянный падать зелёный  
жить человек  
белый свет  
зажечь появление лампа источник  
погаснуть фонарь луч  
гореть тьма Год  
дом новый конец  
гореть новый конец

Образ слов для ключа "сетка"  
сквозь мяч натянутый рука окно  
металлический панцирный кровать ворота ячейка  
железный вокруг тарифный ряд проволочный глаз  
вещание новый двор лицо ящик

Образ слов для ключа "пара"  
слово жизнь место понять  
час неделя человек метр  
друг метр ботинок ребёнок давать  
хороший день говорить участник победител  
месяц хочет число оставаться  
супружеский Год сапог оставаться  
оба элемента рук

Образ слов для ключа "точка"  
вопрос время разный пространство  
наиболее возможность мёртвый  
торговый существовать решение  
высокий новый горячий  
экономический чёрный

Образ слов для ключа "чаша"  
блюдо каменный испить  
полный неупиватель выпить  
золотой сейсеребряный край  
пить терпение гнев  
вес дно находиться вино рука  
кровь миновать наполнить капля

# Вычисление РМІ (Pointwise Mutual Information) и $\chi^2$ (хи-квадрат) для оценки значимости связей между словами.

пара :

```
pmi = 5.23, хи-квадрат = 49.5 p-value: 0.00 у пары ('победитель', 'динамо') : 3
pmi = 5.86, хи-квадрат = 148.6 p-value: 0.00 у пары ('познакомиться', 'оба') : 3
pmi = 4.47, хи-квадрат = 52.8 p-value: 0.00 у пары ('супружеский', 'оба') : 3
pmi = 4.47, хи-квадрат = 76.5 p-value: 0.00 у пары ('оба', 'оба') : 3
pmi = 6.44, хи-квадрат = 659.1 p-value: 0.00 у пары ('длина', 'смотреть') : 3
pmi = 4.07, хи-квадрат = 62.5 p-value: 0.00 у пары ('хотеть', 'слово') : 3
pmi = 4.61, хи-квадрат = 94.6 p-value: 0.00 у пары ('участник', 'рука') : 3
pmi = 6.44, хи-квадрат = 614.9 p-value: 0.00 у пары ('грузовой', 'поезд') : 3
pmi = 6.44, хи-квадрат = 472.1 p-value: 0.00 у пары ('грузовой', 'сутки') : 3
pmi = 6.44, хи-квадрат = 948.7 p-value: 0.00 у пары ('поезд', 'сутки') : 3
```

каток :

```
pmi = 4.17, хи-квадрат = 165.8 p-value: 0.00 у пары ('зимой', 'заливать') : 6
pmi = 3.99, хи-квадрат = 130.2 p-value: 0.00 у пары ('искусственный', 'лёд') : 4
pmi = 5.05, хи-квадрат = 402.2 p-value: 0.00 у пары ('патриарший', 'пруд') : 3
pmi = 6.12, хи-квадрат = 334.4 p-value: 0.00 у пары ('теннисный', 'корт') : 3
pmi = 5.43, хи-квадрат = 174.2 p-value: 0.00 у пары ('гладкий', 'ледовый') : 3
pmi = 5.56, хи-квадрат = 142.7 p-value: 0.00 у пары ('фильм', 'скрипка') : 3
pmi = 6.41, хи-квадрат = 673.0 p-value: 0.00 у пары ('ярко', 'осветить') : 3
```

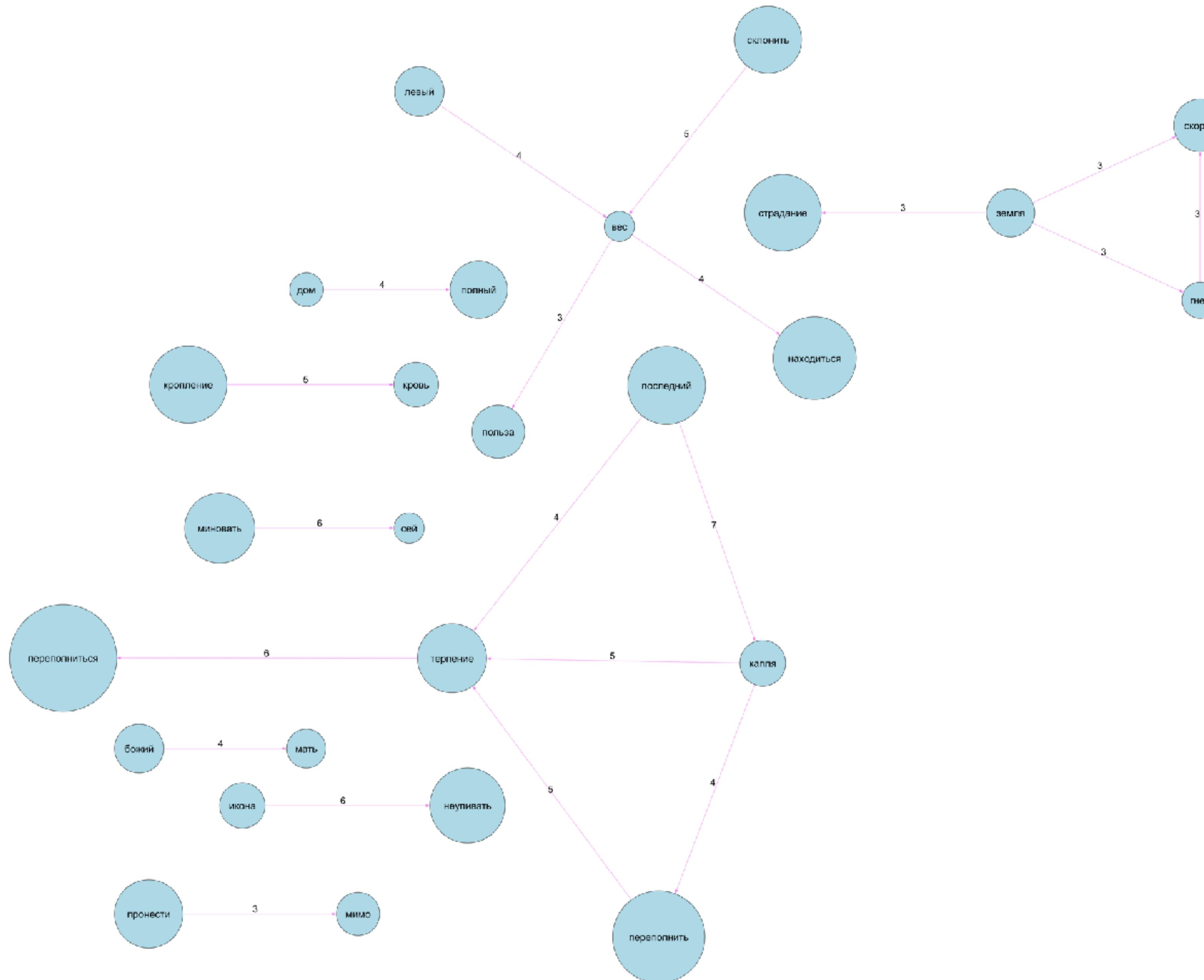
кость :

```
pmi = 4.17, хи-квадрат = 88.9 p-value: 0.00 у пары ('башня', 'слоновый') : 3
pmi = 5.89, хи-квадрат = 587.3 p-value: 0.00 у пары ('мёртвый', 'идея') : 4
pmi = 7.50, хи-квадрат = 1760.1 p-value: 0.00 у пары ('электрический', 'провод') : 3
pmi = 6.40, хи-квадрат = 1172.1 p-value: 0.00 у пары ('провод', 'провод') : 3
pmi = 7.50, хи-квадрат = 1760.1 p-value: 0.00 у пары ('провод', 'опознать') : 3
pmi = 4.17, хи-квадрат = 110.5 p-value: 0.00 у пары ('череп', 'скрестить') : 3
pmi = 6.29, хи-квадрат = 158.0 p-value: 0.00 у пары ('мужской', 'род') : 3
pmi = 5.01, хи-квадрат = 196.6 p-value: 0.00 у пары ('металл', 'дерево') : 5
pmi = 5.42, хи-квадрат = 196.6 p-value: 0.00 у пары ('металл', 'камень') : 5
pmi = 4.32, хи-квадрат = 80.2 p-value: 0.00 у пары ('дерево', 'камень') : 4
pmi = 5.01, хи-квадрат = 105.9 p-value: 0.00 у пары ('резьба', 'дерево') : 3
pmi = 5.89, хи-квадрат = 158.6 p-value: 0.00 у пары ('металл', 'стекло') : 4
pmi = 3.61, хи-квадрат = 43.5 p-value: 0.00 у пары ('цвет', 'слоновый') : 4
pmi = 5.20, хи-квадрат = 184.9 p-value: 0.00 у пары ('материал', 'железо') : 3
pmi = 4.59, хи-квадрат = 45.9 p-value: 0.00 у пары ('различный', 'животное') : 3
pmi = 5.89, хи-квадрат = 203.3 p-value: 0.00 у пары ('цветной', 'металл') : 3
pmi = 5.01, хи-квадрат = 73.4 p-value: 0.00 у пары ('цветной', 'дерево') : 3
pmi = 5.42, хи-квадрат = 73.4 p-value: 0.00 у пары ('цветной', 'камень') : 3
pmi = 6.11, хи-квадрат = 99.3 p-value: 0.00 у пары ('цветной', 'стекло') : 3
pmi = 4.73, хи-квадрат = 54.3 p-value: 0.00 у пары ('дерево', 'стекло') : 3
pmi = 5.13, хи-квадрат = 113.2 p-value: 0.00 у пары ('камень', 'стекло') : 3
```

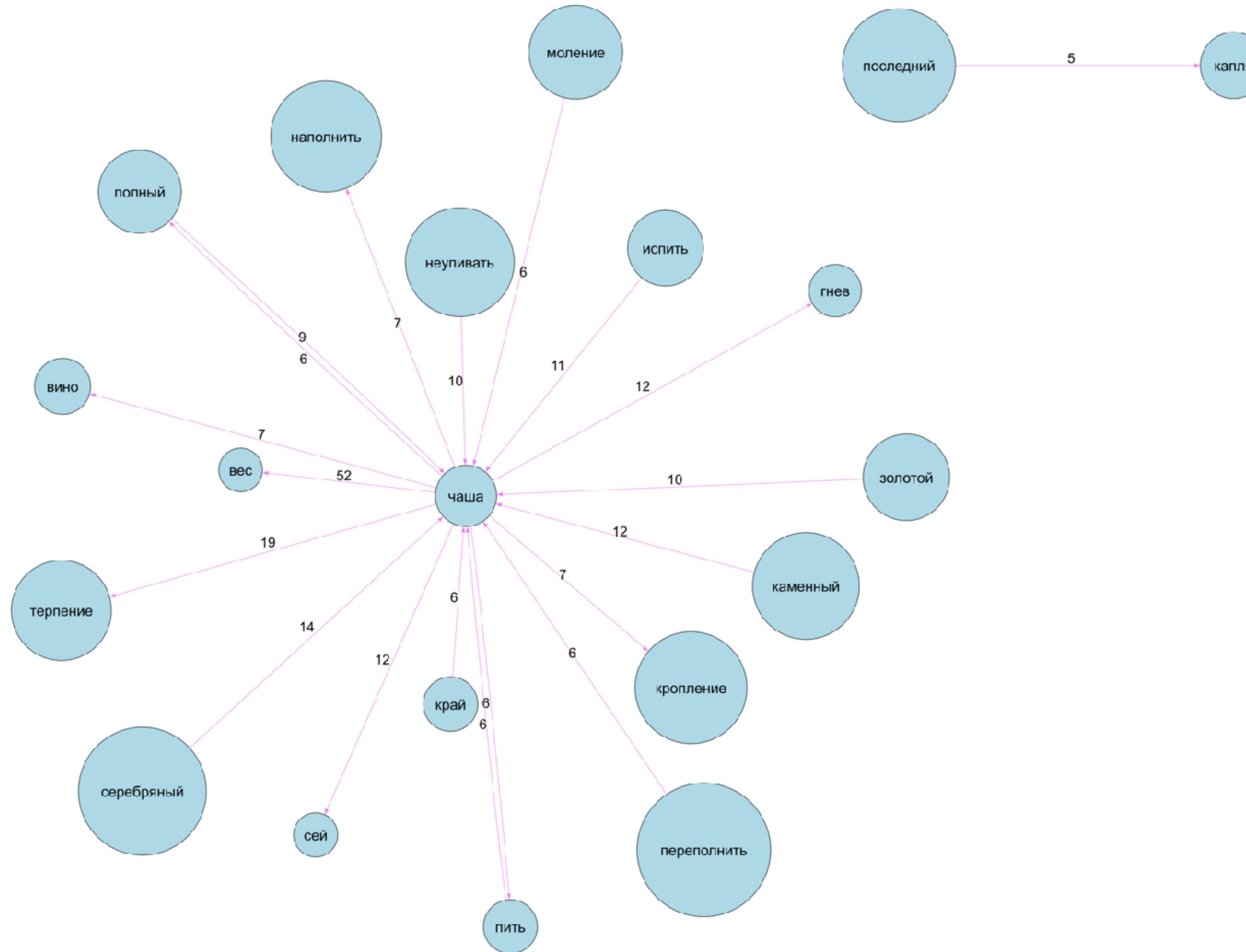
свет :

```
pmi = 5.75, хи-квадрат = 333.3 p-value: 0.00 у пары ('комбинационный', 'рассеяние') : 4
pmi = 4.95, хи-квадрат = 107.5 p-value: 0.00 у пары ('рассеянный', 'частота') : 3
pmi = 6.15, хи-квадрат = 343.9 p-value: 0.00 у пары ('длина', 'волна') : 4
pmi = 5.64, хи-квадрат = 184.1 p-value: 0.00 у пары ('люминесцентный', 'лампа') : 3
pmi = 4.83, хи-квадрат = 338.5 p-value: 0.00 у пары ('лампа', 'дневный') : 4
pmi = 5.34, хи-квадрат = 105.4 p-value: 0.00 у пары ('линза', 'источник') : 3
pmi = 5.46, хи-квадрат = 142.0 p-value: 0.00 у пары ('сознание', 'освещать') : 3
pmi = 3.95, хи-квадрат = 96.3 p-value: 0.00 у пары ('дом', 'гореть') : 3
pmi = 5.14, хи-квадрат = 389.3 p-value: 0.00 у пары ('лунный', 'санкхие') : 4
pmi = 4.69, хи-квадрат = 115.4 p-value: 0.00 у пары ('светить', 'тьма') : 3
```

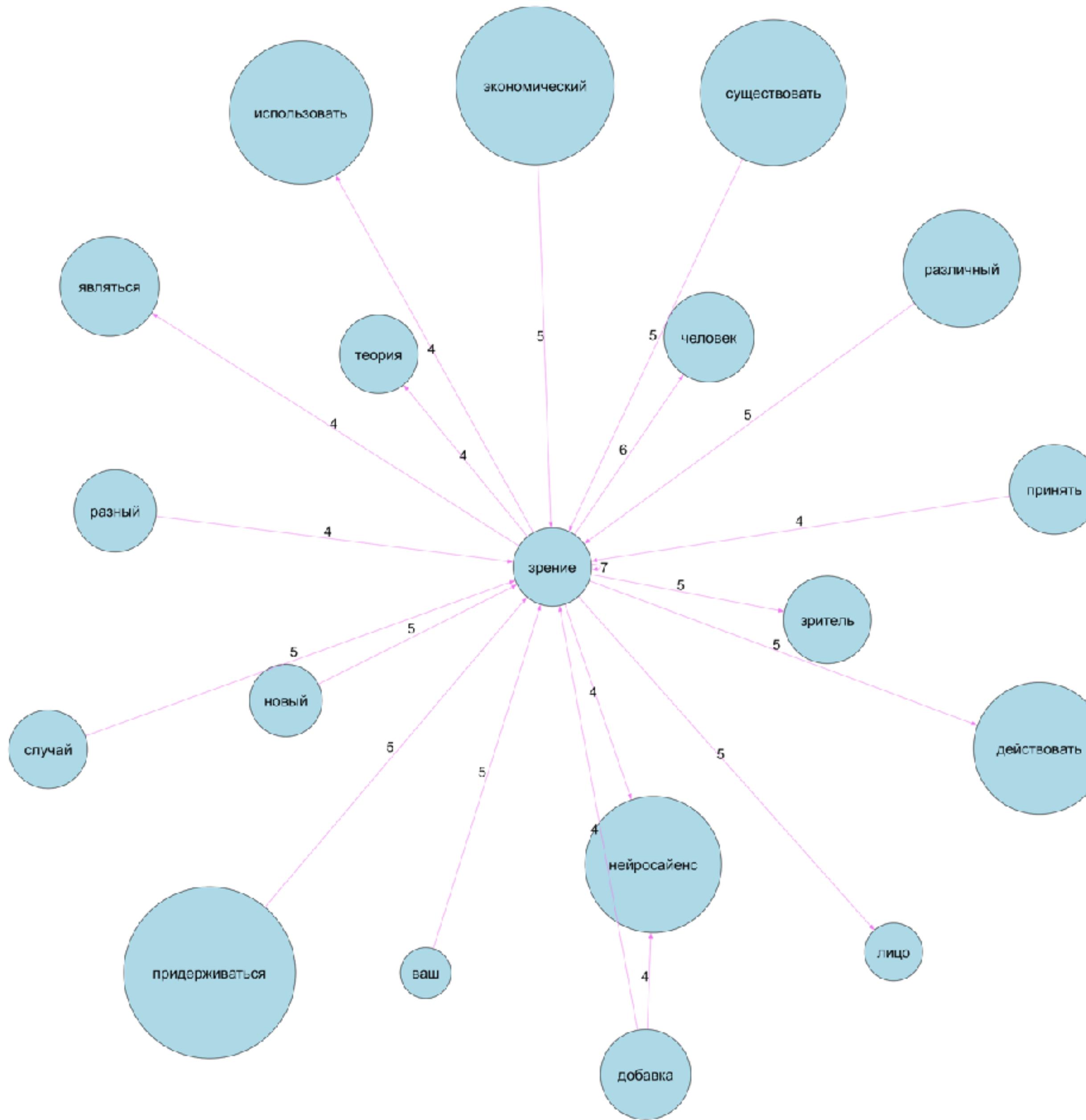
# Без ключевого слова.



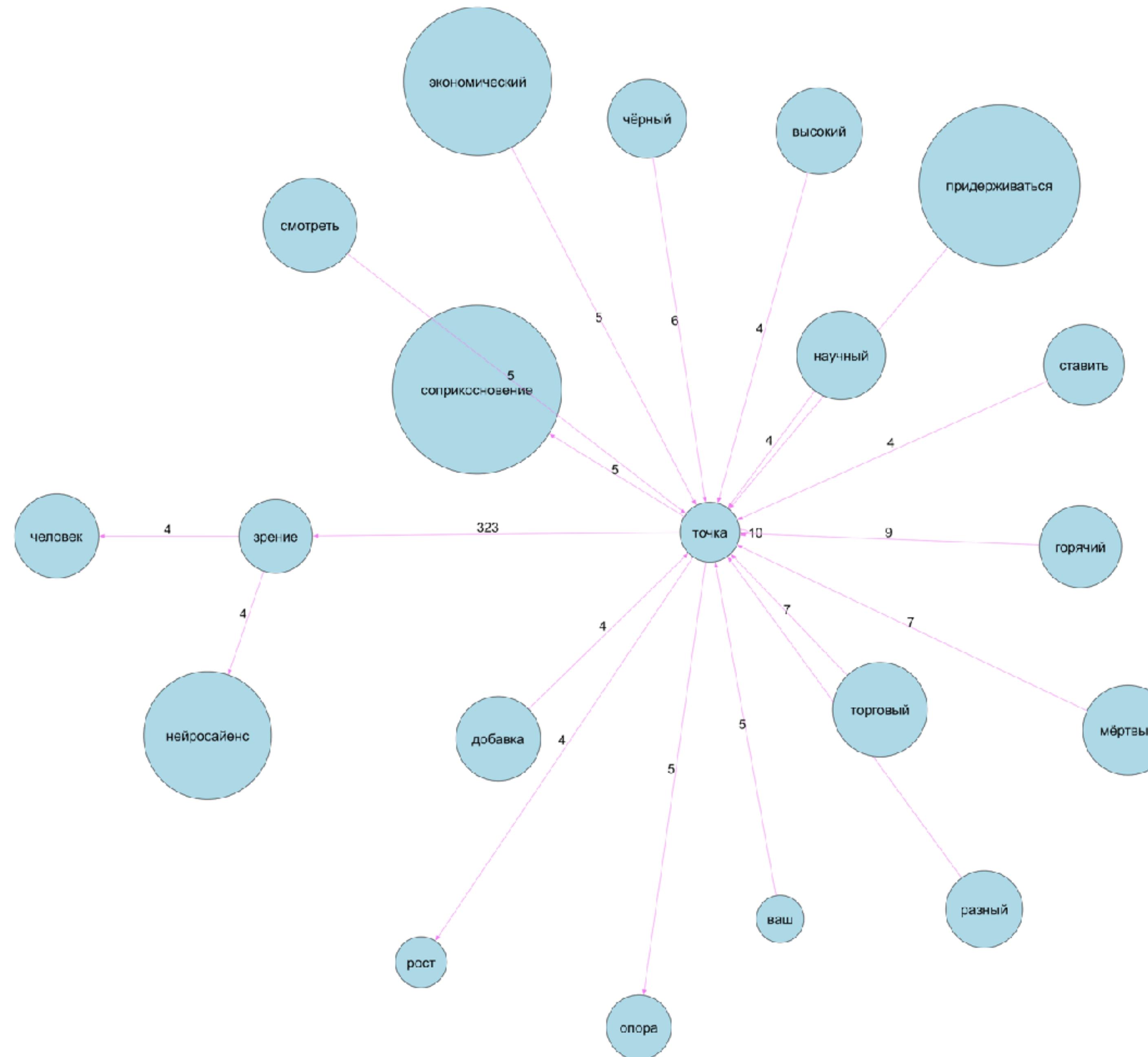
# С ключевым словом.



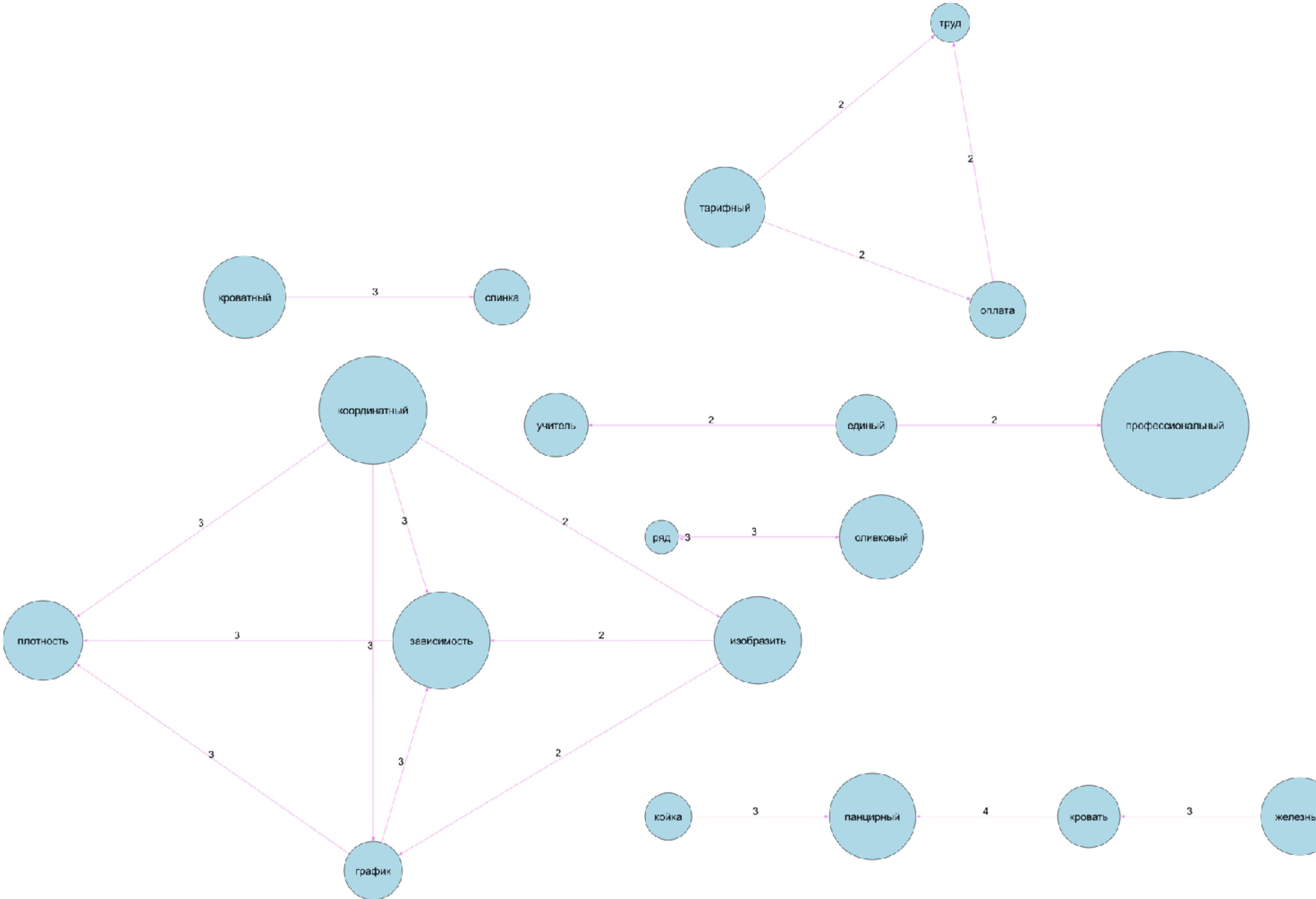
# Без ключевого слова.



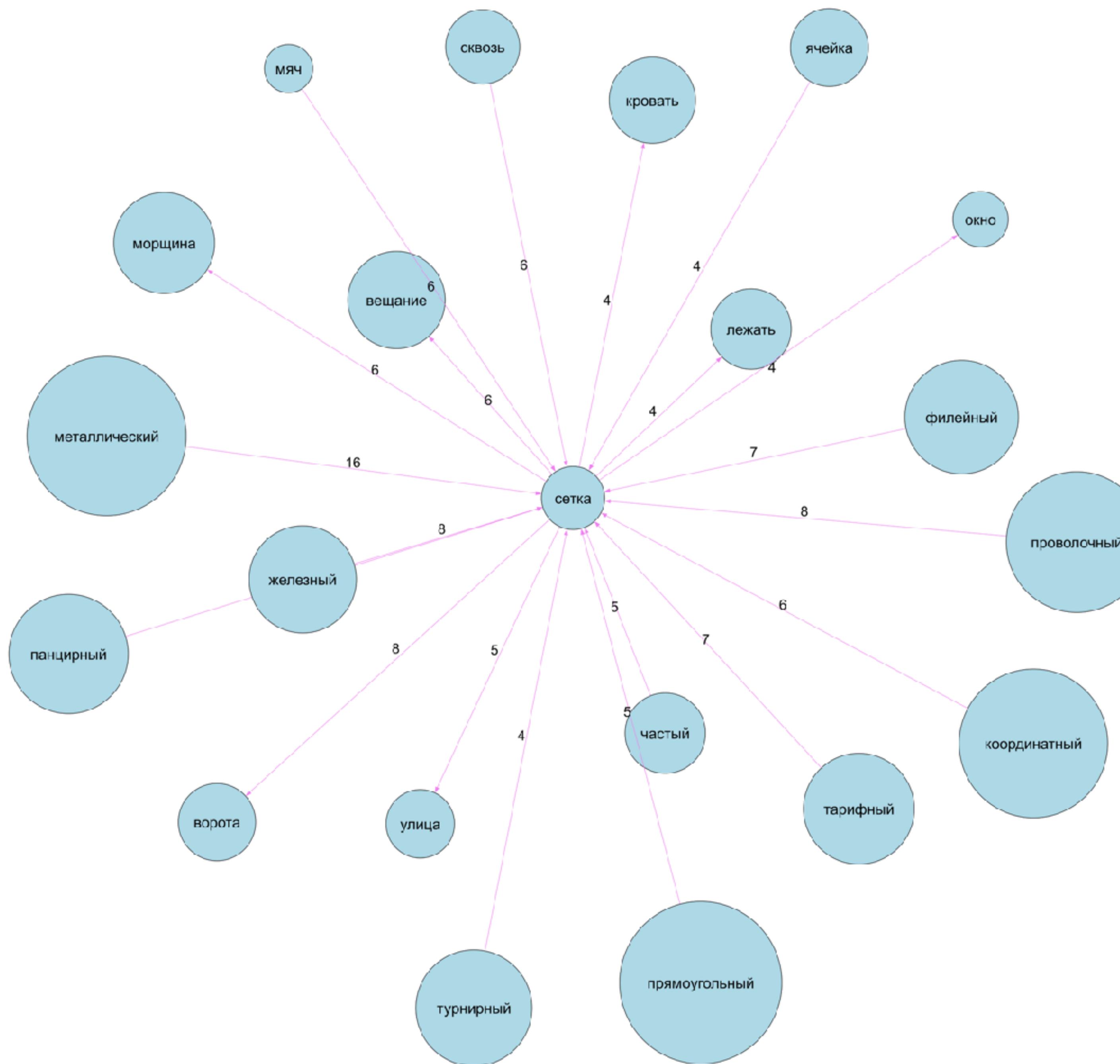
# С ключевым словом.



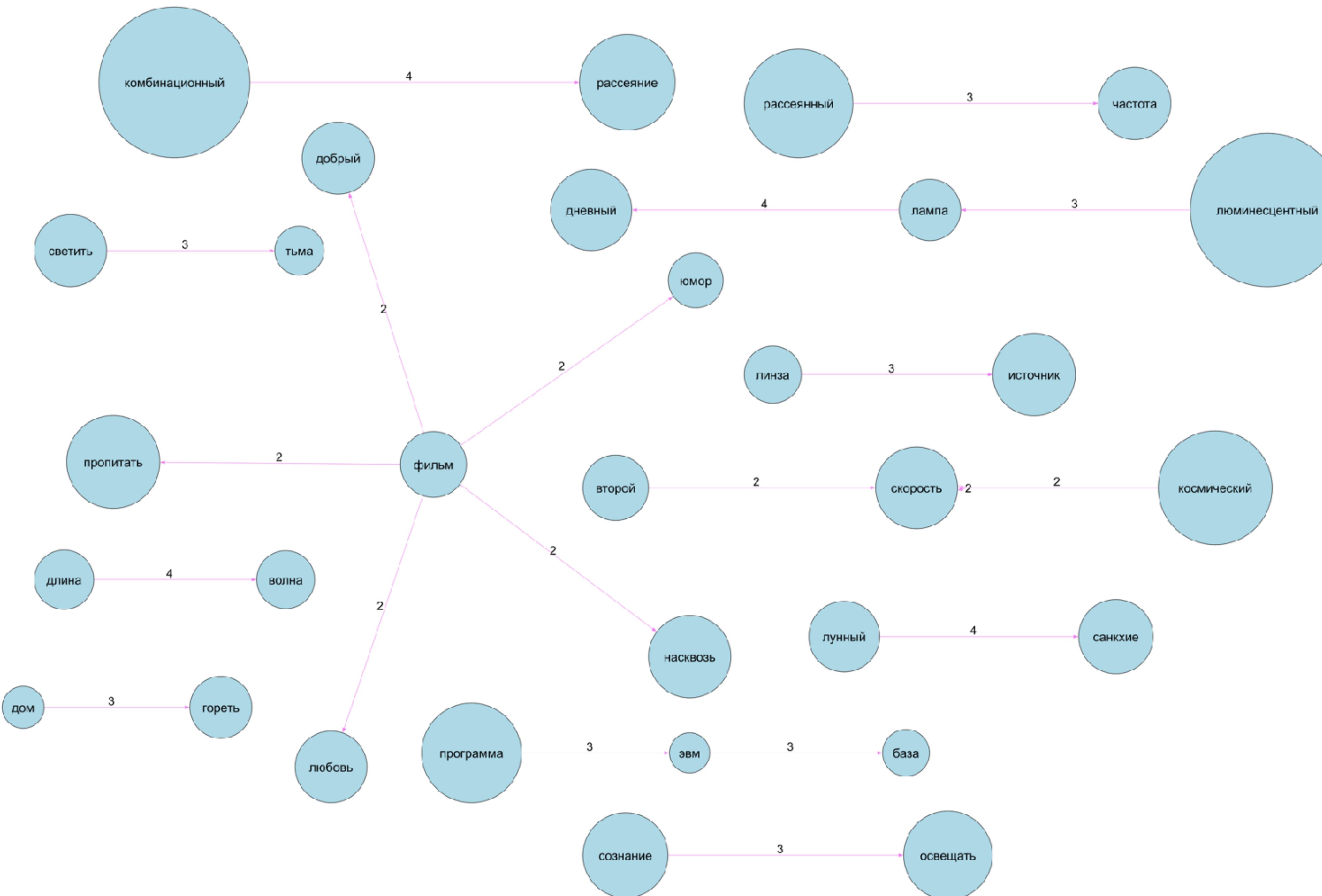
# Без ключевого слова.



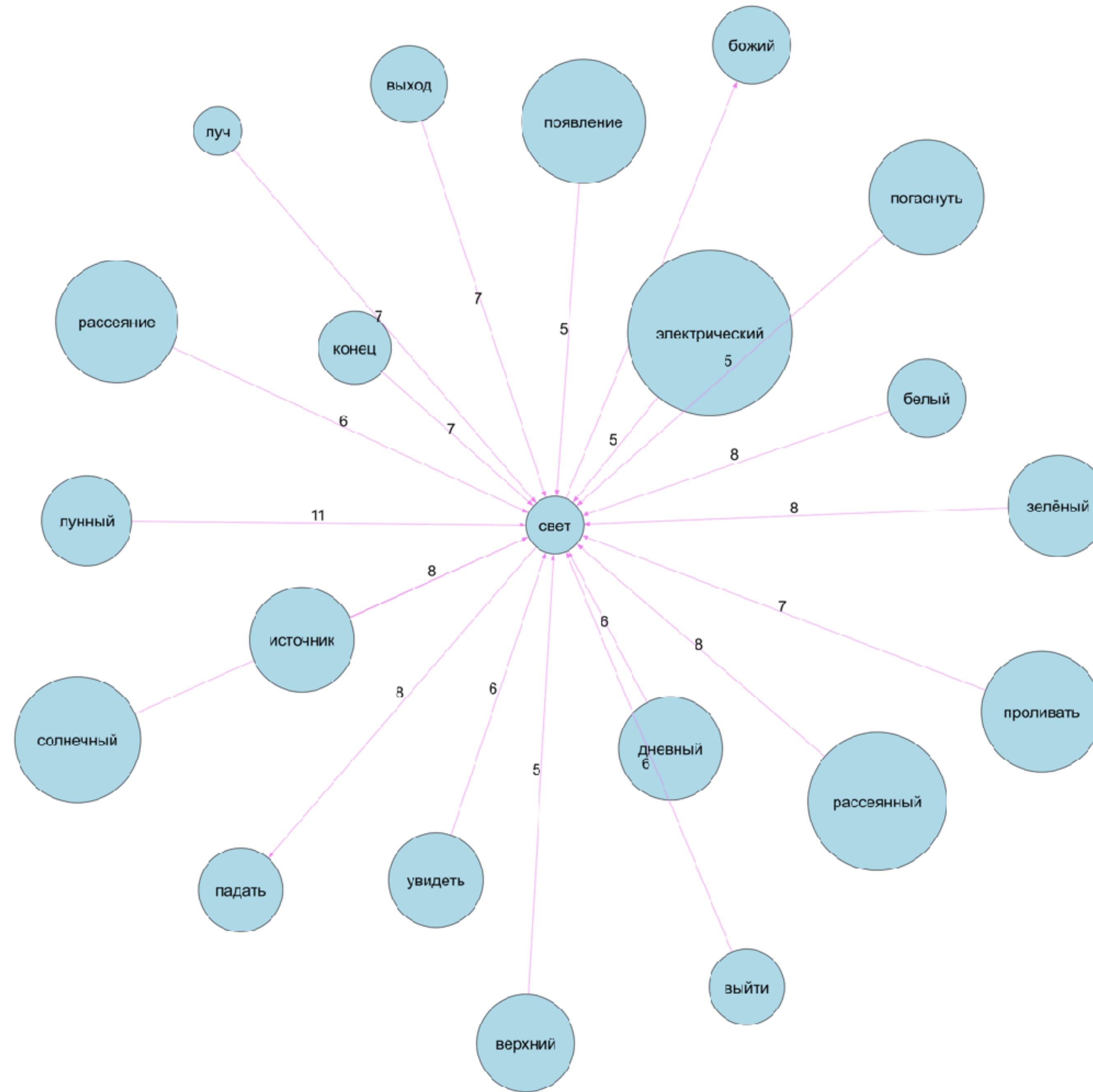
# С ключевым словом.



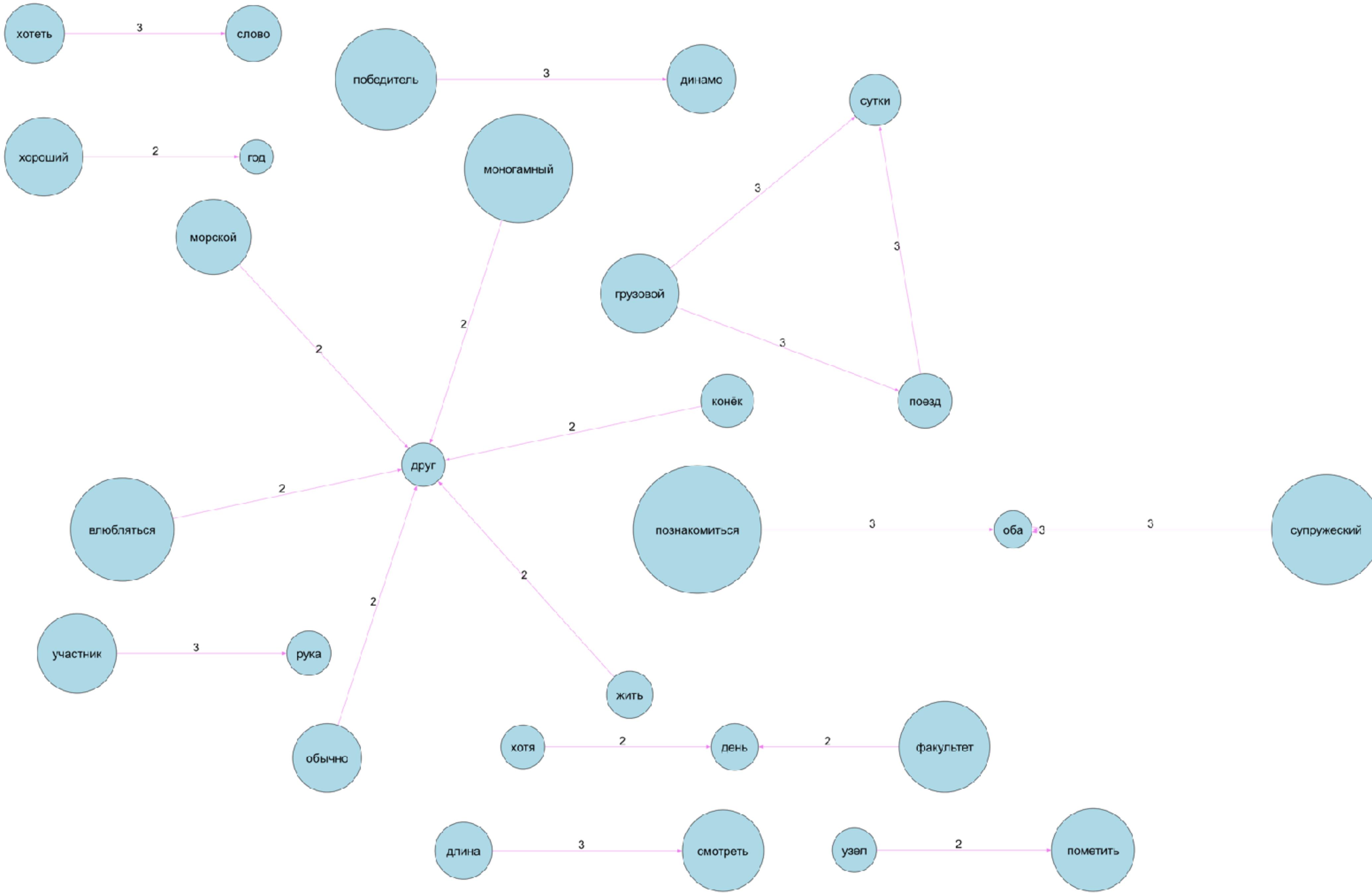
# Без ключевого слова.



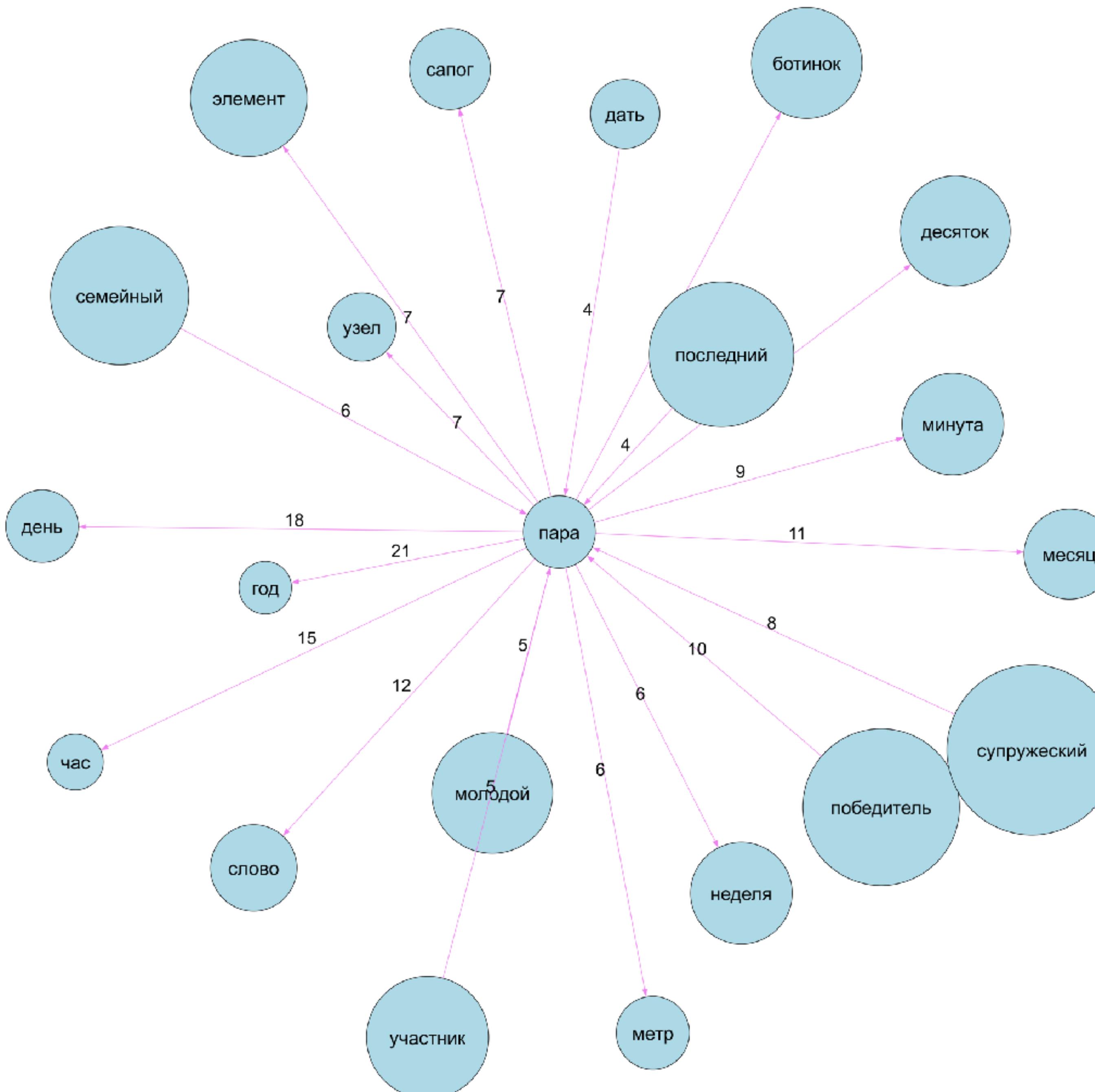
# С ключевым словом.



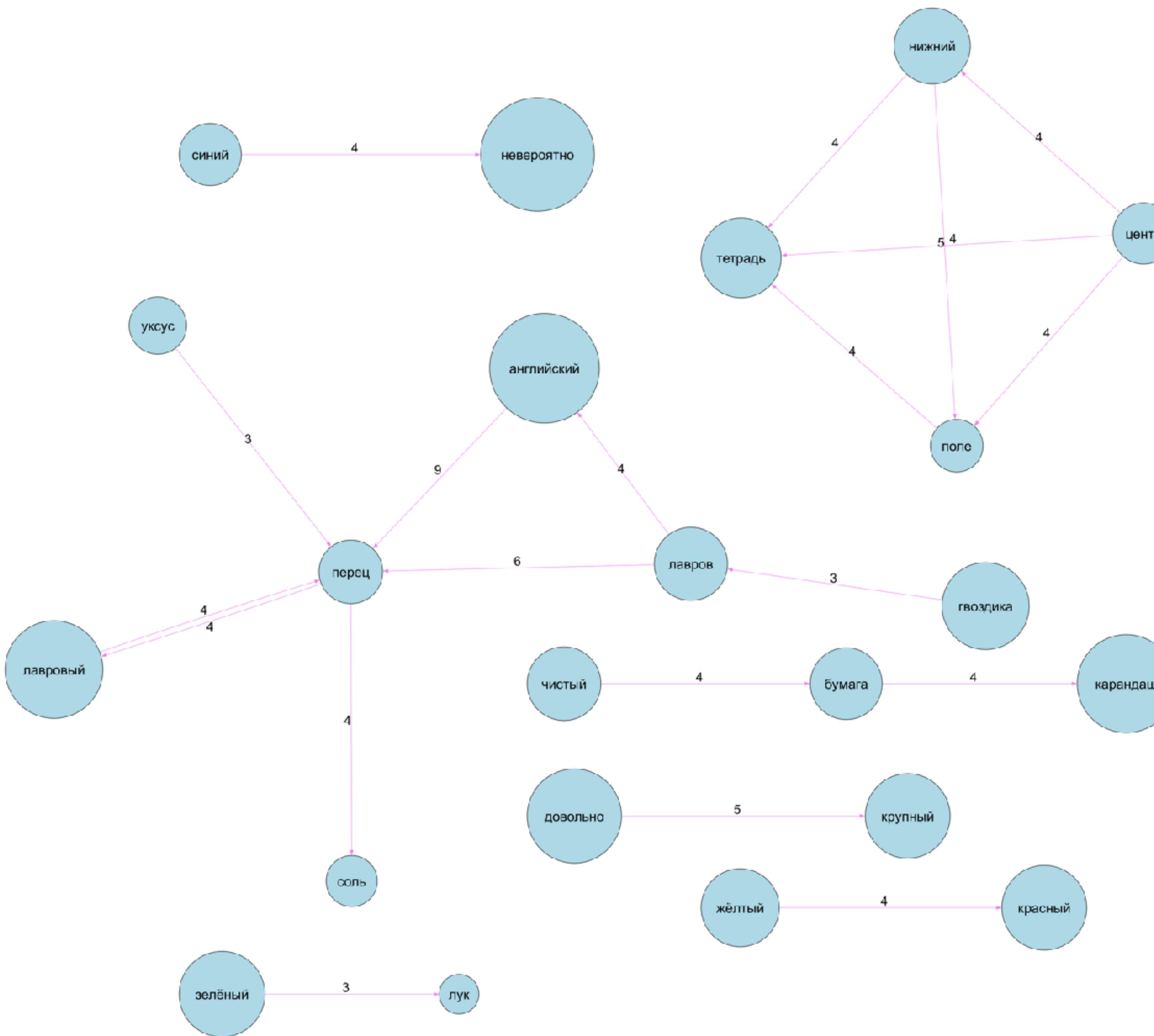
# Без ключевого слова.



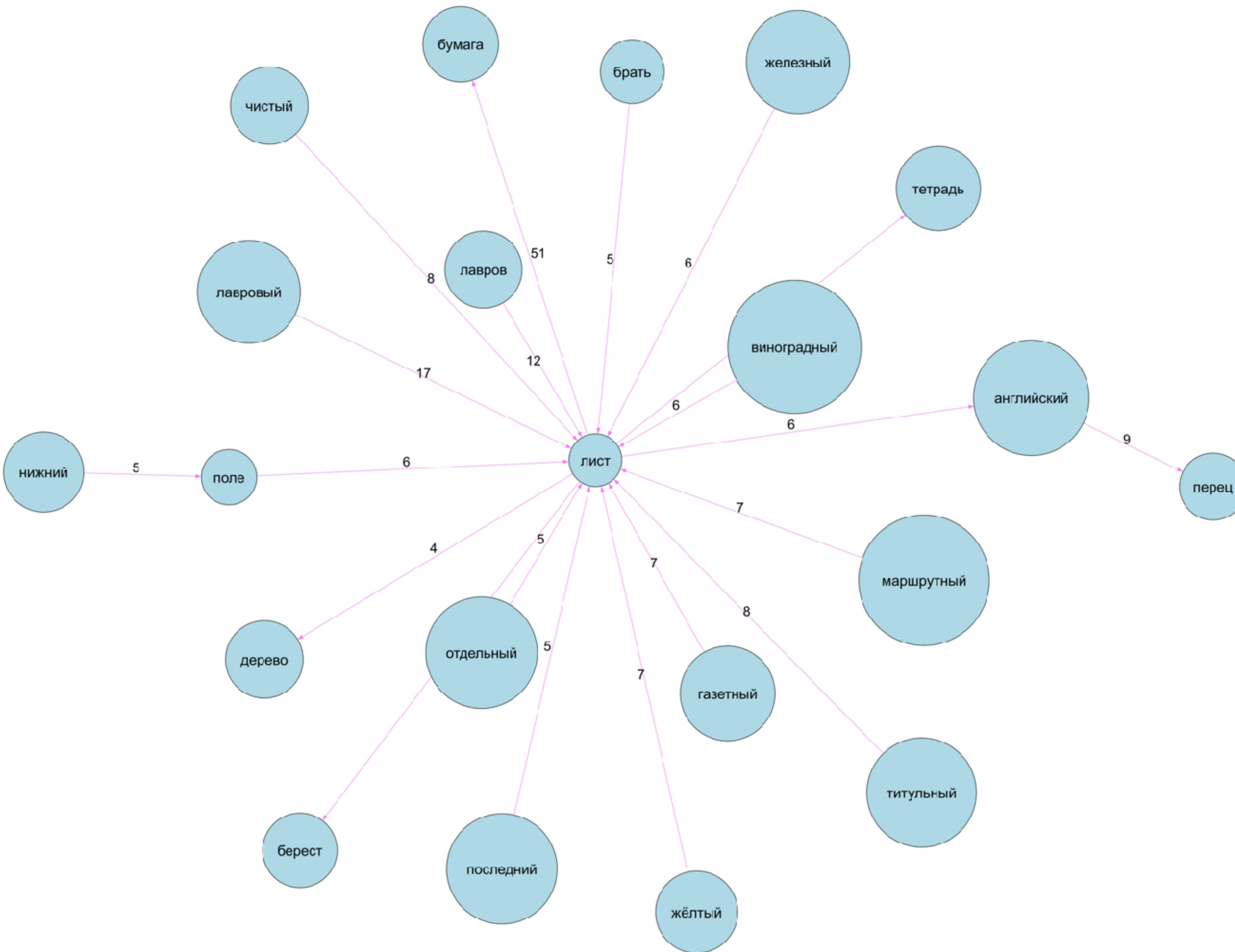
# С ключевым словом.



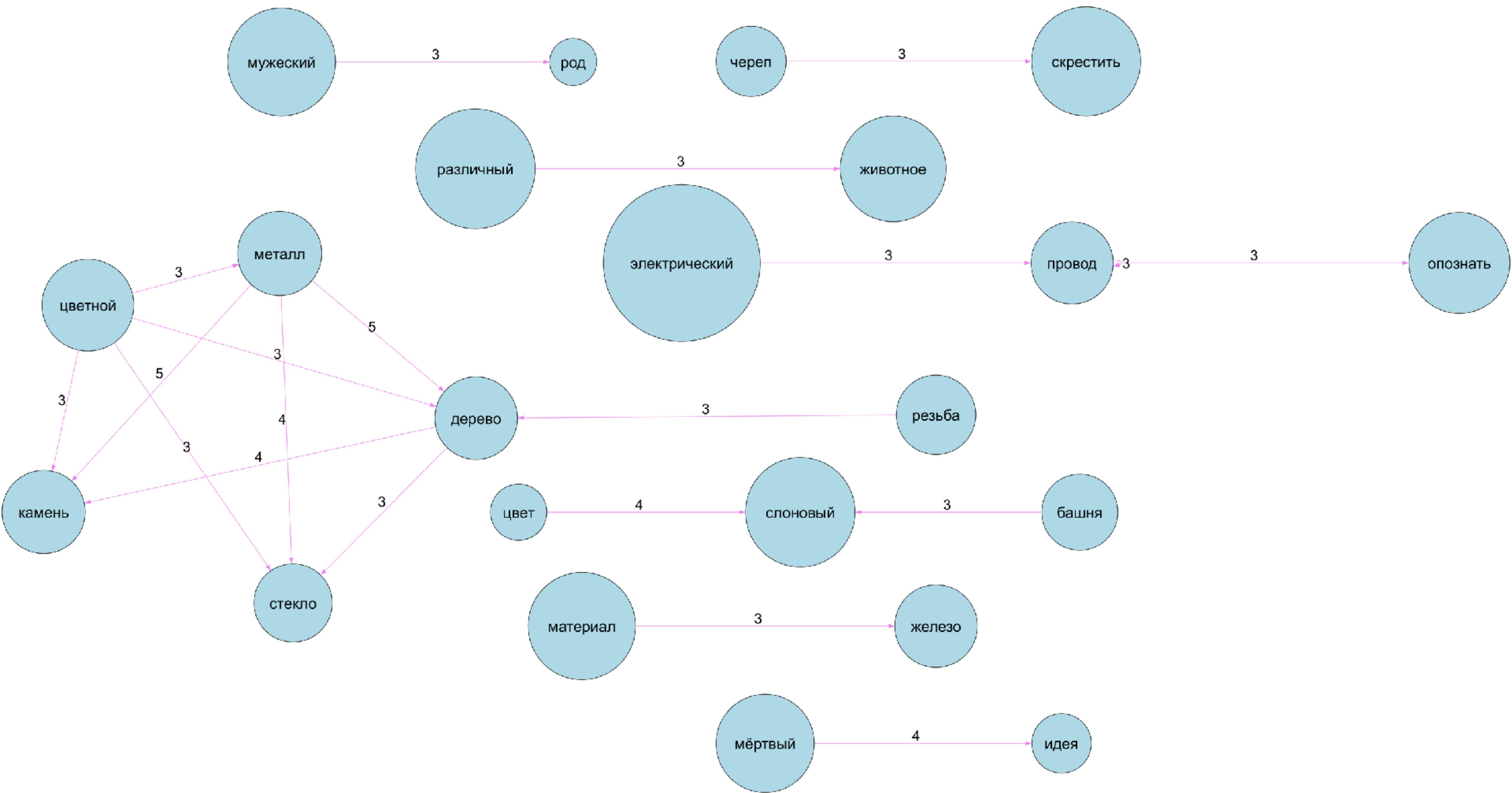
# Без ключевого слова.



# С ключевым словом.



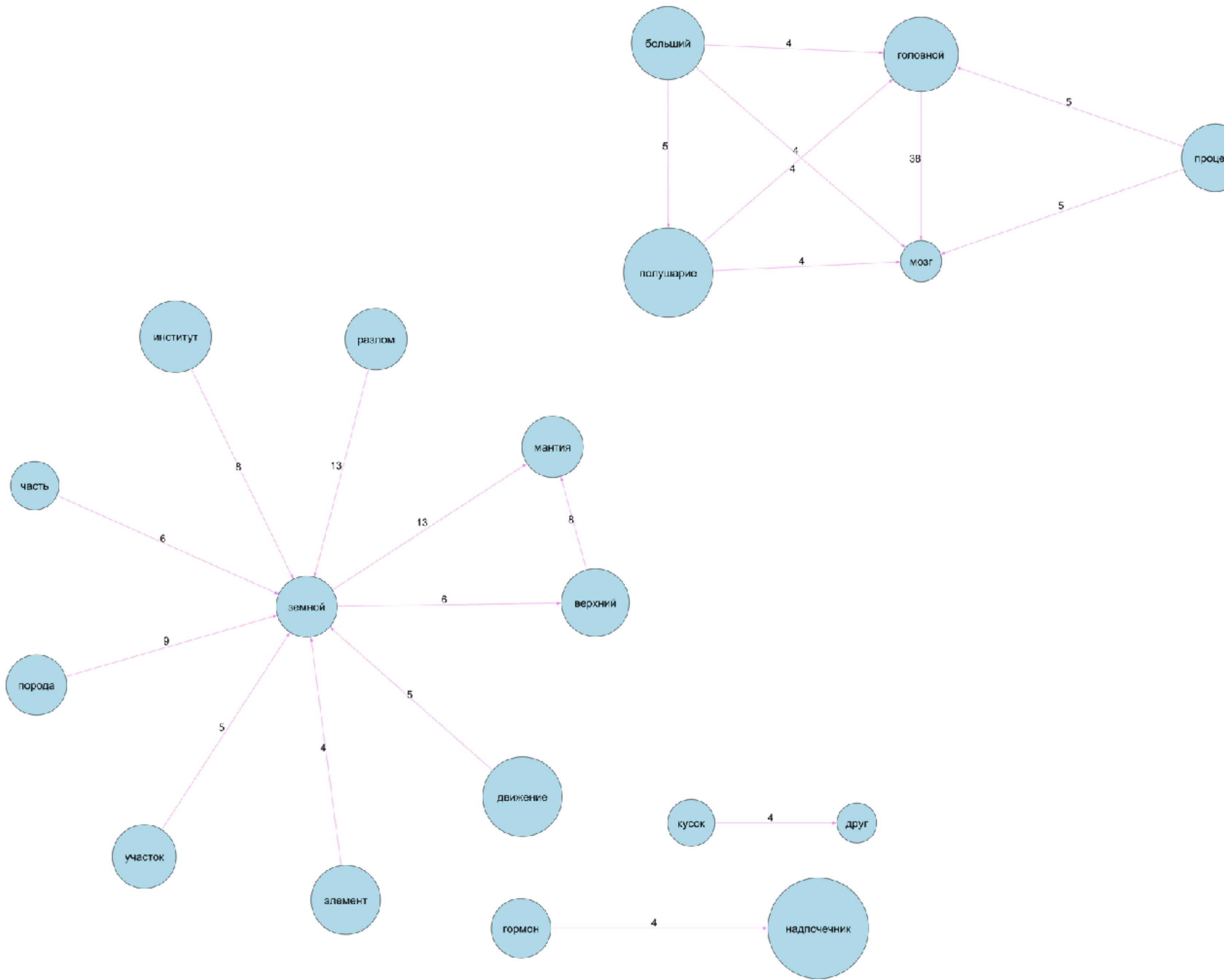
# Без ключевого слова.



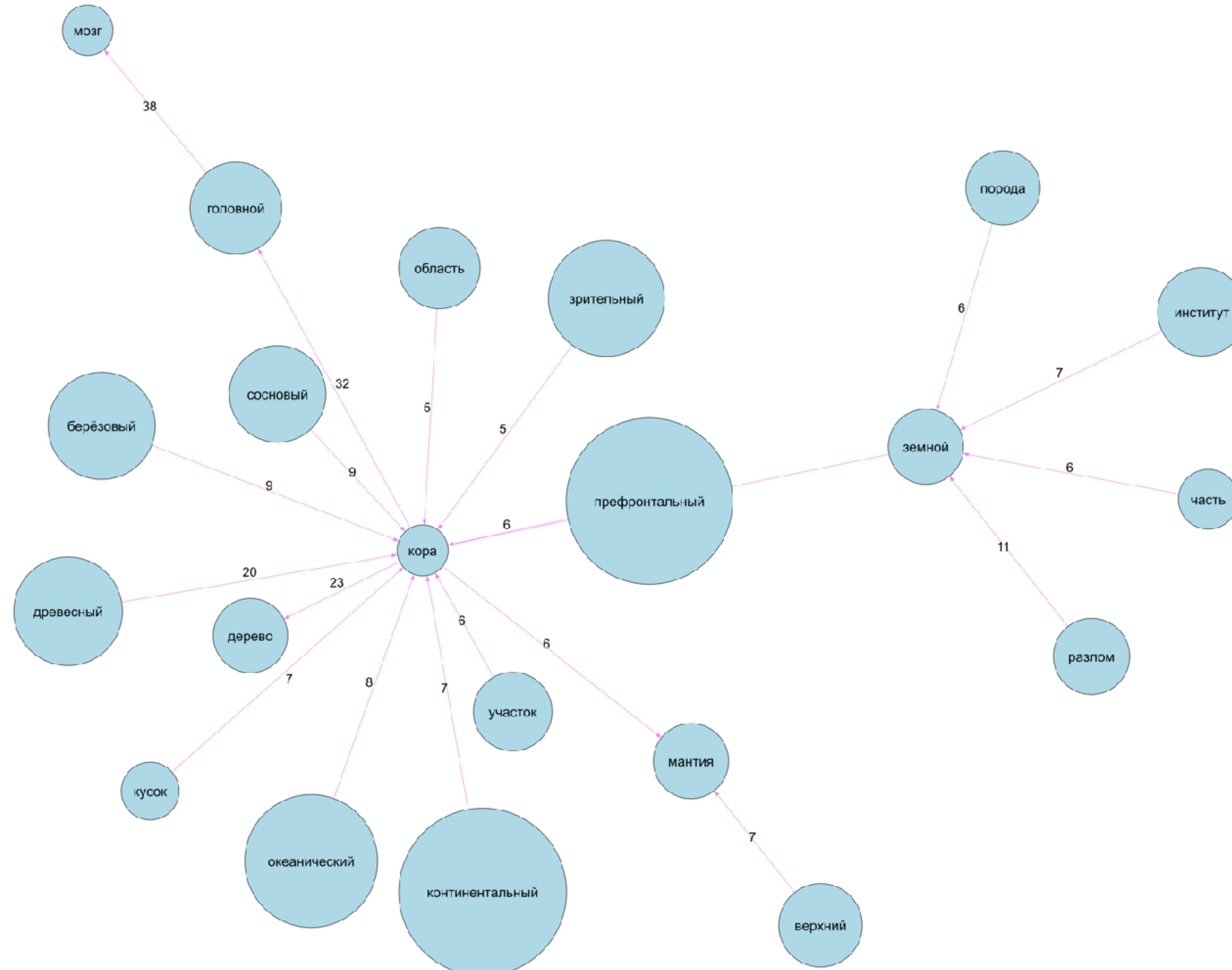
# С ключевым словом.



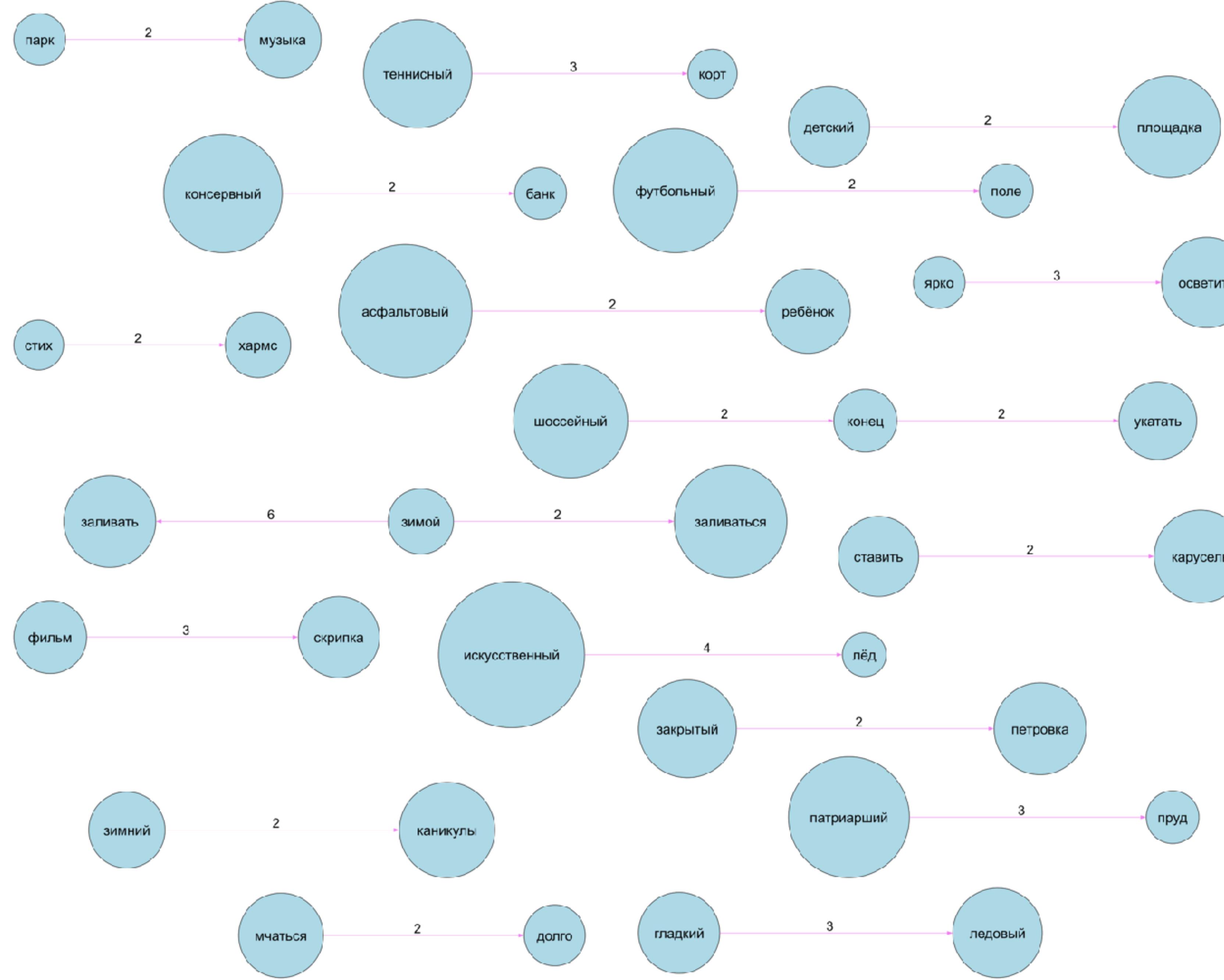
# Без ключевого слова.



# С ключевым словом.



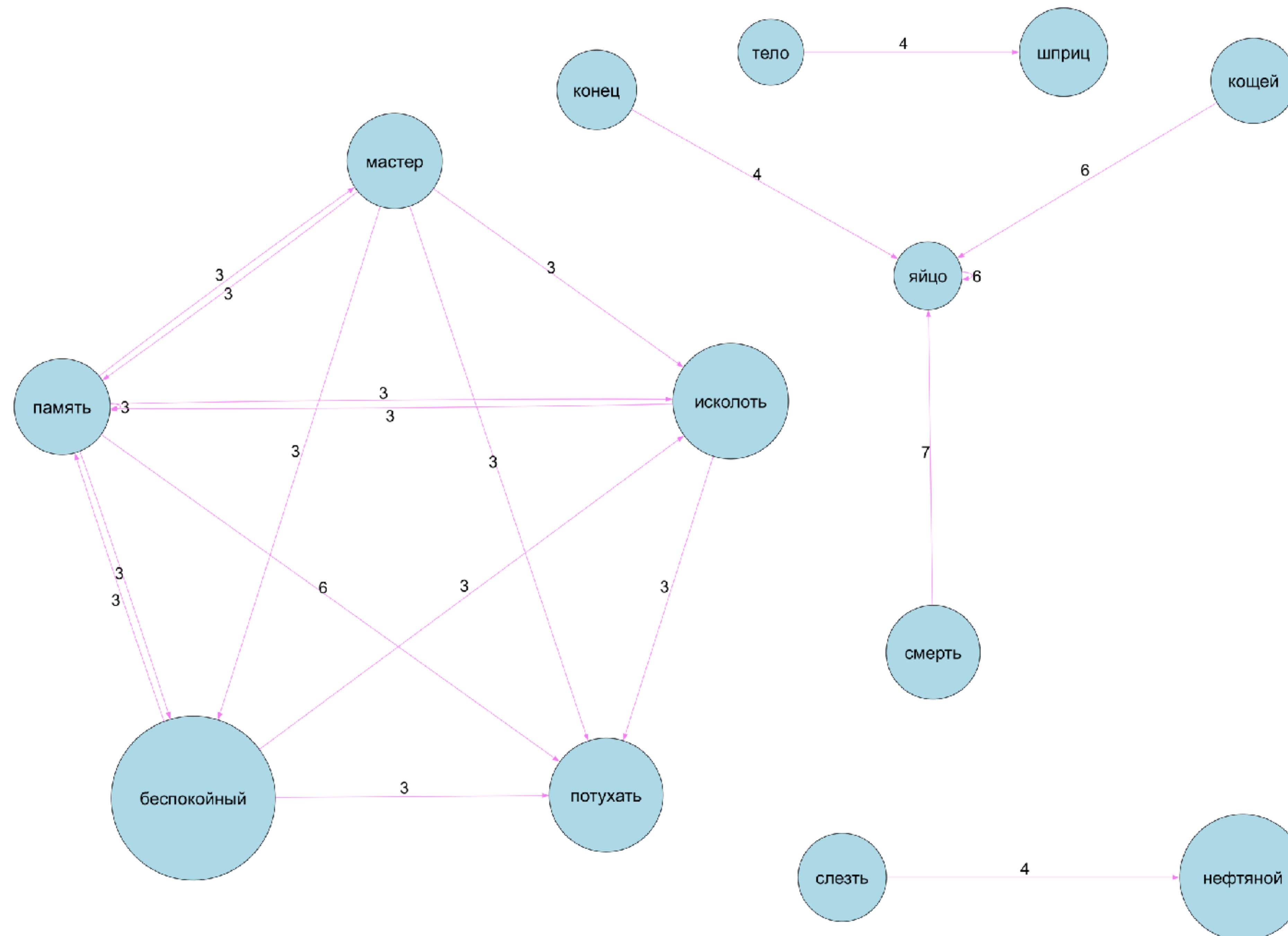
# Без ключевого слова.



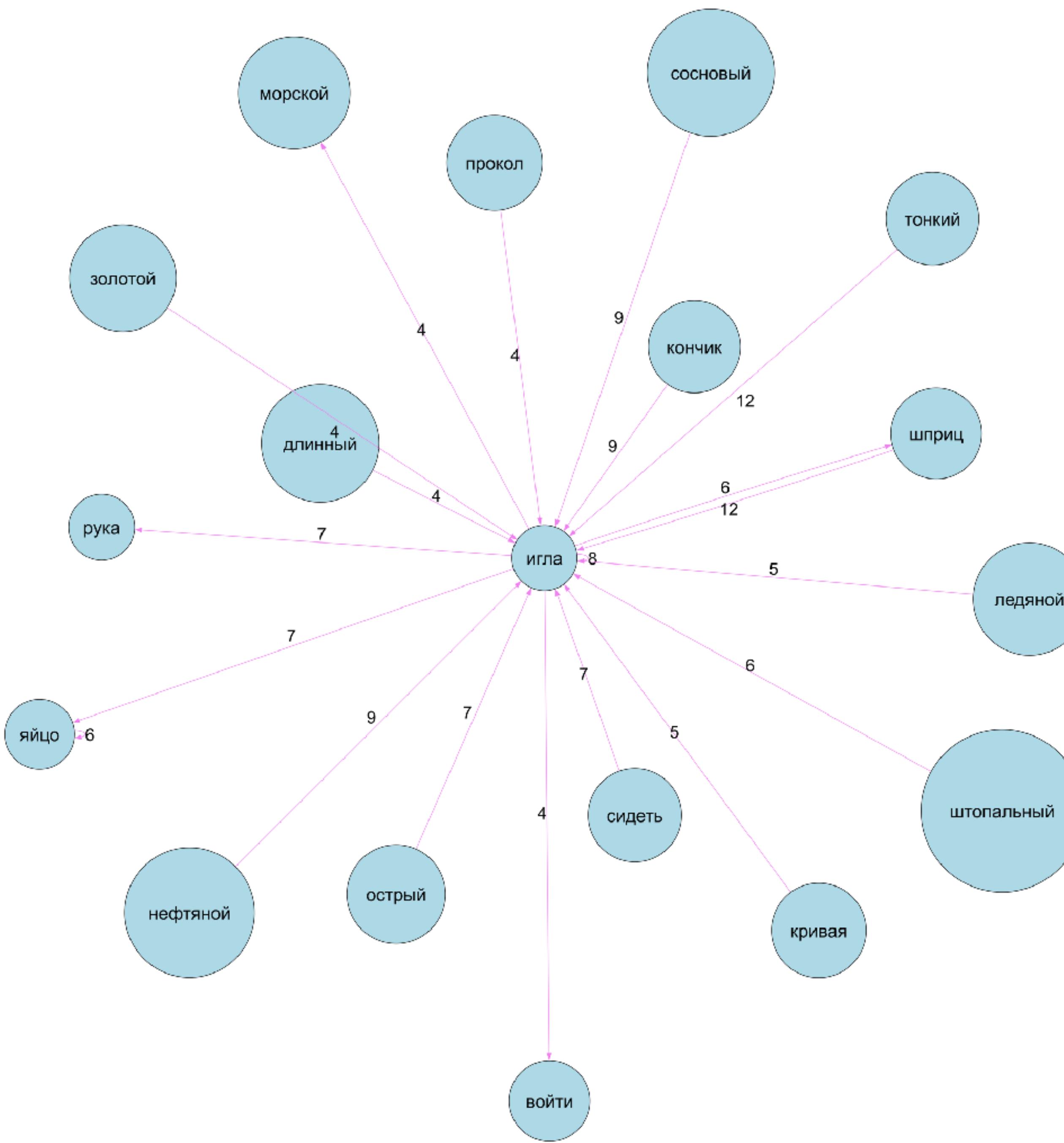
# С ключевым словом.



# Без ключевого слова.

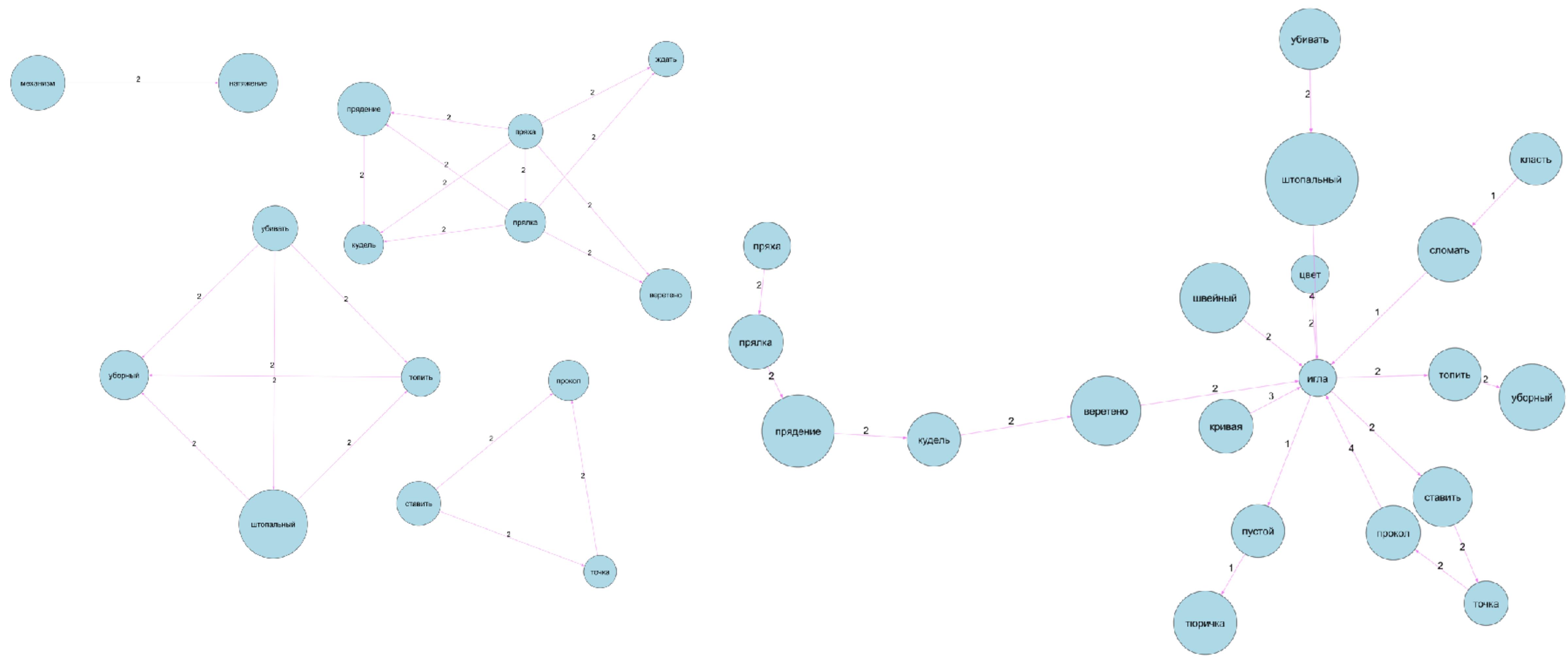


# С ключевым словом.

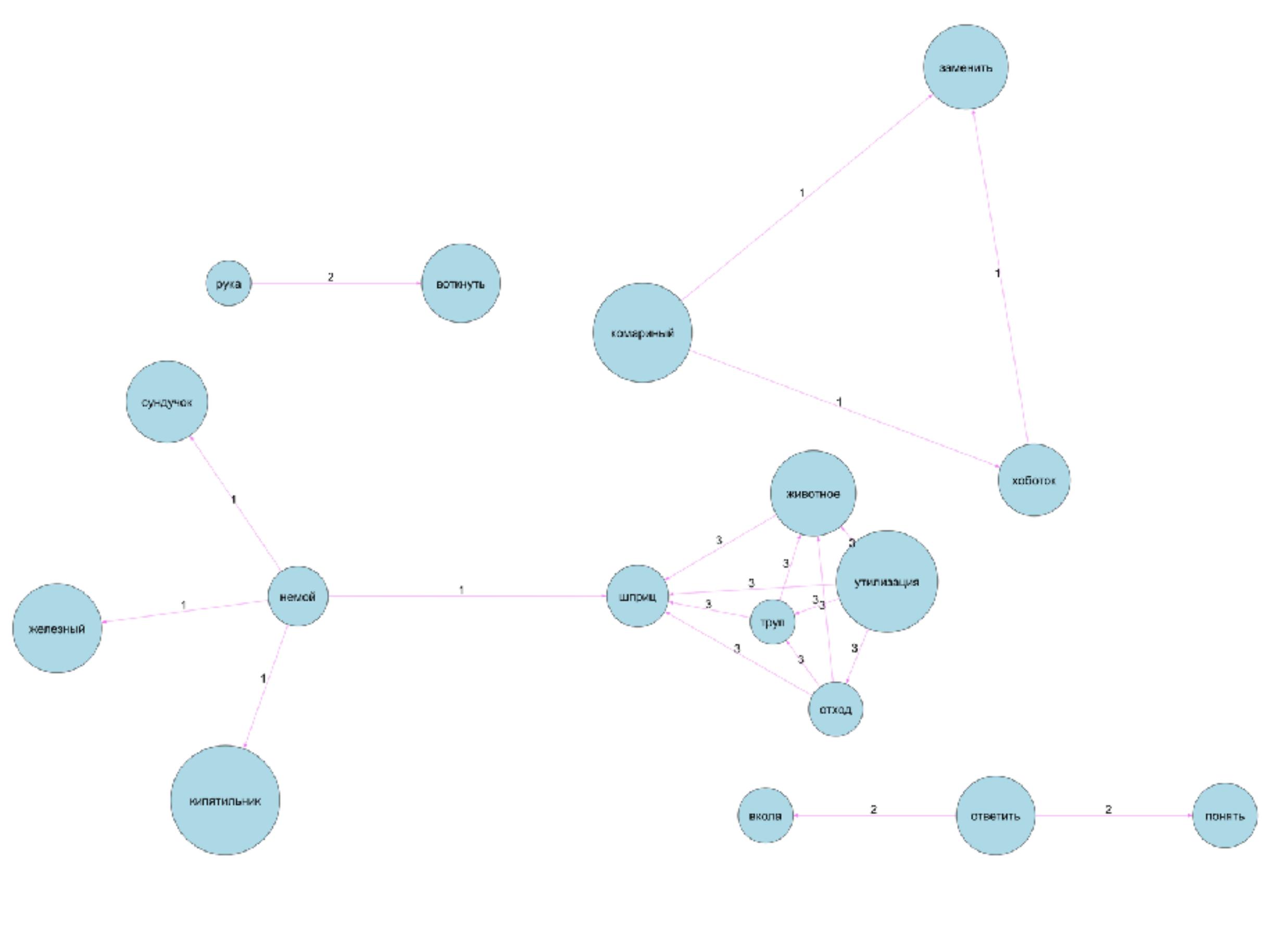


# Анализ размеченных данных, группируя их по значениям слов и создавая визуализации для разных значений одного и того же слова.

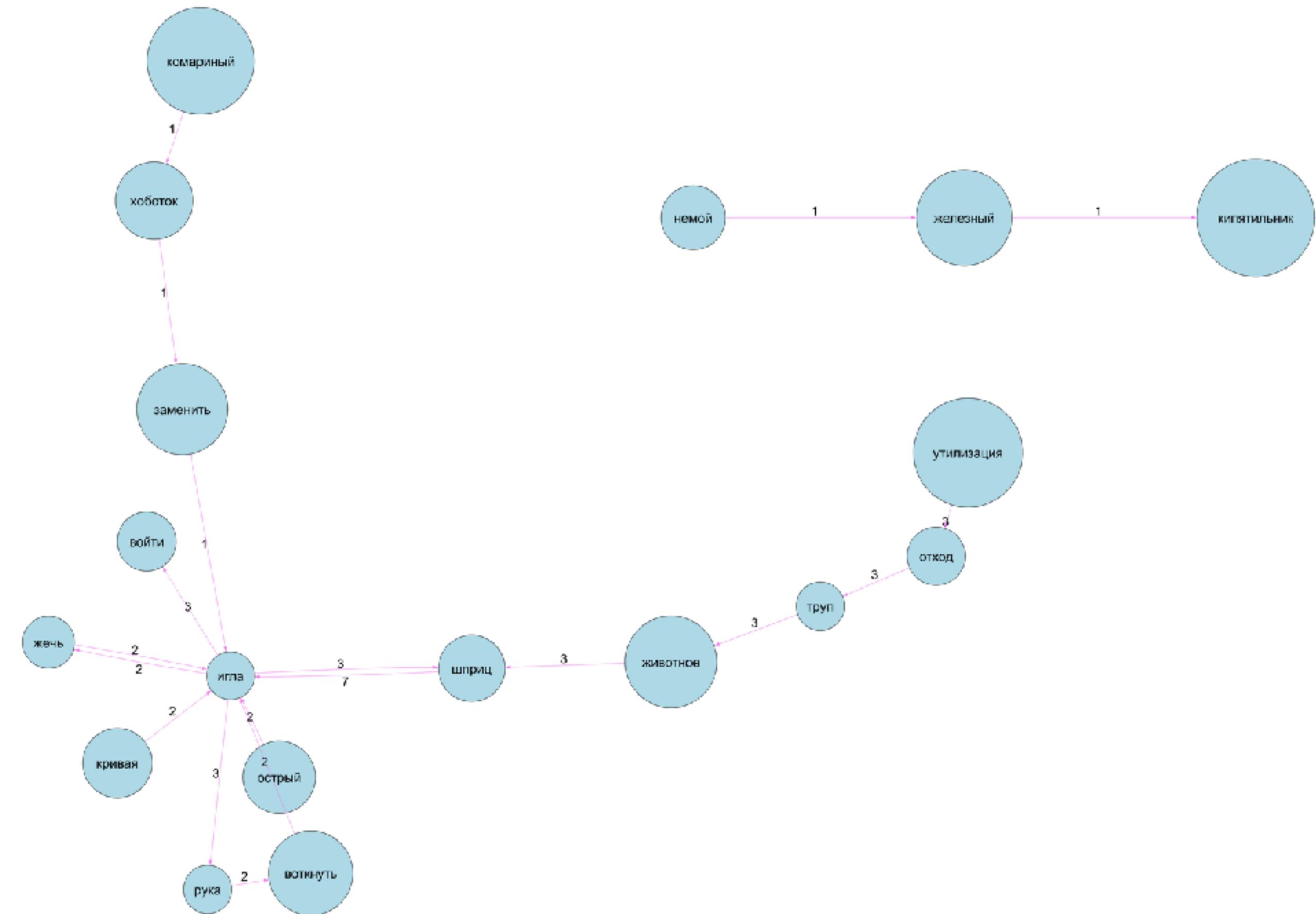
	<b>игла</b>	<b>значение «игла»</b>	<b>пара</b>	<b>значение «пара»</b>	<b>каток</b>	<b>значение «каток»</b>	<b>лист</b>	<b>значение «лист»</b>
1	и клали в них сломанные иглы, пустые тюрички (...)	для шитья	Могу дать пару советов.	общее указание, небольшой набор советов	принято, поэтому не улица, а каток!	метафора, скользкая дорога	выкладываем на лист салата.	часть блюда
2	голове, будто подушечка, утыканная швейными иг...	для шитья	На пары ходить, лекции перечитывать, готовиться к	занятия	Тяжелый ящик, установленный на одинаковых катк...	элемент механического устройства	Перечислить всех просто не хватит листа.	бумажный
3	этих несчастных заразилось через героиновую иглу.	наркотик	А еще ходить на пары все, и если можно понять	занятия	В воскресенье на олимпийском катке в Хамаре (Н...	ледовая площадка	Сверху положить листья эстрагона и зеленый лук.	часть блюда



# Игла для шитья. Без ключевого слова.

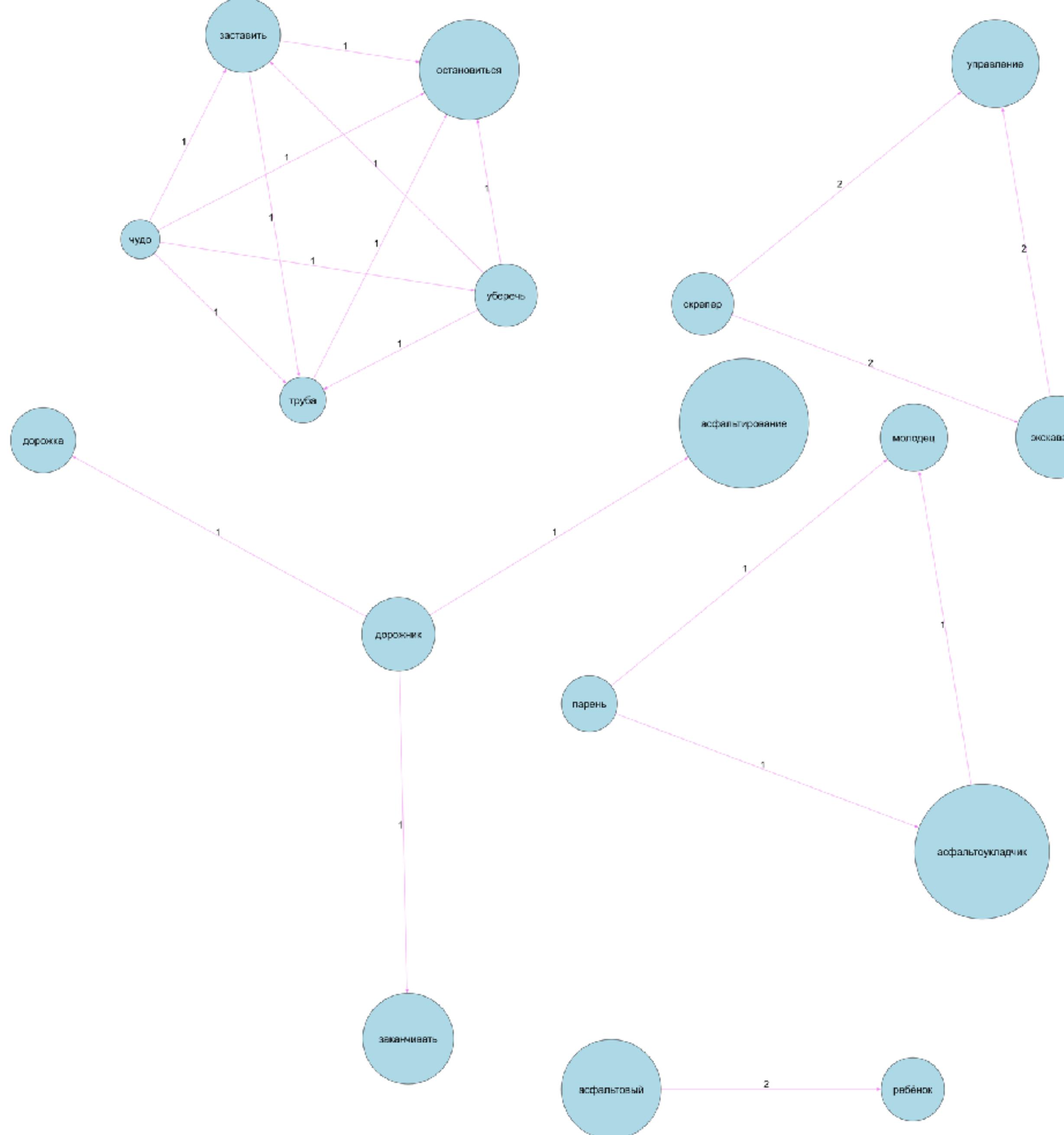


Игла медицинская.  
Без ключевого слова.

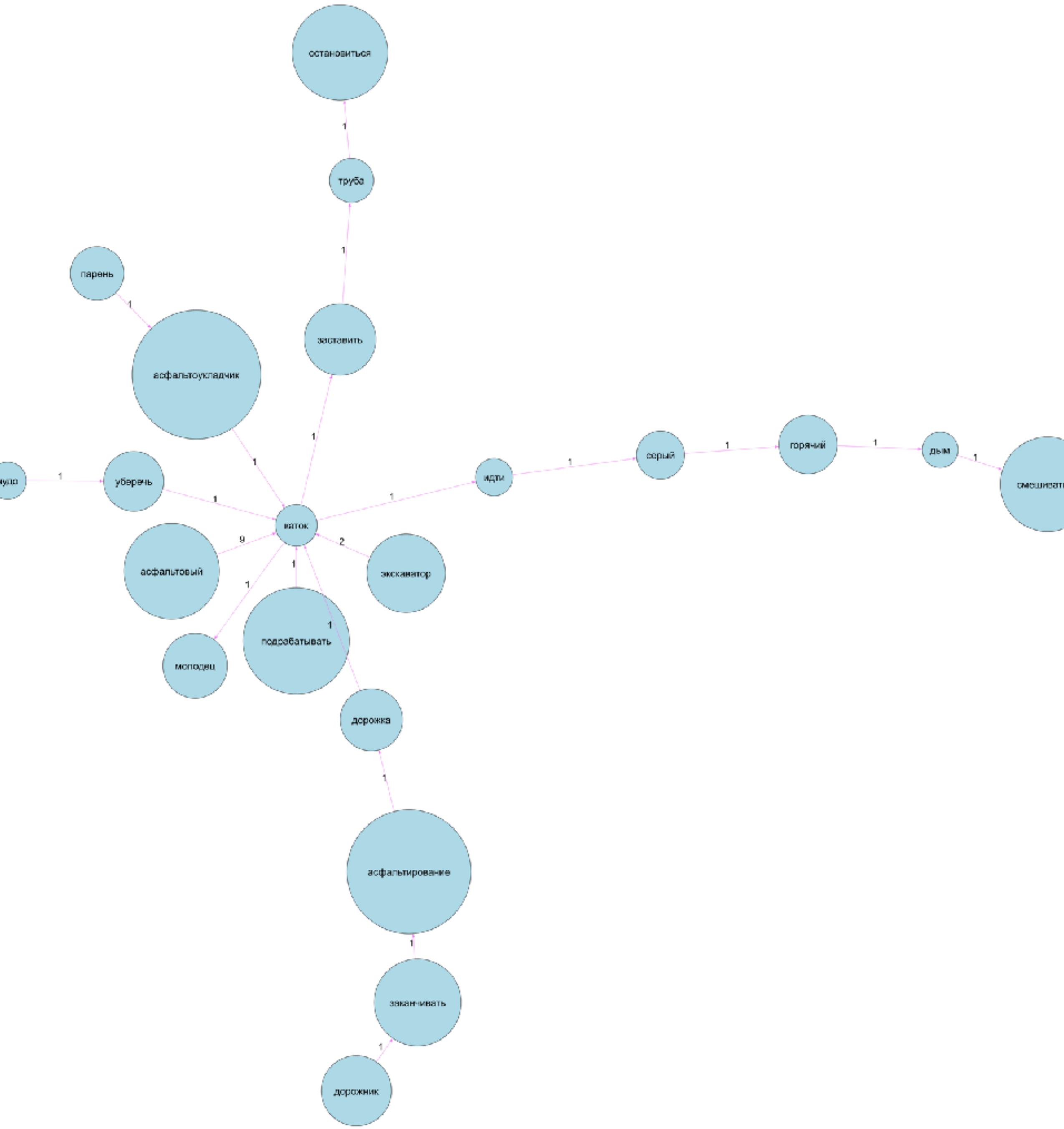


Игла медицинская.  
С ключевым словом.

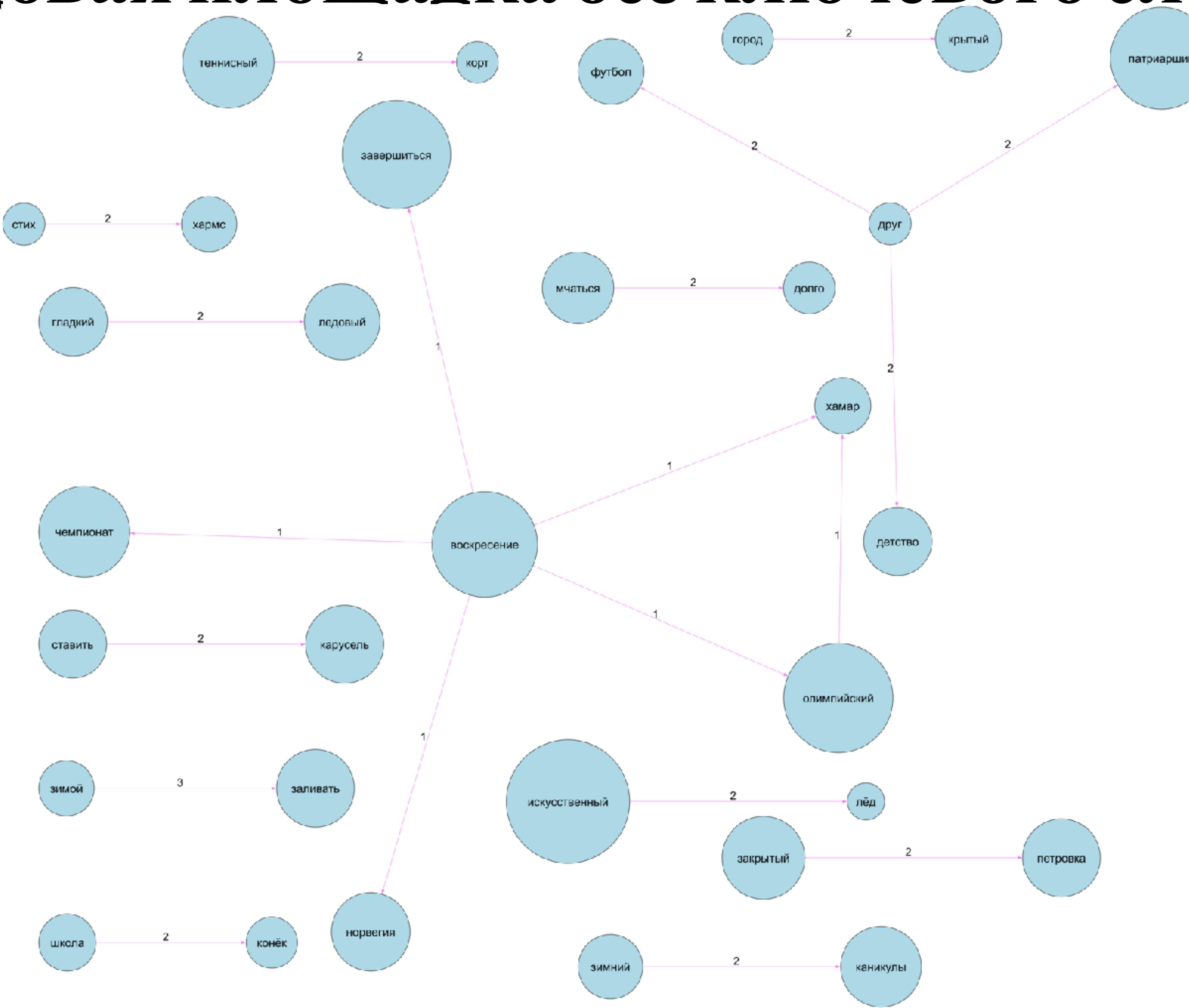
# Дорожный каток без ключевого слова.



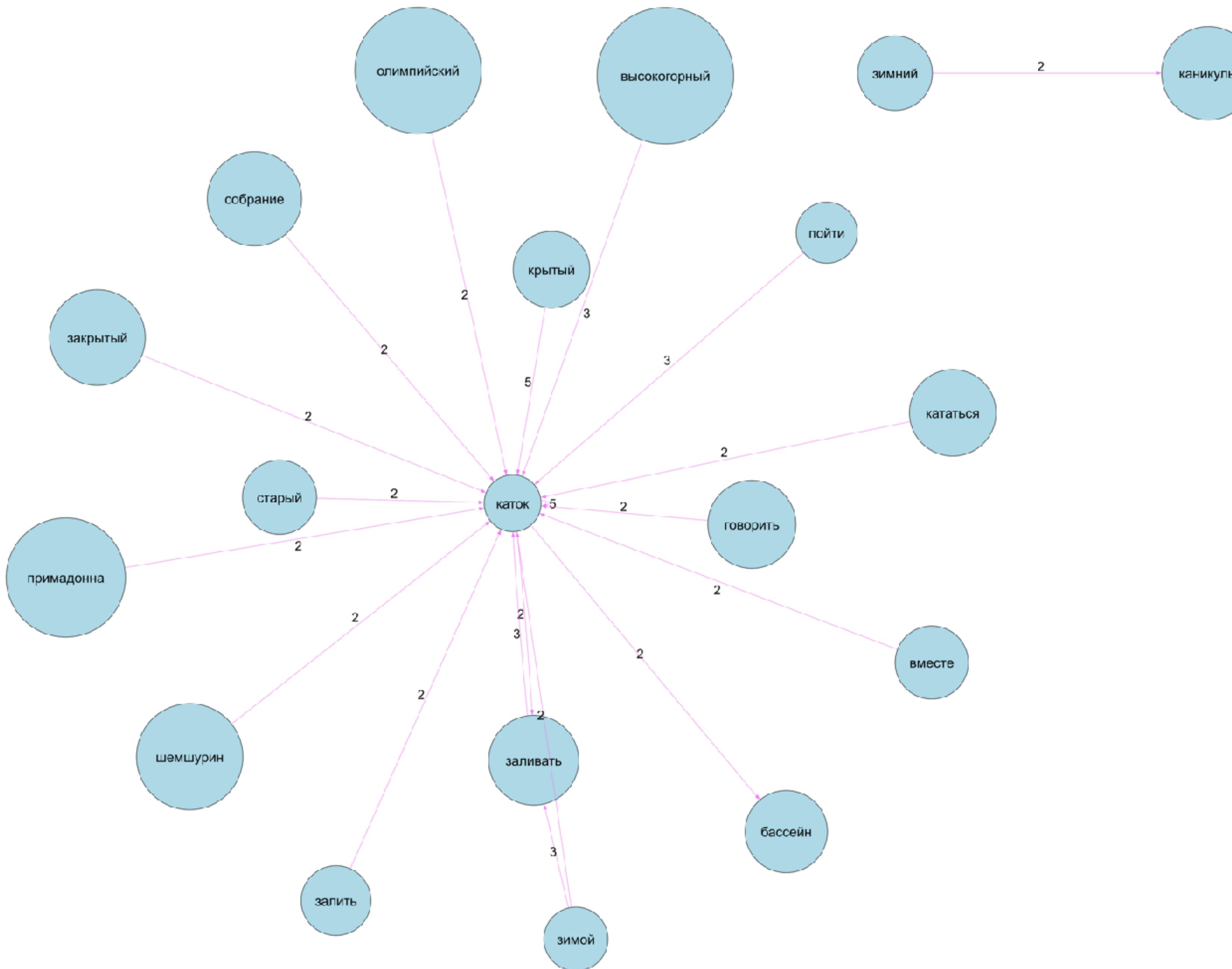
# Дорожный каток с ключевым словом.



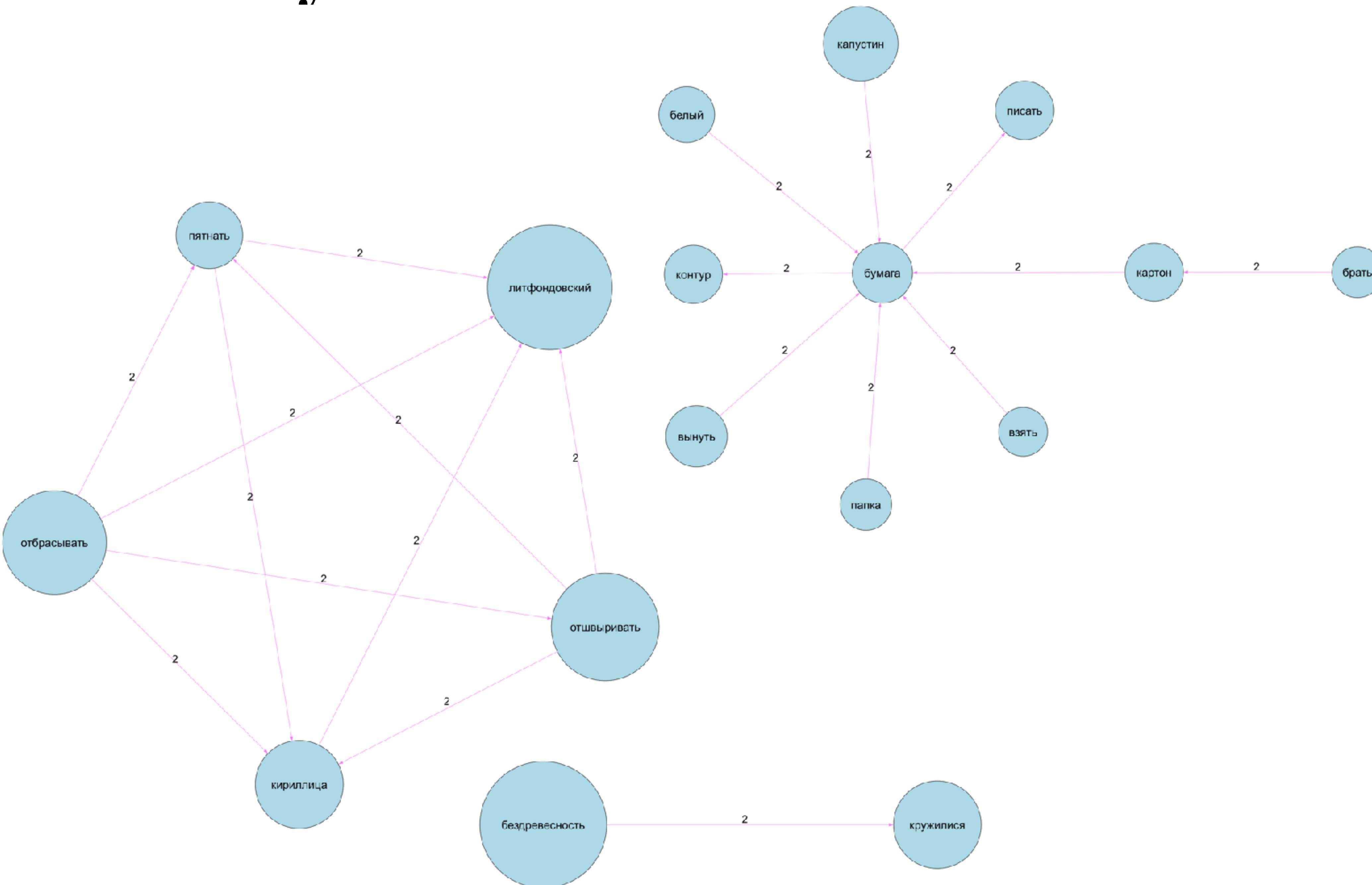
# Ледовая площадка без ключевого слова.



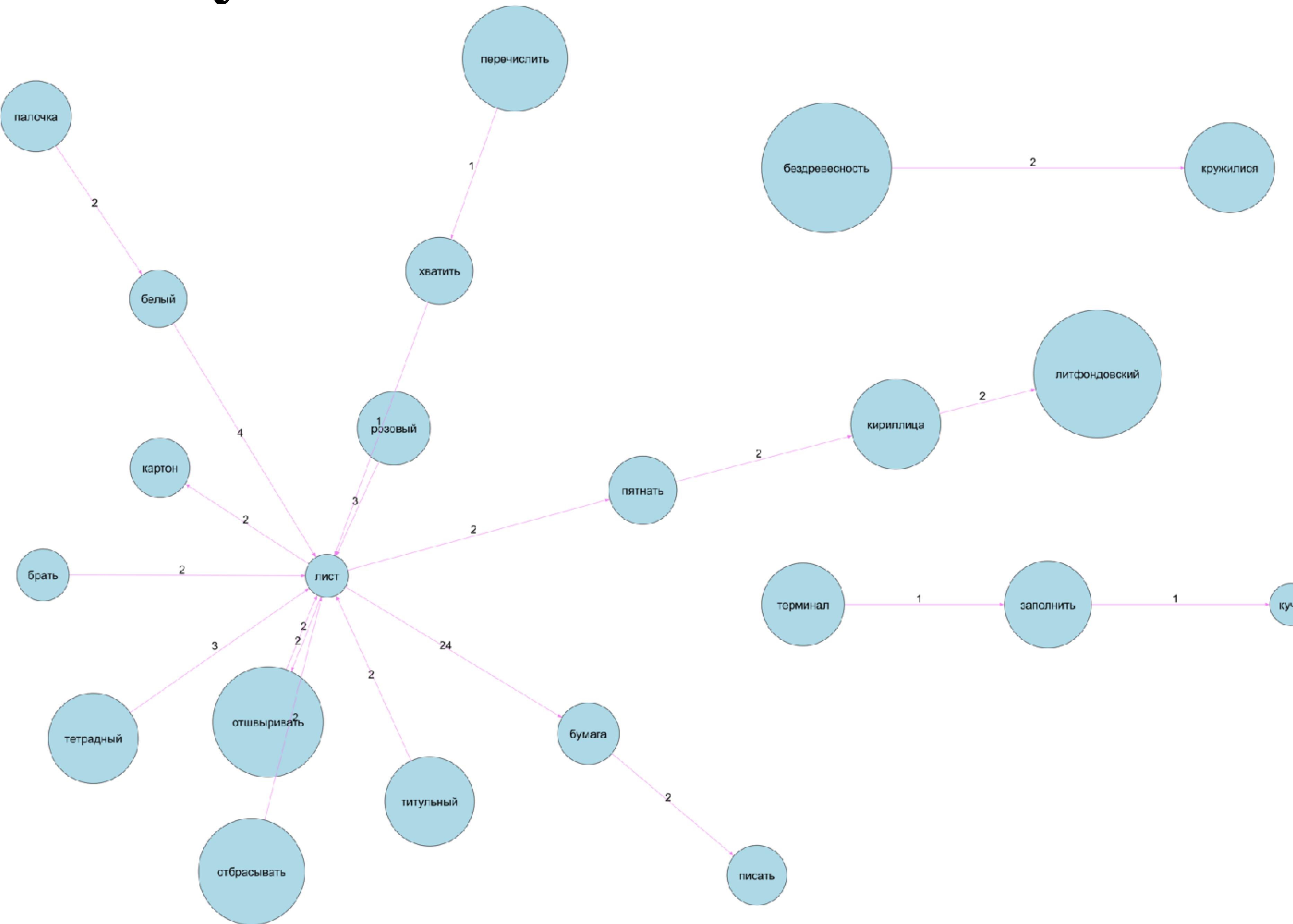
# Ледовая площадка с ключевым словом.



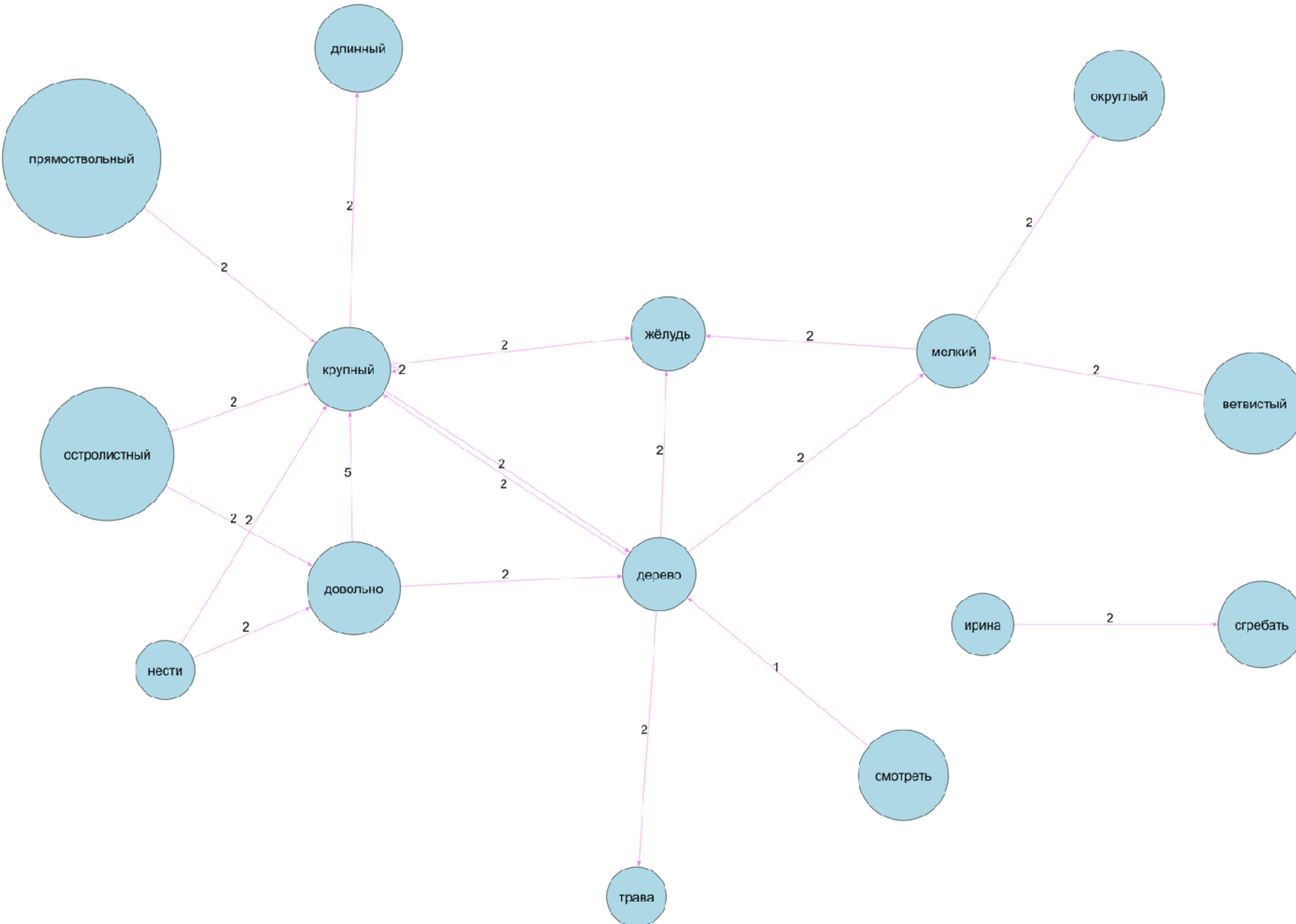
# Лист бумажный без ключевого слова.



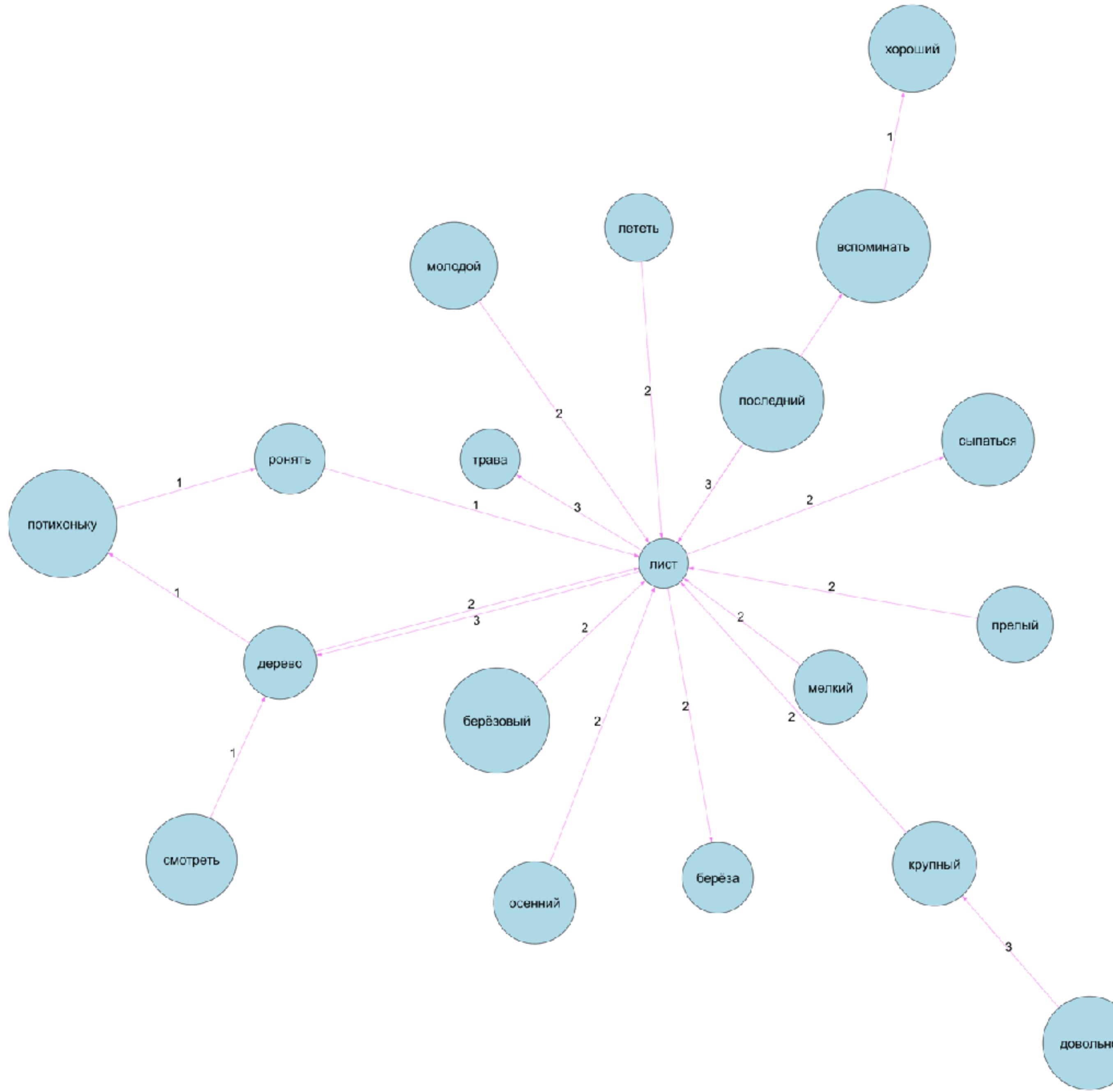
# Лист бумажный с ключевым словом.



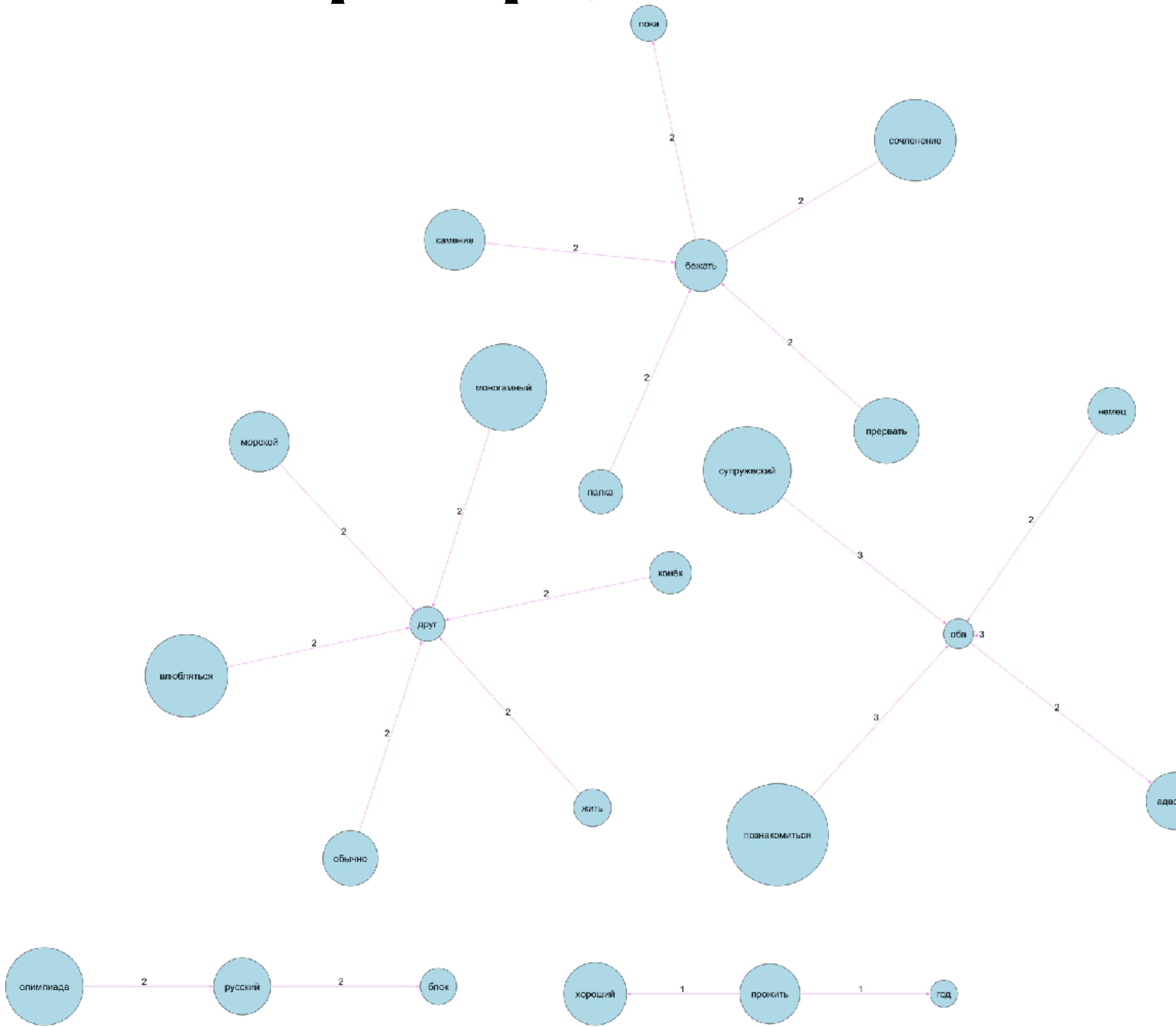
# Лист дерево без ключевого слова.



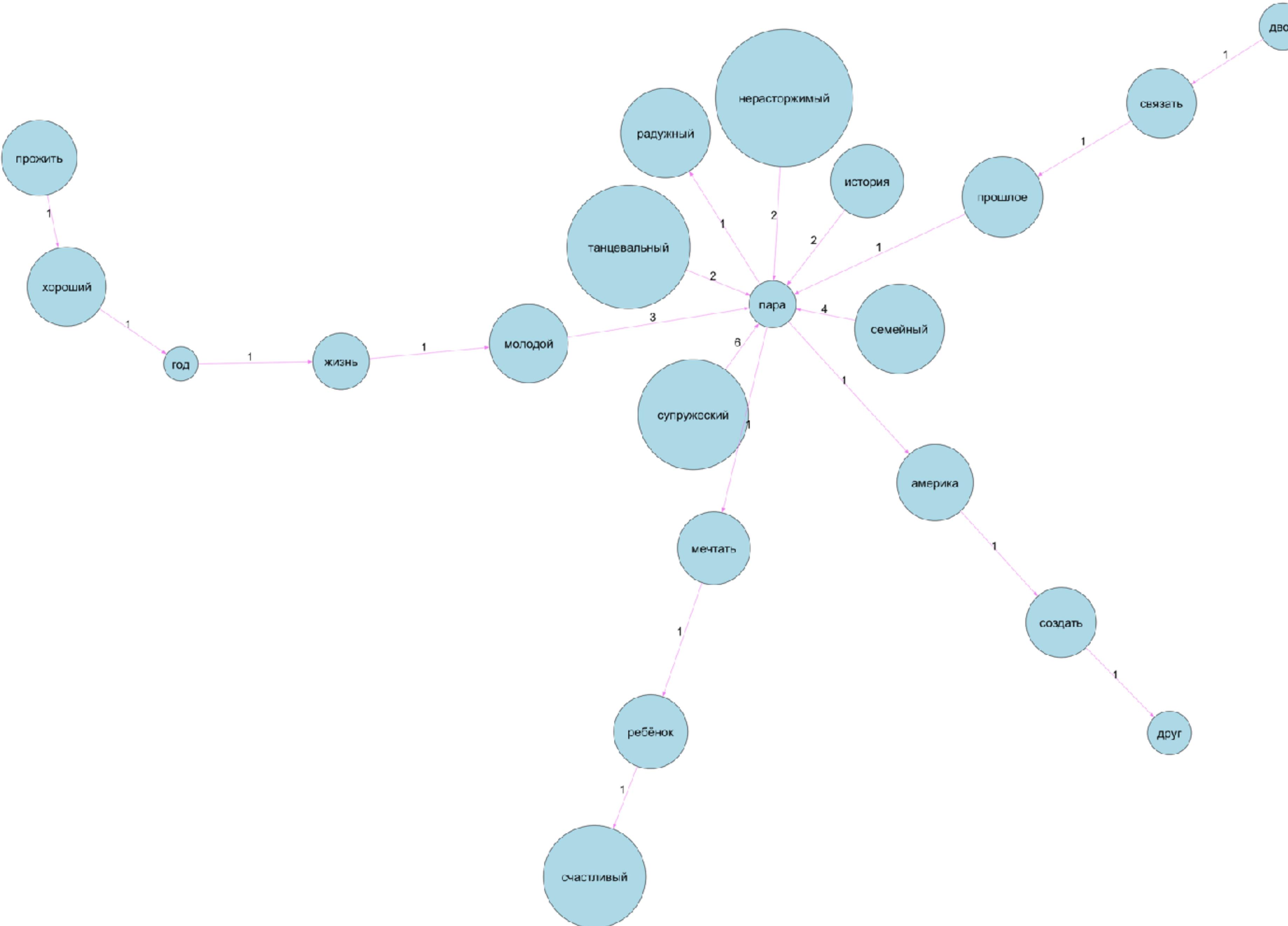
# Лист дерево с ключевым словом.



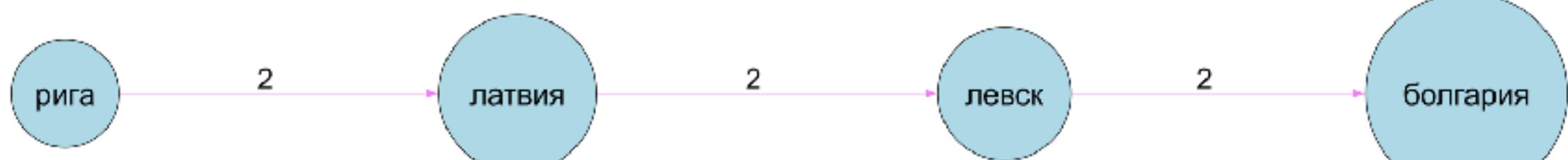
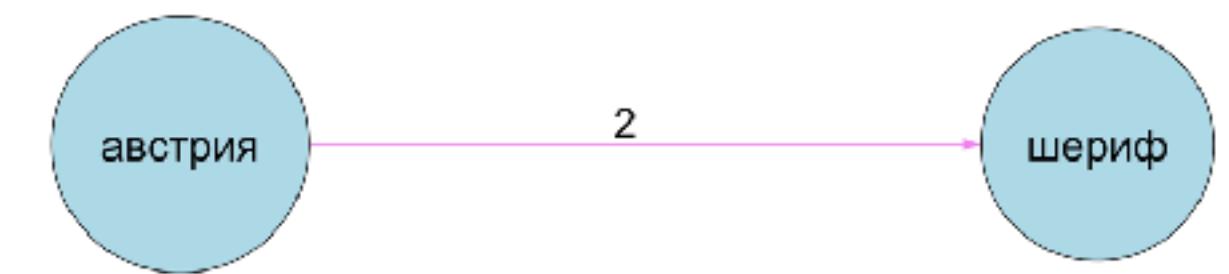
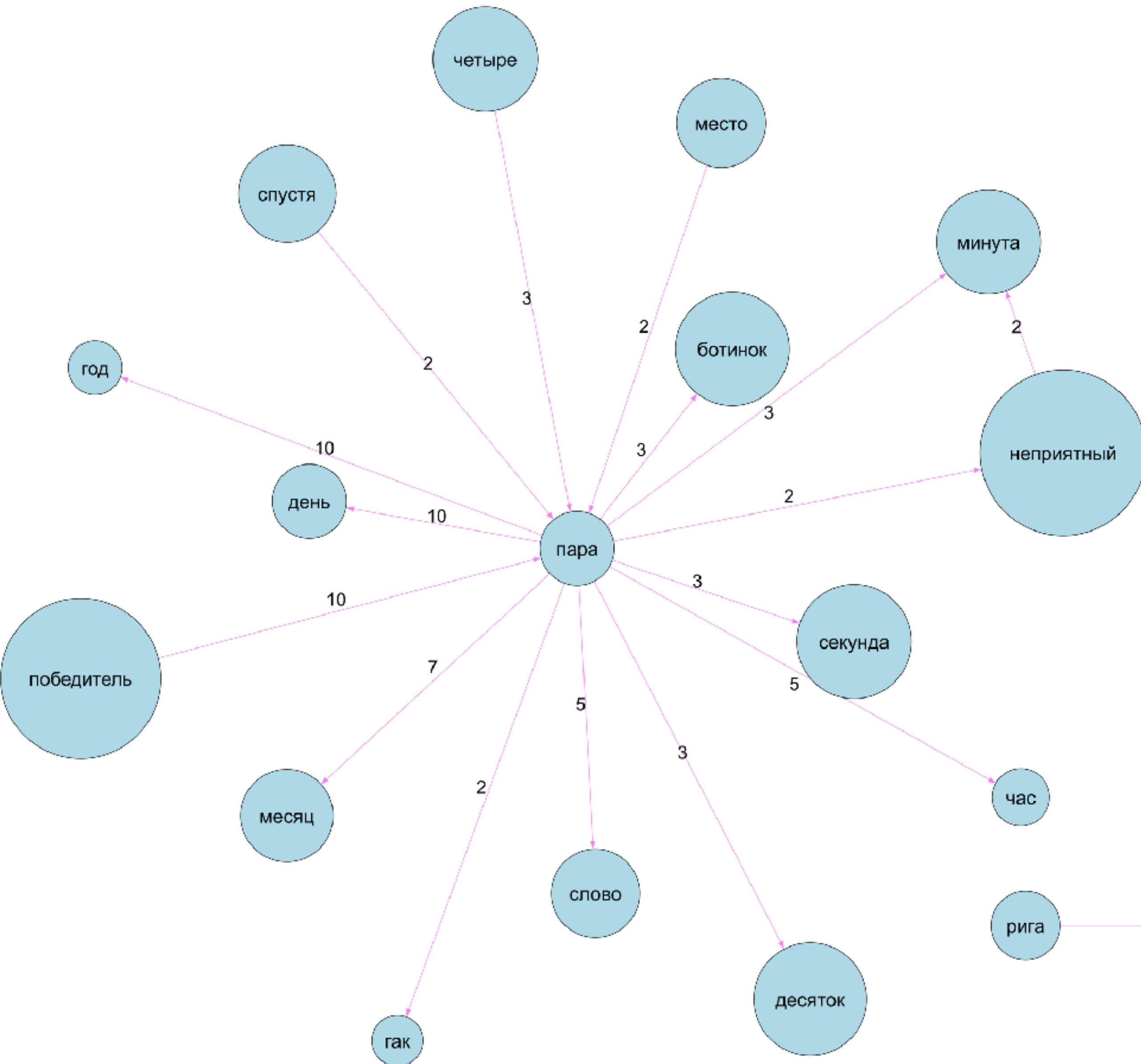
# Пара - два партнера, без ключевого слова.



# Пара - два партнера, с ключевым словом.



# Пара - несколько, количество. С ключевым словом.



**Спасибо за внимание!**