**Уровень 1**

**1 файл**

**Визуализация данных**

**Задание 1**: законспектировать основные рекомендации по созданию визуализации

1. Определить цель визуализации: перед началом работы необходимо четко определить, какую информацию вы хотите передать через визуализацию и какую именно аудиторию вы планируете привлечь.

2. Выбрать подходящий тип визуализации: в зависимости от цели и типа данных, выберите наиболее подходящий тип визуализации, например, диаграммы, графики, тепловые карты и т.д.

3. Обеспечить четкость и понятность: визуализация должна быть легко читаемой и понятной для аудитории. Используйте понятные метки, легенды, цветовую схему и другие элементы для улучшения понимания данных.

4. Избегать избыточной информации: не перегружайте визуализацию лишней информацией, это может затруднить ее восприятие. Сосредоточьтесь на ключевых данных, которые необходимо передать.

5. Проверить точность данных: перед публикацией визуализации убедитесь, что данные корректны и точны. Неправильные данные могут привести к неверному выводу.

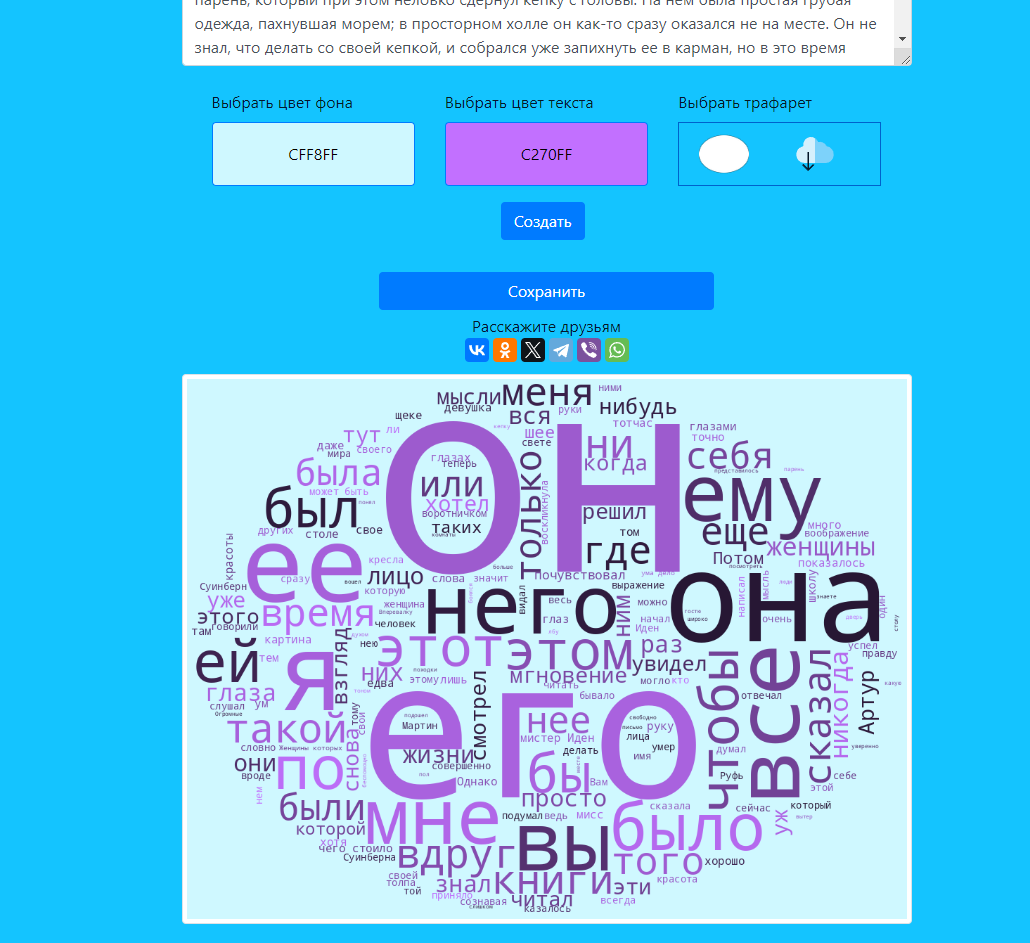
6. Протестировать визуализацию: перед окончательным представлением визуализации протестируйте ее на различных аудиториях, чтобы убедиться, что она эффективно передает нужную информацию.

7. Использовать специализированные инструменты: для создания профессиональных визуализаций в компьютерной лингвистике используйте специализированные инструменты и программное обеспечение, такие как Tableau, Python с библиотекой Matplotlib или R с пакетом ggplot2.

**Задание 2**: воспользоваться указанными программами для создания собственных облаков слов. Найти как минимум 3 аналогичных ресурса.

Word Clouds.

Первая глава произведения Джека Лондона «Мартин Иден»



Word it out

Статья про скатов



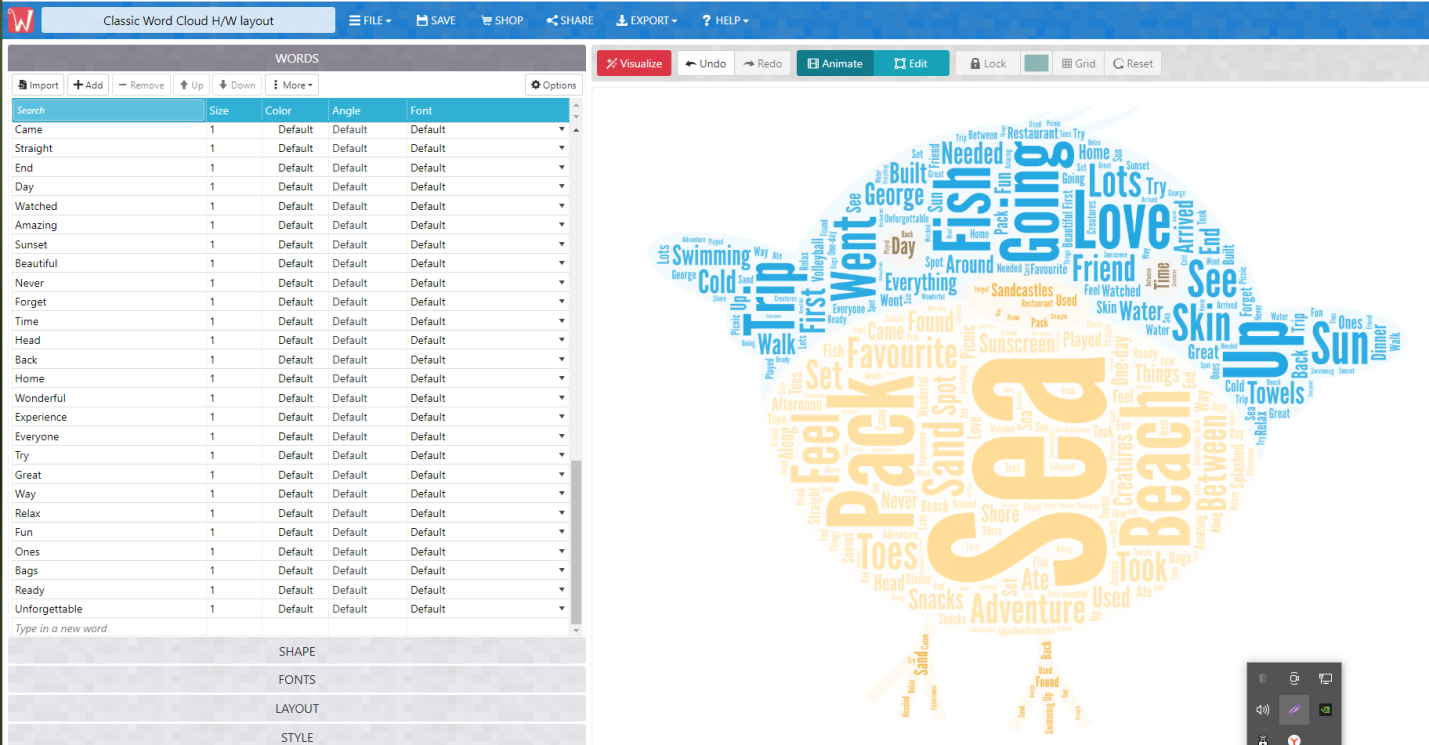
Word Cloud by Jason Davies

Текст про Гавайские острова



World art

Текст про путешествие по морю



**TagCrowd** (tagcrowd.com): TagCrowd - это онлайн-инструмент для создания облаков слов, который позволяет быстро и легко визуализировать ключевые слова из текста. Пользователи могут настроить шрифт, цвета и другие параметры облака слов.

**WordClouds** (wordclouds.com): WordClouds предоставляет возможность создания красочных и креативных облаков слов из текста или URL. Инструмент также позволяет настраивать форму облака слов, цвета и размер шрифта.

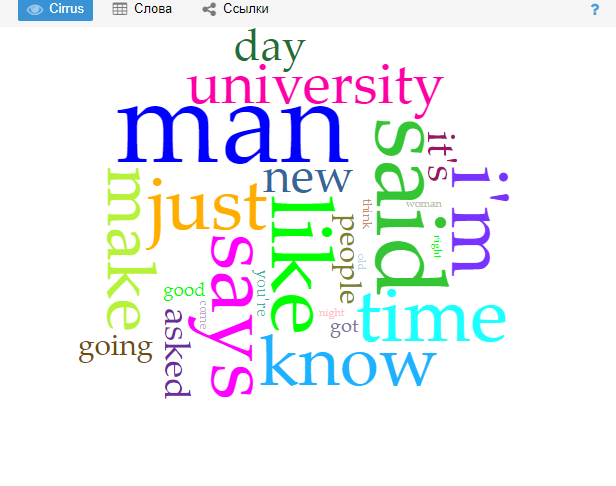
**VocabGrabber** (visualthesaurus.com/vocabgrabber): VocabGrabber - это онлайн-инструмент, который помогает анализировать текст и создавать облака слов на основе ключевых терминов. Пользователи могут увидеть частоту использования слов и их связи в тексте.

Задание 3

3.1.До того, как убрали стоп-слова



После того, как убрали стоп-слова



3.2

1993 и 1996 годы



3.3

Больше всего словоупотреблений в1989 году. Этот файл самый большой. Самый маленький файл – 1997 год.



3.4

Лексическая плотность меньше всего в 1988-1989(19%), 1991(21%) годах



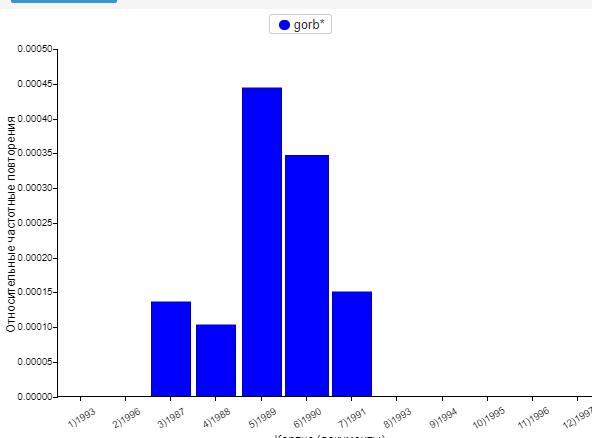
3.5

Значимые слова





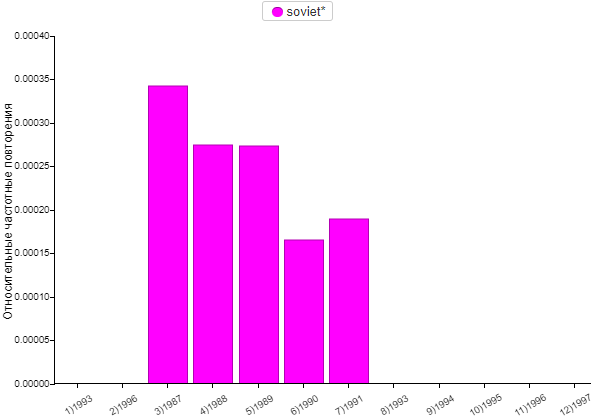
3.6



Постепенно прекратило употребляться

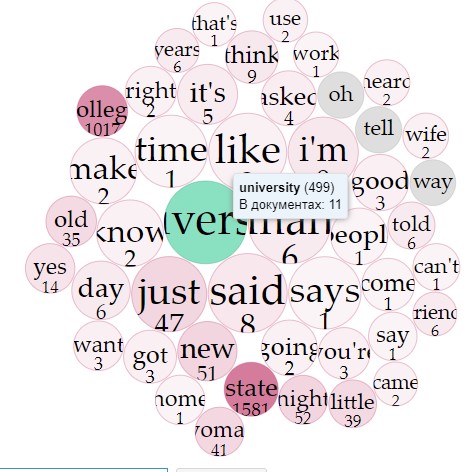


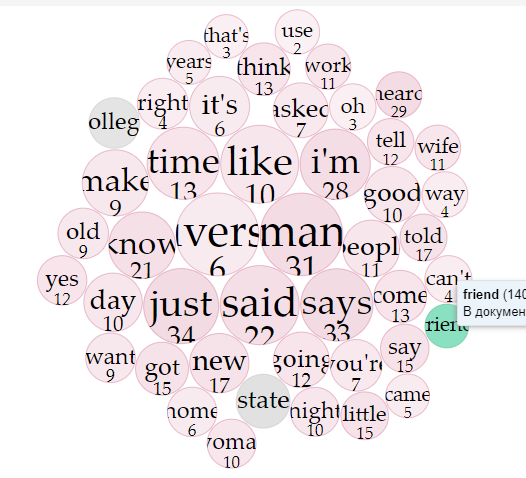
В 1990 употреблялось больше всего



Активно употреблялось несколько лет подряд.

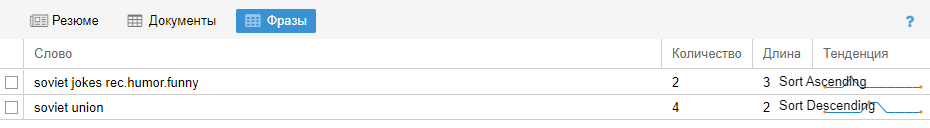
3.7







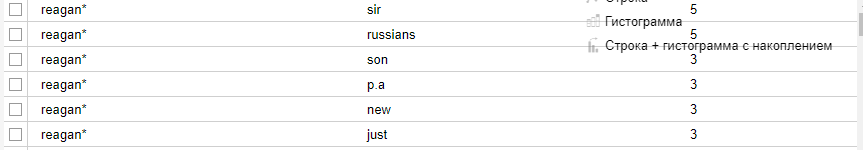
3.8



3.9

Walks into the bar чаще всего употребляется с man

3.10

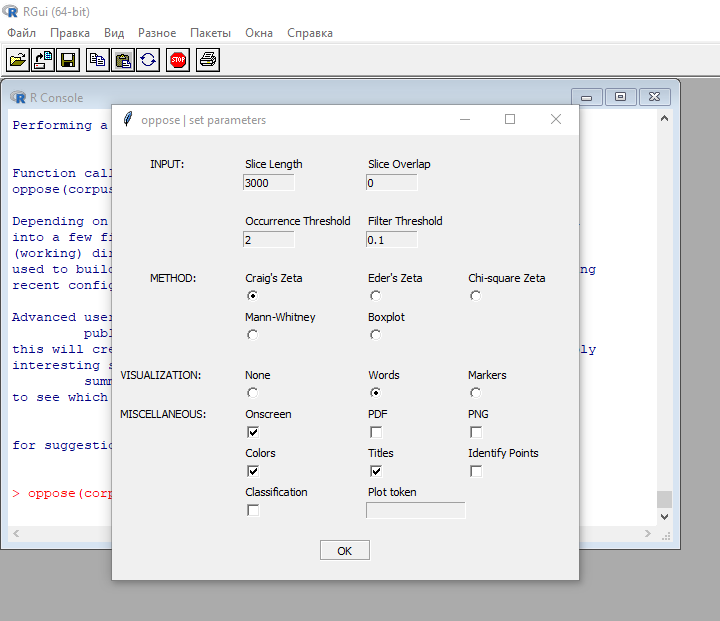


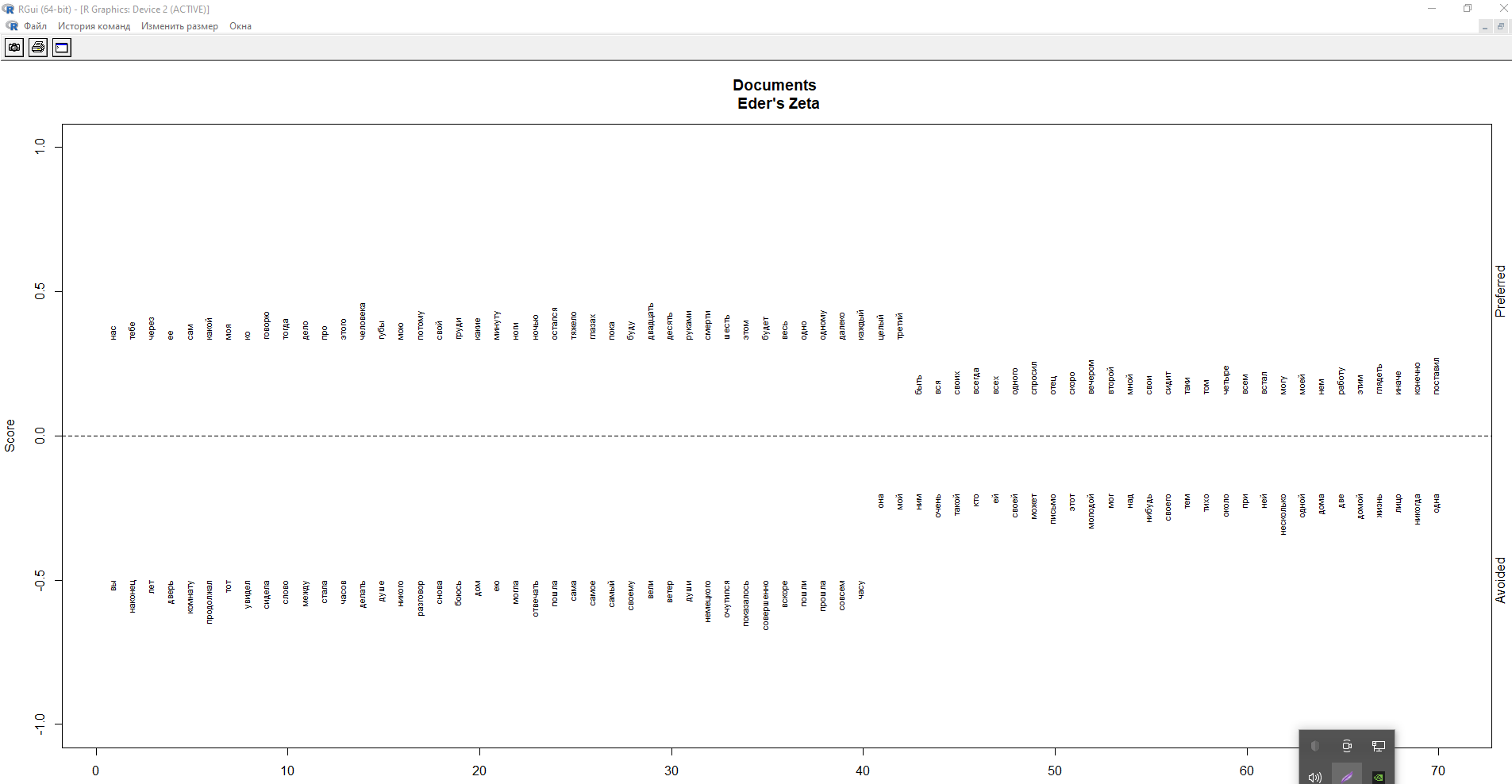
Стилеметрия и тематическое моделирование



Как мы видим, стили некоторых авторов похожи (Шолохов и Распутин).

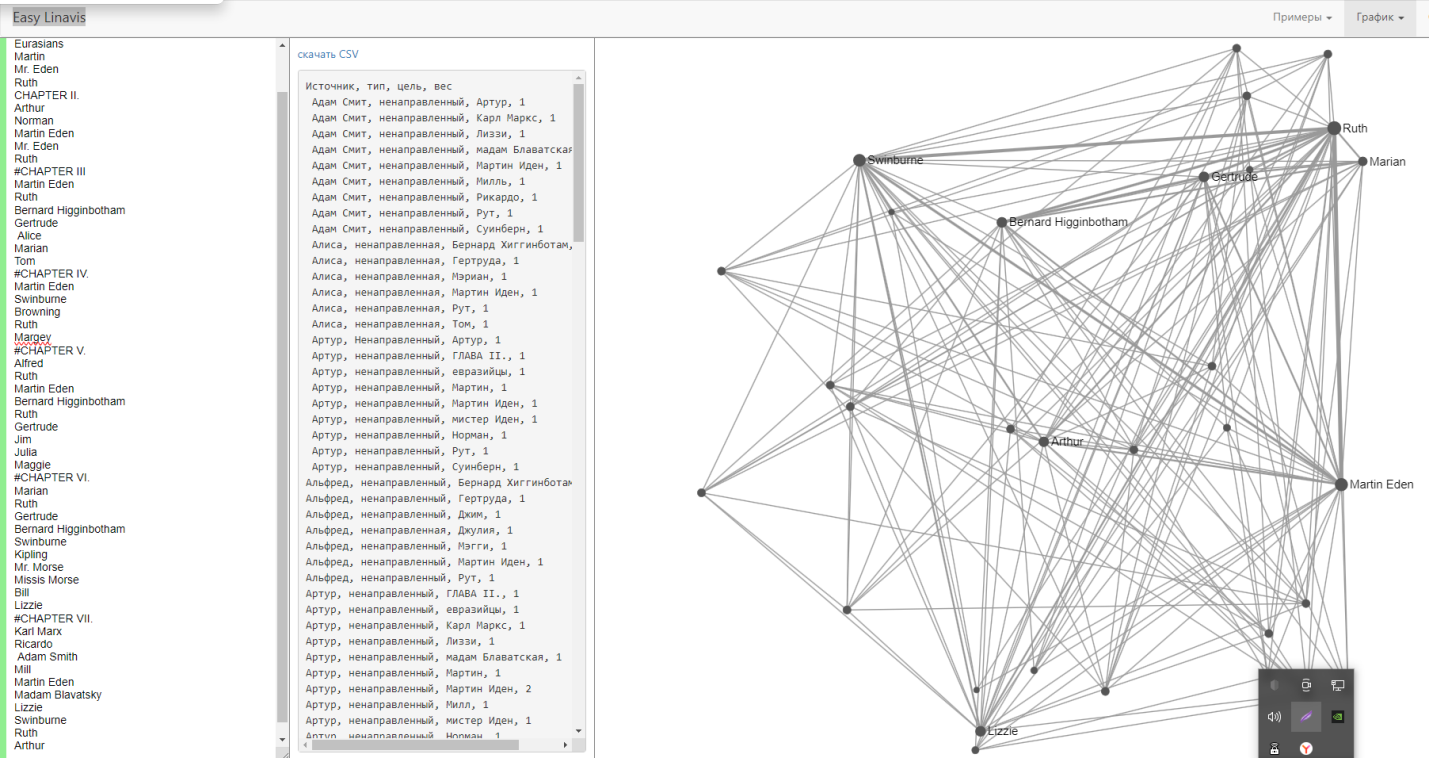
Oppose





Функция генерирует вектор слов, которым тестируемый автор отдает значительное предпочтение, и другой вектор, содержащий слова, которых значительно избегают.

Сетевой анализ  
Easy Linavis



**Уровень 2**

**Задание 3**

1. Ранг: Ранг - это порядковый номер слова по убыванию его частоты в тексте или корпусе. Он позволяет оценить важность слова в контексте анализа. Например, наиболее часто употребляемые слова будут иметь более низкий ранг.

2. Относительные частоты: Относительные частоты показывают долю конкретного слова или явления относительно общего количества слов или явлений в тексте. Они позволяют сравнивать частоту различных элементов и выявлять закономерности.

3. Показатель R (range): Показатель R, или диапазон, представляет разницу между наибольшим и наименьшим значениями в выборке. Он может использоваться для измерения вариации в частоте употребления слов или явлений.

4. Коэффициент Жуйяна: Коэффициент Жуйяна используется для оценки равномерности распределения частоты слов в тексте. Он вычисляется как отношение суммы квадратов частот слов к квадрату суммы всех частот.

5. ARF (Average Rank Frequency): ARF представляет собой средний ранг слова в тексте или корпусе. Эта метрика позволяет оценить общую значимость слова на основе его ранга.

6. Критерий логарифмического правдоподобия: Критерий логарифмического правдоподобия используется для оценки степени соответствия модели частотного распределения данным. Он позволяет оценить точность модели и ее способность предсказывать частоту слов.

7. Критерий Йенсена-Шеннона: Критерий Йенсена-Шеннона используется для измерения разнообразия текста на основе энтропии распределения частот слов. Он позволяет определить степень разнообразия лексического состава текста.

**Задание 4**

Глоттохронология - это научная дисциплина, занимающаяся реконструкцией исторических языковых связей на основе сравнительного анализа лексики и фонетики. Одним из методов глоттохронологии является метод Сводеша-Старостины, который использует количественные данные для оценки времени разделения языков.

Формула, используемая в методе Сводеша-Старостины, выглядит следующим образом:

T = log(2)/R

Где:

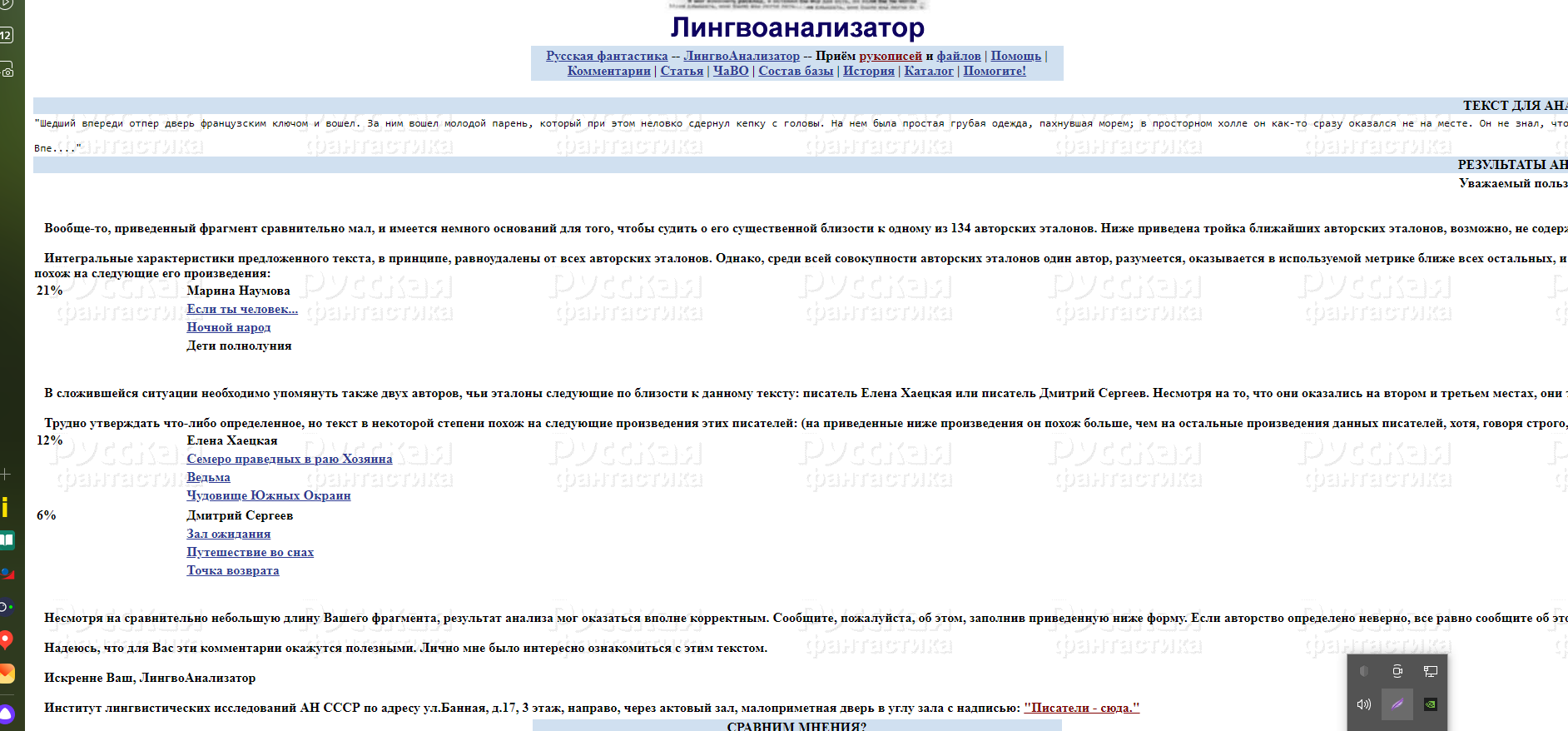
- T - время разделения двух языков (в тысячелетиях);

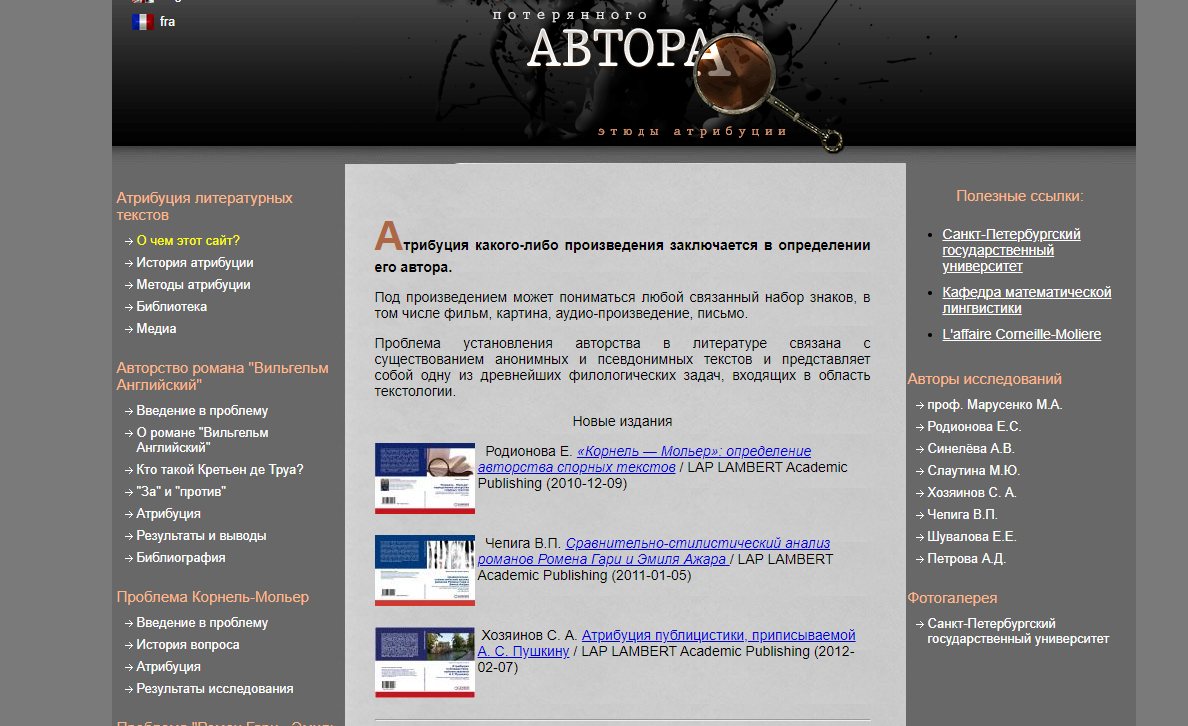
- R - средняя доля лексики, сохраняющейся в двух языках за определенное время.

Эта формула позволяет оценить примерное время разделения двух языков на основе степени сходства их лексики. Чем выше значение R, тем более близки языки по лексике и тем менее времени прошло с момента их разделения.

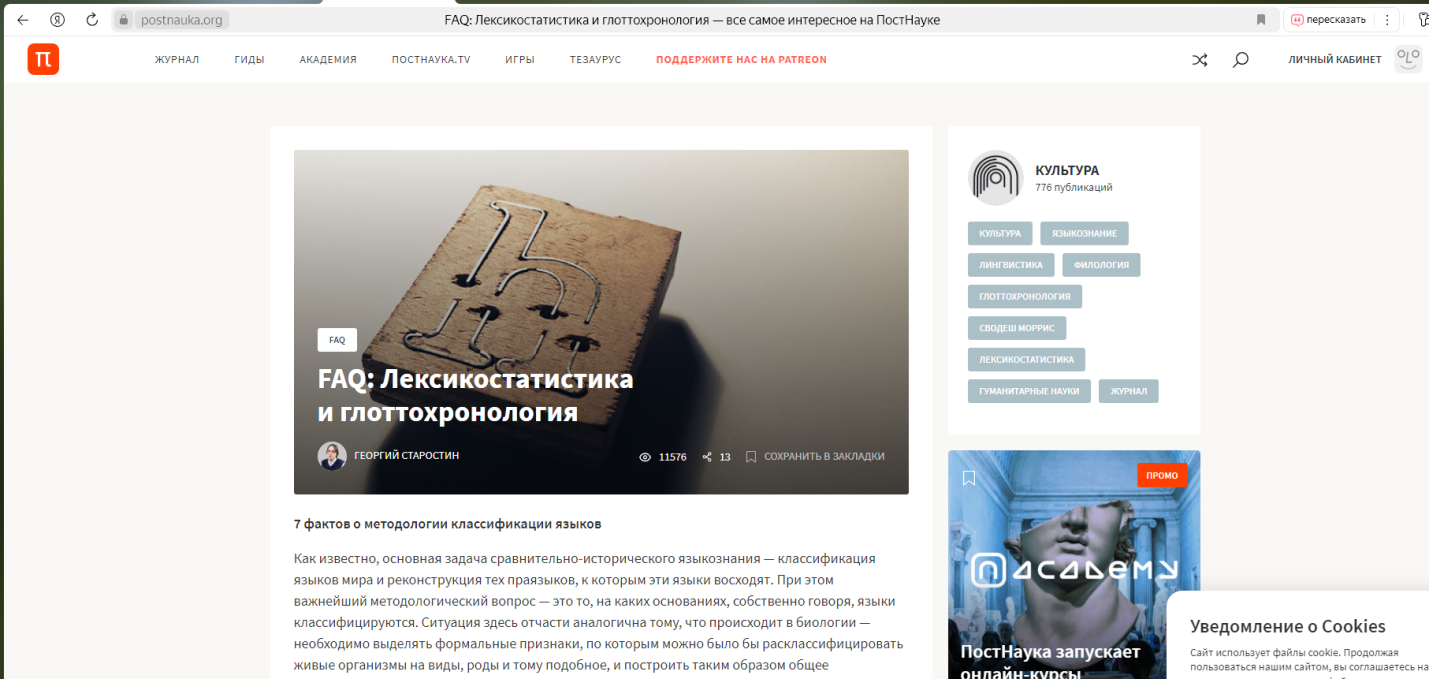
Метод Сводеша-Старостины и соответствующие формулы используются для реконструкции исторических языковых связей и определения времени разделения языковых групп.

**Задание 5**

1. http://www.rusf.ru/books/analysis/ - Этот ресурс представляет собой сайт, посвященный вопросу определения авторства текстов. Автор Д. В. Хмелев разработал лингвоанализатор для определения авторства анонимных или псевдонимных произведений. 

2. http://attribution.corneille-moliere.com/?p=page19&m=main&l=rus - Этот информационный ресурс также посвящен проблеме определения авторства анонимных и псевдонимных произведений. Он может содержать методы и инструменты для анализа текстов и определения их авторства. 

3. http://postnauka.ru/faq/11053 - Ссылка на статью или материал на портале "ПостНаука" по вопросам авторства или другим лингвистическим темам.



4. http://starling.rinet.ru/Texts/textsr.htm - Этот ресурс, возможно, содержит тексты на различных языках для анализа и сравнения.



5. http://dict.ruslang.ru/freq.php - Новейший академический частотный словарь русского языка, который вероятно содержит информацию о частотности употребления слов в русском языке.

6. http://www.artint.ru/projects/frqlist.php - Частотный словарь одного из авторов предыдущего словаря, возможно, предоставляет данные о частотности слов в текстах этого автора.

7. http://morpher.ru/Russian/Stats.aspx - Частотный словарь, созданный программистом для сравнения с предыдущими двумя. Этот ресурс вероятно содержит статистические данные о частотности слов в русском языке и может использоваться для сравнительного анализа.



Каждый из этих ресурсов может быть полезен для анализа текстов, определения авторства, изучения частотности слов и других лингвистических исследований.

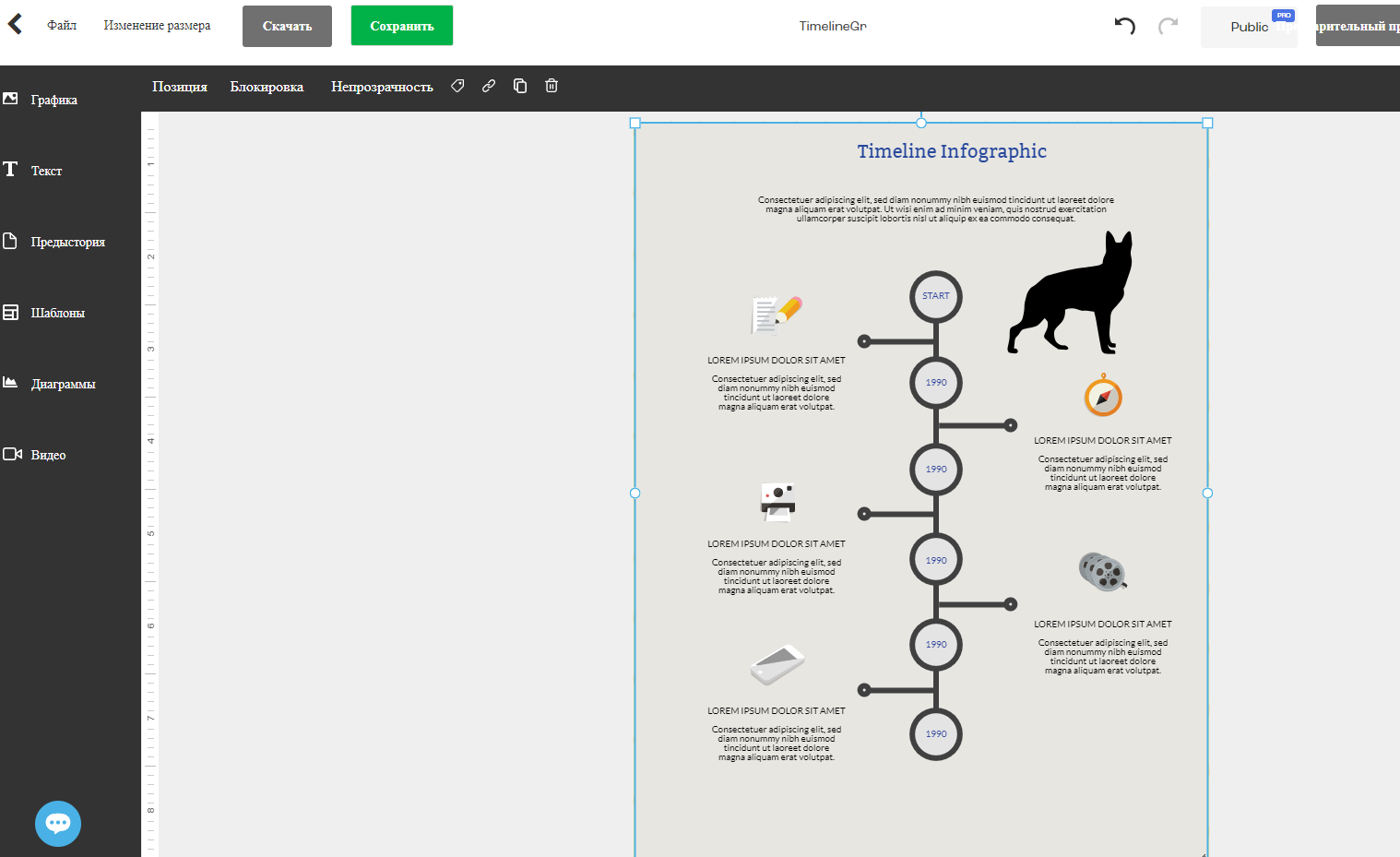
**Задание 2**

1. Coggle - это онлайн-инструмент для создания диаграмм и ментальных карт. С его помощью пользователи могут визуализировать идеи, организовывать информацию и делиться своими мыслями с другими. Coggle предлагает простой и интуитивно понятный интерфейс, который позволяет создавать креативные и структурированные диаграммы.

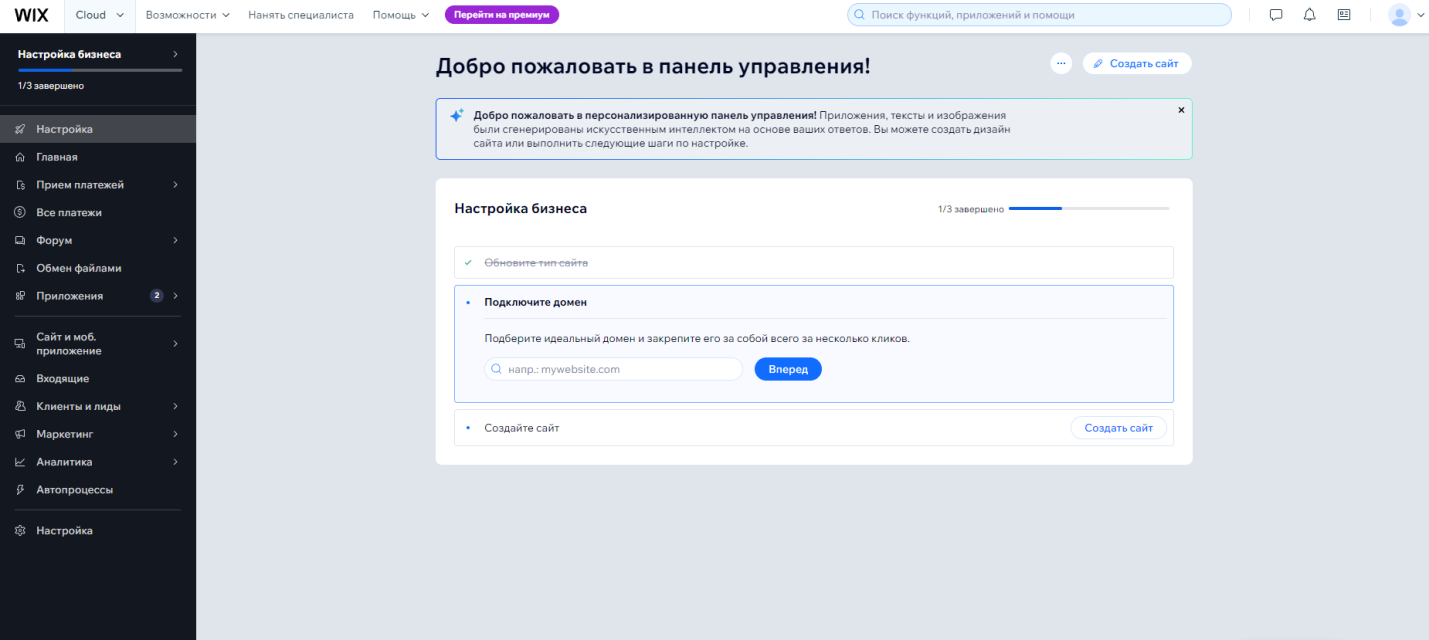
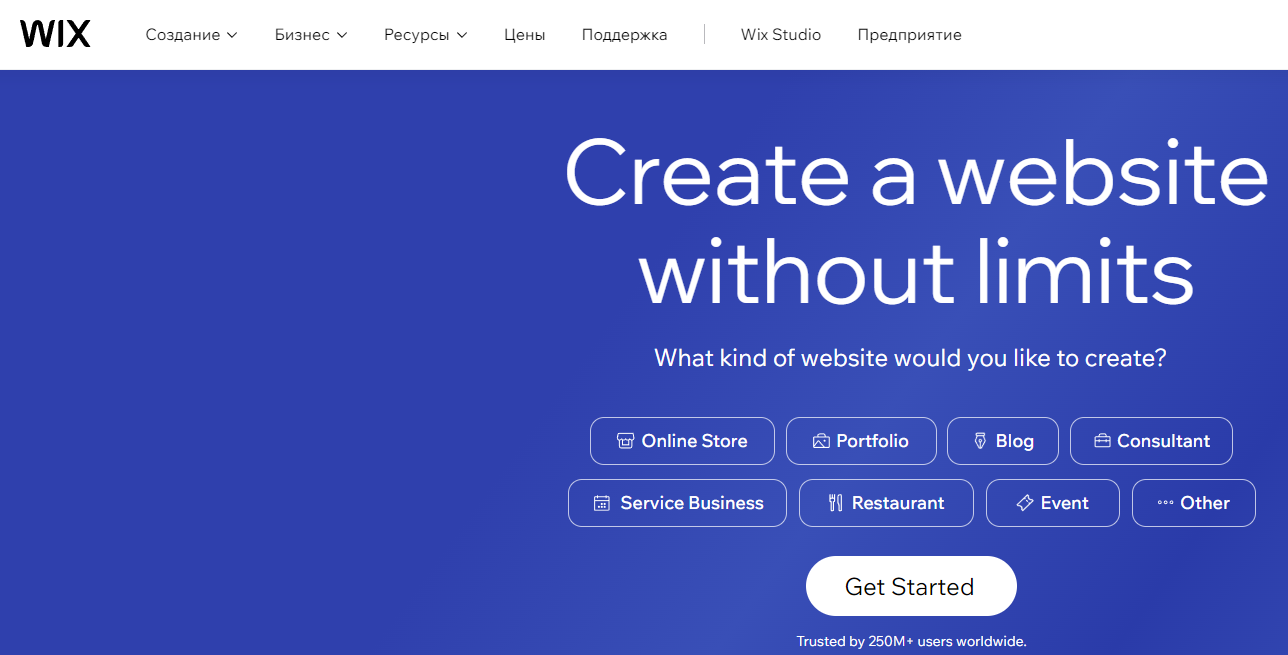
2. Canva - это популярный графический дизайнерский инструмент, который позволяет пользователям создавать профессионально выглядящие дизайны без необходимости обладать специальными навыками дизайна. Canva предлагает широкий выбор шаблонов, изображений, шрифтов и элементов дизайна, чтобы помочь пользователям воплотить свои творческие идеи.

4. Wix - это платформа для создания веб-сайтов с помощью графического интерфейса. Wix предлагает широкий выбор шаблонов, инструментов для редактирования и настройки дизайна, функциональности для добавления контента, а также возможность оптимизации сайта для мобильных устройств. С помощью Wix пользователи могут создавать профессионально выглядящие веб-сайты без необходимости знаний веб-разработки.

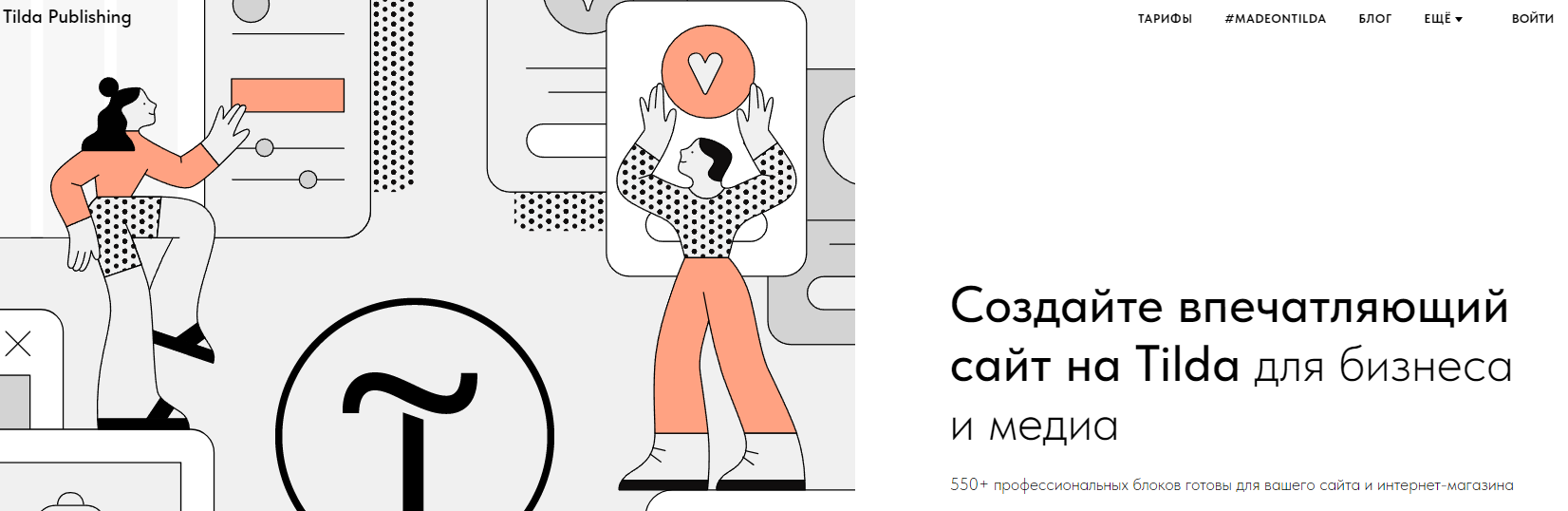
3. Easel.ly - это онлайн-инструмент для создания информационных графиков и инфографики. С его помощью пользователи могут легко и быстро создавать привлекательные и информативные графические изображения для презентаций, отчетов, блогов и других целей. Easel.ly предлагает различные шаблоны, графические элементы и инструменты для редактирования, чтобы помочь пользователям создавать убедительные визуальные материалы.

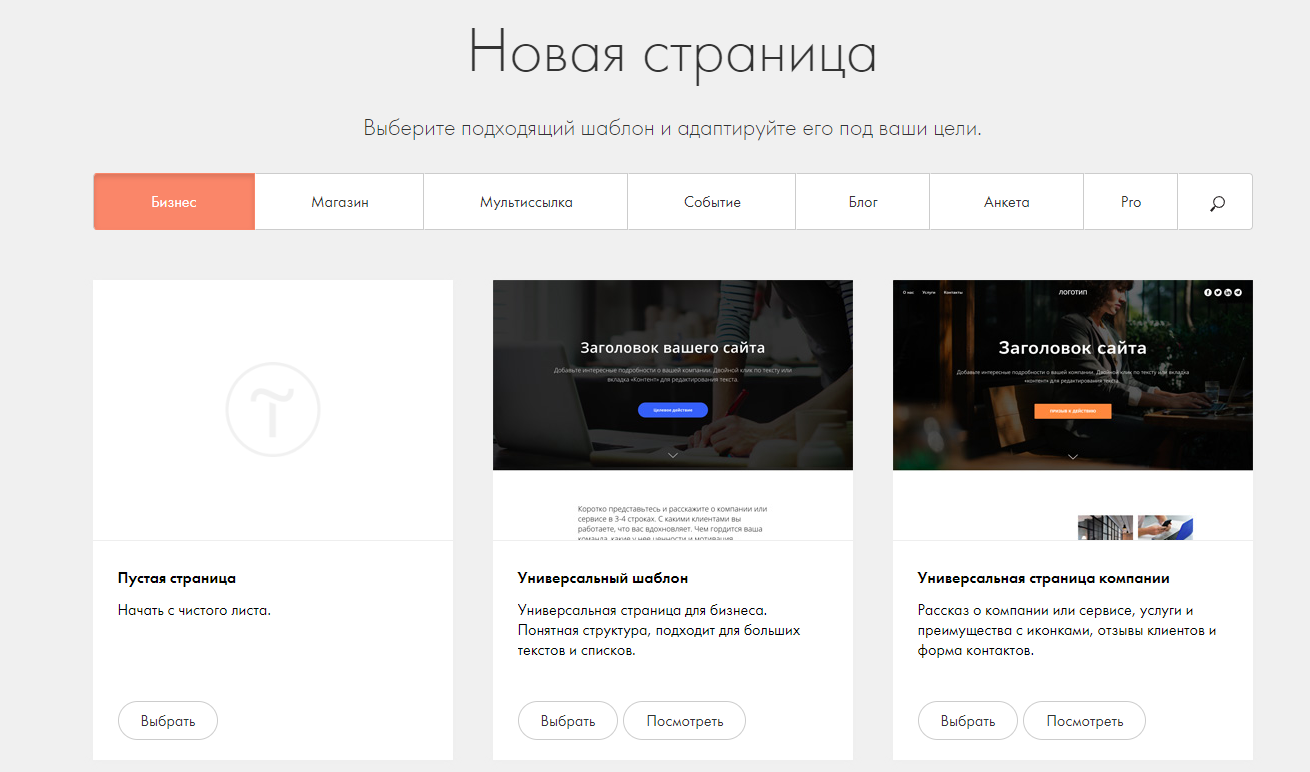
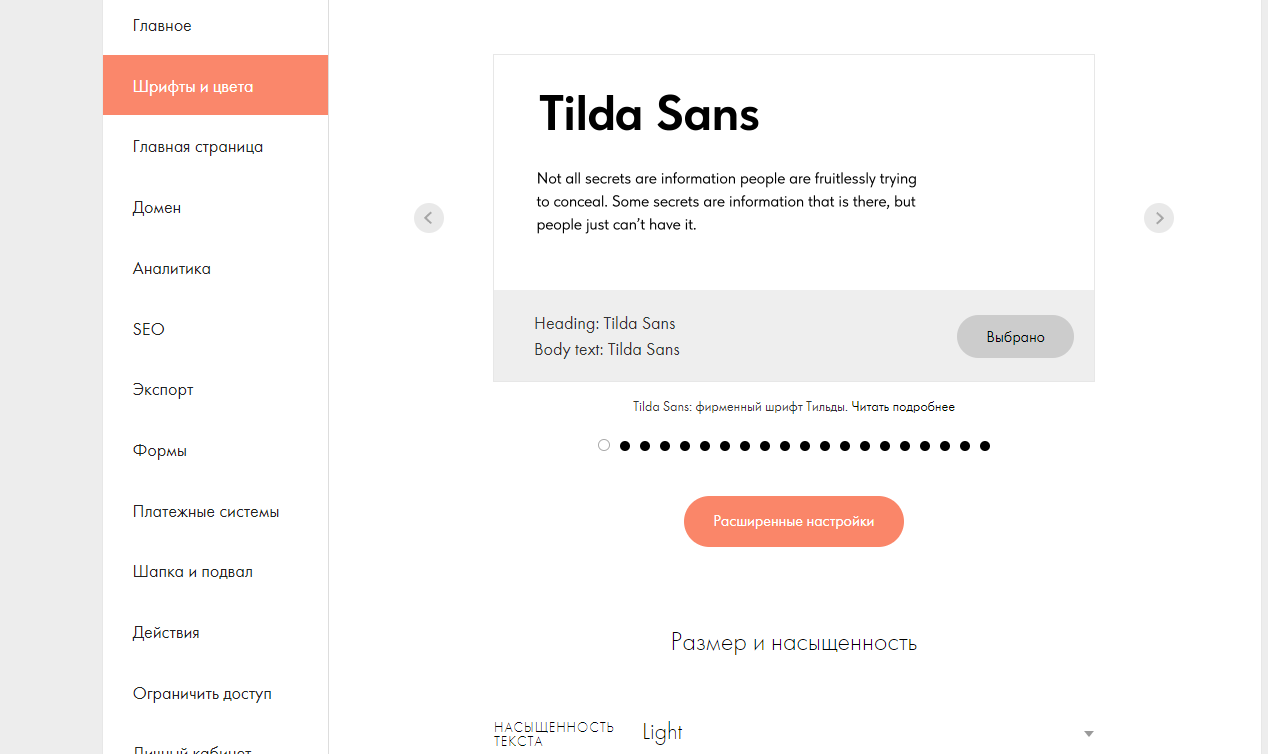
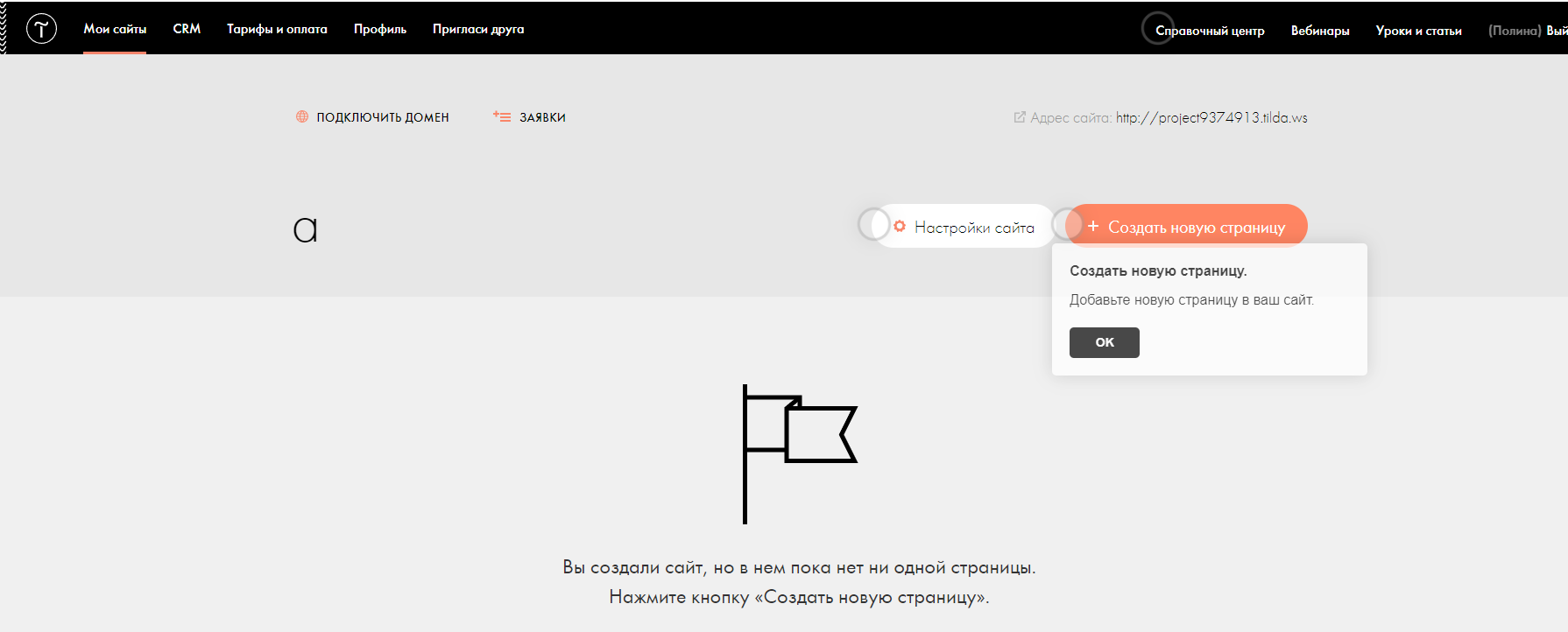


4.Wix предлагает широкий выбор шаблонов, инструментов для редактирования и настройки дизайна, функциональности для добавления контента, а также возможность оптимизации сайта для мобильных устройств. С помощью Wix пользователи могут создавать профессионально выглядящие веб-сайты без необходимости знаний веб-разработки.

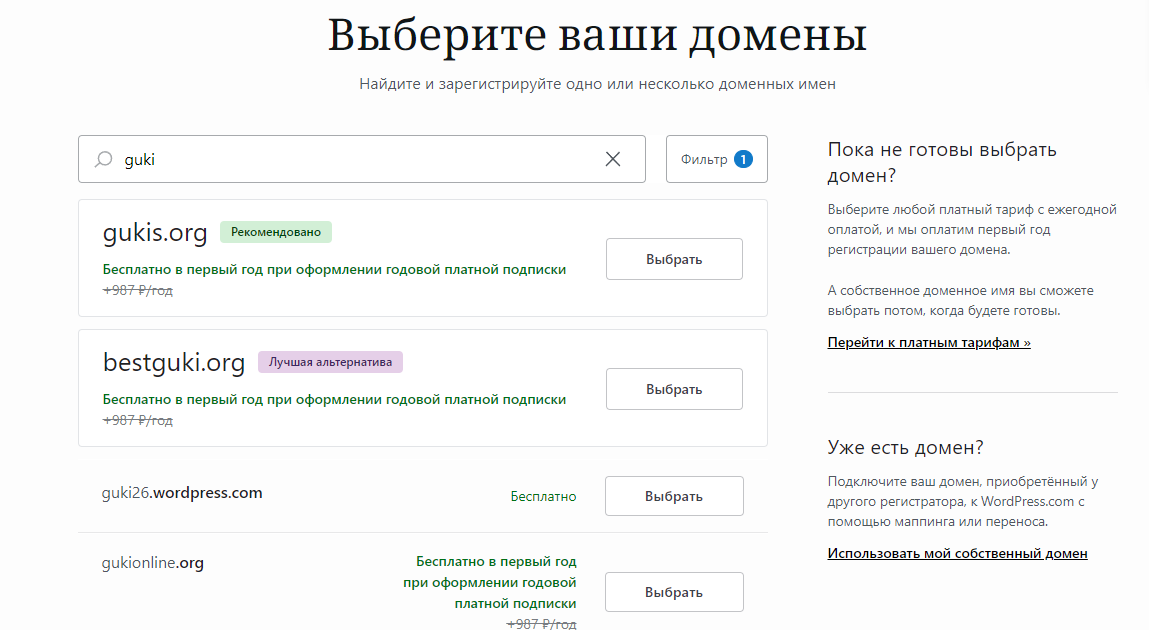


5. Tilda - это сервис для создания лендингов и веб-сайтов с помощью интуитивного визуального редактора. Tilda предлагает большой выбор блоков, элементов дизайна и шаблонов, чтобы помочь пользователям создавать красивые и функциональные веб-страницы. Сервис также предоставляет инструменты для анализа и оптимизации сайта.





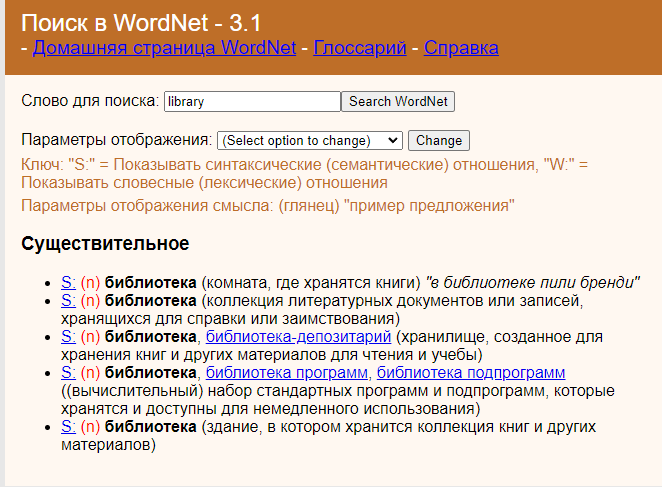
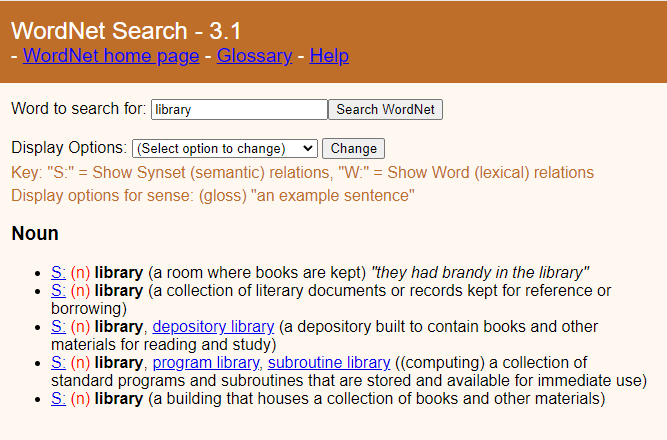
6. WordPress - это самая популярная платформа для создания блогов и веб-сайтов. WordPress предлагает гибкую систему управления контентом, широкий выбор тем и плагинов, а также возможность расширения функциональности сайта. С помощью WordPress пользователи могут легко создавать и управлять своими веб-проектами, независимо от их опыта веб-разработки.



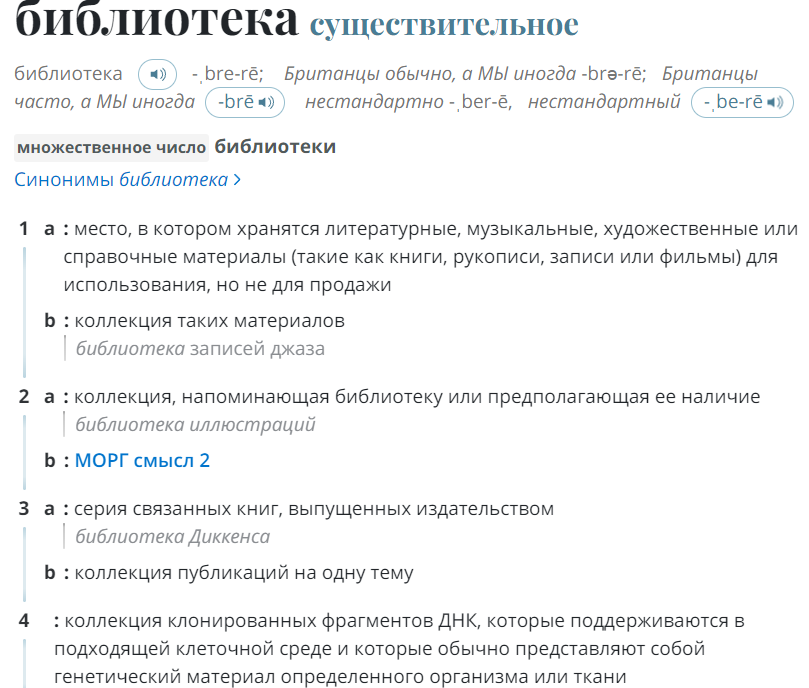
Уровень 3

Семантические сети

1. В WordNet было найдено 5 синсетов



**В другом словаре 4**

****

Несколько значений WordNet соответствуют одному значению в другом словаре

2.

****

**В WordNet части, где живут люди, связанные с понятием "house" (дом) как места, включают следующие синсеты:**

**1. "room" - комната**

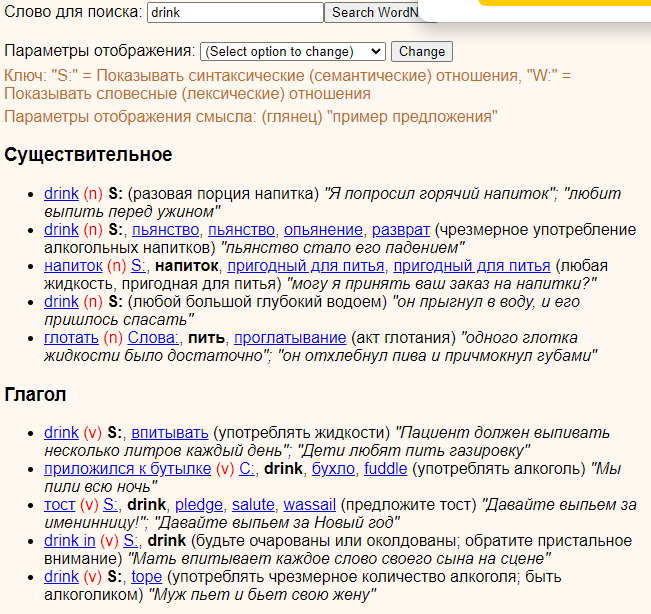
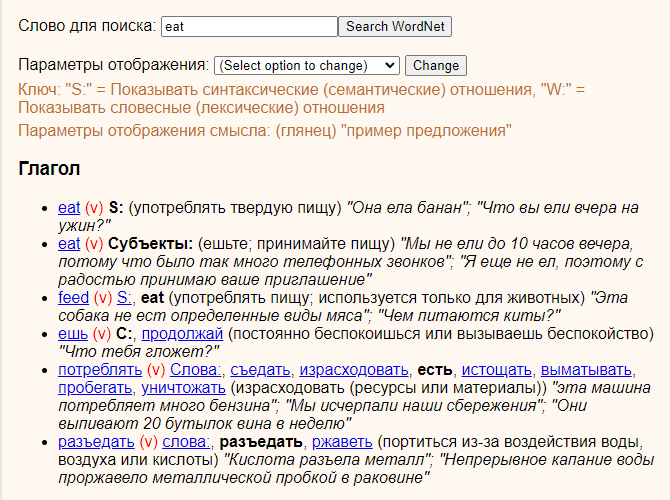
**2. "dwelling" - жилище**

**3. "residence" - местожительство**

**4. "home" - дом**

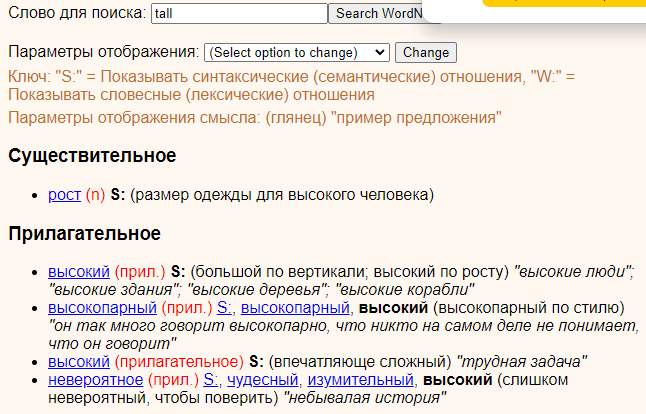
**Элементы, такие как ванны, кухни и спальни, относятся к синсету "room" (комната) в WordNet. Связь между этими элементами и понятием "house" (дом) как места, где живут люди, можно описать отношением "часть - целое". То есть ванны, кухни и спальни являются частями дома, где люди живут.**

**Синсет "room" (комната) связан с исходным синсетом "house" (дом) как места, где живут люди, посредством отношения включения. Комнаты являются частями дома, и вместе они образуют пространство, где люди проживают и проводят время. 3.Library, house имеют значение помещения.**

**4. **

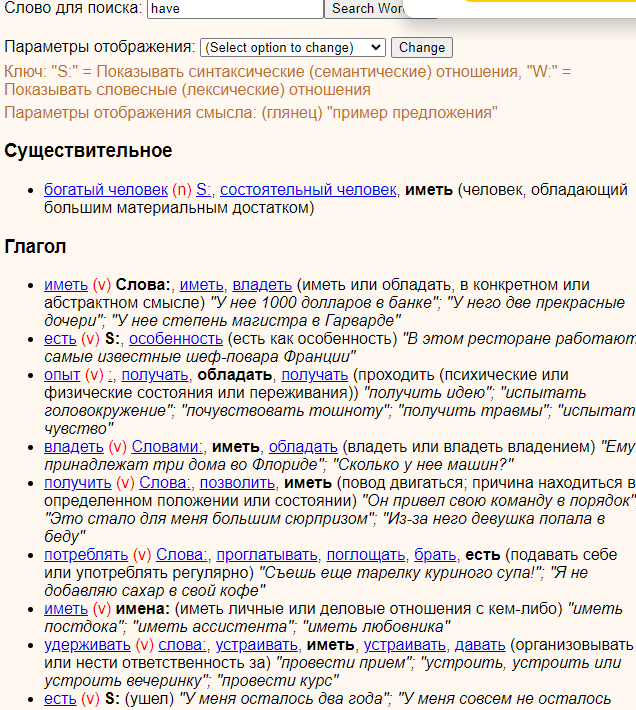
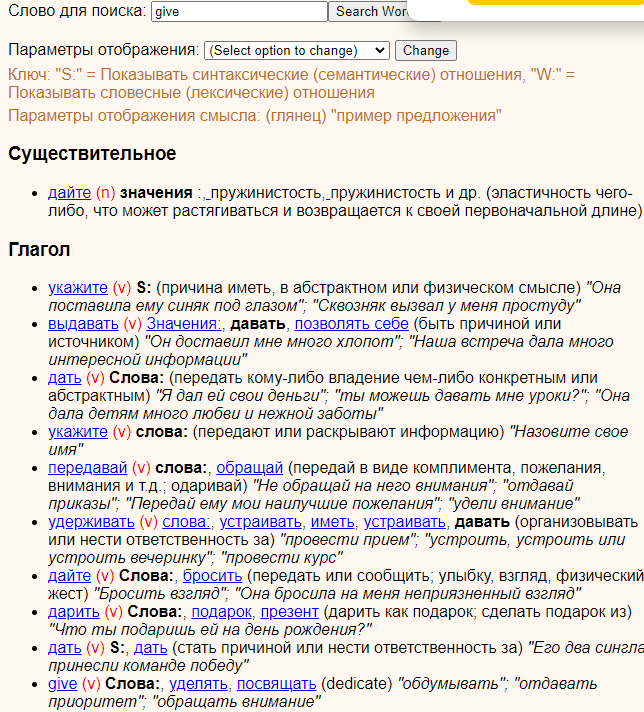
**Первые значения этих слов употребление пищи внутрь человеком**

**Eat означает прием пищи человеком или животным, разъедание металла, израсходовать что-то, беспокоиться о чем-либо**

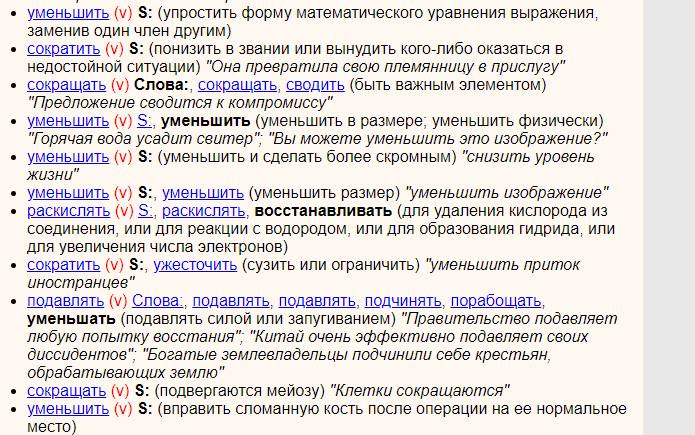
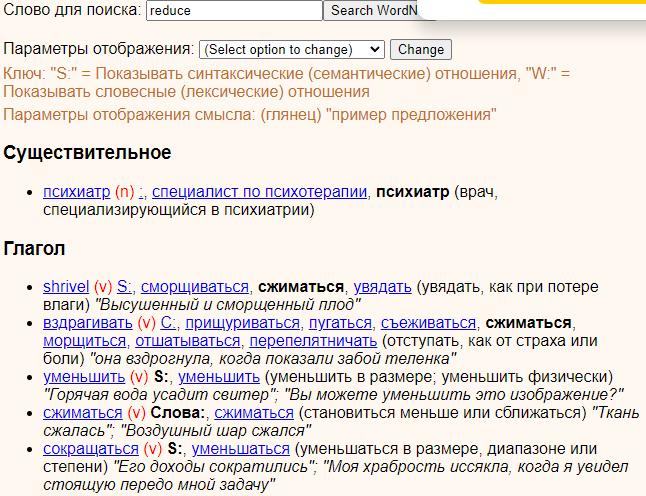
**5. **

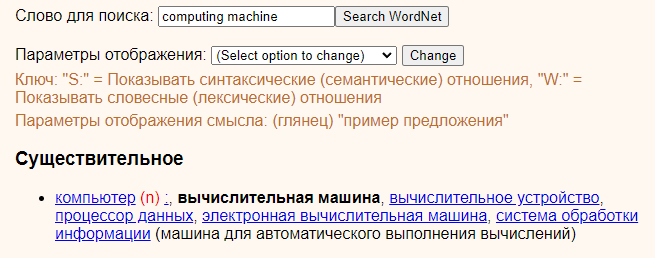
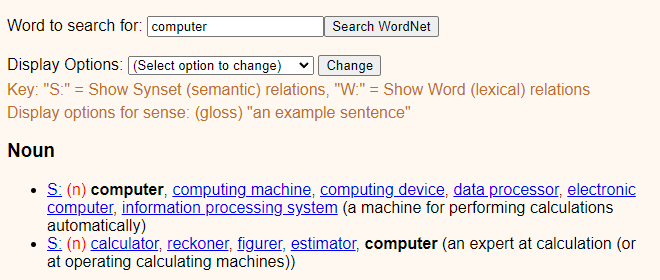
**Рост(размер)**

**6.**

****

**Глаголы "give" (давать) и "have" (иметь) связаны между собой в контексте передачи владения или собственности. Когда один человек дает что-то другому, это означает, что он передает этот объект в собственность или владение другого человека. Следовательно, глагол "give" выражает действие передачи, а глагол "have" указывает на наличие чего-то у кого-то после передачи.Слово "shrink" входит в синсет "shrink, reduce", который описывает действие уменьшения размера или объема чего-либо. Слово "shrink" также может встречаться в других синсетах, например, в синсете, описывающем психологическое состояние сжатия или ухода от реальности.**

****

**7. **

**Задание 2**

**Осмысление фактов, цифр и измерений – это искусство визуализации данных. Чтобы превратить цифры в знания задача состоит в том, чтобы правильно их представить данные.**

**Многие из нас принадлежат к "поколению PowerPoint" — именно здесь лежат корни нашего понимания визуализации и представления данных. PowerPoint мог бы стать самым мощным инструментом на вашем компьютере. Но это не так. Бесчисленные инновации терпят неудачу, потому что их сторонники используют PowerPoint так, как того хочет Microsoft, а не как следует. Сет Годин, эксперт по маркетингу Нет сомнений в том, что PowerPoint был, по крайней мере, частью проблемы, поскольку повлиял на целое поколение. Еще в 90-е годы к нему должна была прилагаться предупреждающая надпись и хороший набор инструкций по дизайну.**

**Столбчатые диаграммы хороши для сравнения, в то время как линейные диаграммы лучше подходят для отображения тенденций. Точечные диаграммы хороши для описания взаимосвязей и распределений, но круговые диаграммы следует использовать только для простых композиций, а не для сравнений или распределений. Существует диаграмма выбора диаграмм, созданная доктором Эндрю Абелой, которая поможет вам выбрать диаграмму, соответствующую вашему типу данных.**

**Таблицы, по сути, являются источником для всех диаграмм. Их лучше всего использовать для сравнения, составления структуры или анализа взаимосвязей, когда имеется всего несколько переменных и точек данных. Не имеет особого смысла создавать диаграмму, если данные могут быть легко интерпретированы из таблицы.**

**Столбчатая диаграмма, вероятно, является наиболее часто используемым типом диаграмм. Эту диаграмму лучше всего использовать для сравнения различных значений, когда важны конкретные значения, и ожидается, что пользователи будут искать и сравнивать отдельные значения в каждом столбце. С помощью столбчатых диаграмм вы можете сравнивать значения для разных категорий или сравнивать изменения значений за определенный период времени для одной категории.**

**Линейчатые диаграммы - это, по сути, горизонтальные столбчатые диаграммы. Если у вас длинные названия категорий, лучше всего использовать линейчатые диаграммы, поскольку они дают больше места для длинного текста. Вам также следует использовать линейчатые диаграммы вместо столбчатых, когда количество категорий превышает семь (но не более пятнадцати) или для отображения набора с отрицательными числами.**

**Задание 5**

**В последние годы сетевая парадигма, использующая физические понятия и методы, стала основным и очень эффективным инструментом изучения реальных сложных систем. До этого сетевые структуры изучали в основном математики – специалисты по теории графов, и социологи. Рождением теории графов принято считать 1736 год, когда Леонард Эйлер (1707-1783), живший и работавший тогда в Петербурге, решил известную проблему семи мостов в старом Кенигсберге.**

**Доказательство знаменитой теоремы о четырех красках Аппелем и Хакеном (Kenneth Appel, Wolfgang Haken) в 1976 году – это наиболее известное последнее достижение в теории графов.**

**В отличие от социологов, физики исследуют статистические свойства сетей, например, законы распределение узлов по числу связей. Изучая эти вопросы, было обнаружено большое количество удивительных и интригующих свойств реальных сетей, на которые не обратили внимания математики и социологи. Эти свойства послужили стимулом для разработки новых теорий, моделей, измерений, выявления новых фундаментальных свойств сетей. Именно физические журналы в настоящее время публикуют подавляющее число наиболее важных работ в этой области.**

**Длины всех связей между смежными узлами считаются равными единице. Расстояние ij l между узлами i и j есть длина самого короткого пути (геодезическая линия) между ними в сети. Среднее межузловое расстояние \_ l есть среднее ij l по всем тем парам узлов (i,j) между которыми существует хотя бы один соединяющих их путь.**

**Диаметр сети Dl есть максимальное расстояние между узлами сети. Зависимость \_ l или Dl от размера N сети и зависит от архитектуры сети. Мы будем иметь дело с сетями, в которых эта зависимость растет медленнее, чем в регулярных решетках и фракталах. Кратчайшая длина пути сети и ее диаметр являются нелокальными свойствами сети.**

**Если связи в сети ориентированы, что показывается стрелками, направленными от одного узла к смежному узлу, то такая сеть называется ориентированной. Примером может служить сеть цитирования научных статей. В такой сети узлами являются научные статьи, а направленными (ориентированными) связями являются ссылки на статьи внутри другой статьи.**

**В 1967 году социолог из Гарвардского университета С. Милграм (Stanley Milgram) на основе своего социологического исследования сделал удивившее многих утверждение: каждого человека на земном шаре можно связать с любым другим человеком цепочкой из шести знакомых.**

**Социологические исследования показывают, что люди сгруппированы в малые кластеры друзей, коллег и знакомых и эти кластеры обычно слабо связаны с другими аналогичными кластерами. В социологии это называется кластеризацией. Поэтому чем лучше ваши друзья знают друг друга, тем в меньшей степени вы можете рассчитывать на то, что они помогут вам связаться с незнакомым вам человеком.**

**В 1999 году физик из университета Нотр Дам (США) Л. Барабаши (Laszlo Barabasi) вместе со своей аспиранткой Р. Альберт (Reka Albert) изучали свойства реальных сетей с несколько иной точки зрения. Если Строгатц и Воттс в своем исследовании сетей исходили из феномена «тесного мира», то Барабаши и Альберт решили исследовать закон распределения узлов некоторых реальных сетей по числу связей. Результат также оказался неожиданным.**

**Задание 3**

**В этой статье описывается программное обеспечение "Стилометрия с помощью R" (stylo), гибкий пакет R для высокоуровневого анализа стиля письма в стилометрии. Стилометрия (компьютерная стилистика) занимается количественным изучением стиля письма, например, для проверки авторства, что имеет значительный потенциал в криминалистической практике, а также в исторических исследованиях. В этой статье мы представляем возможности stylo для компьютерного анализа текста на примере ряда примеров из английской и французской литературы. Мы демонстрируем, как этот пакет особенно полезен при предварительном статистическом анализе текстов, например, в отношении авторского стиля письма. Поскольку stylo предоставляет привлекательный графический интерфейс пользователя для высокоуровневого исследовательского анализа, он особенно подходит для аудитории новичков, не имеющих навыков программирования (например, в области цифровых гуманитарных наук). Более опытные пользователи могут воспользоваться нашей реализацией серии стандартных конвейеров для обработки текста**

**Определение авторства играет важную роль в зарождающейся области стилометрии, или компьютерного анализа стиля письма (Juola, 2006; Stamatatos и др., 2000; Stamatatos, 2009; Koppel и др., 2009; Ван Халтерен и др., 2005). Хотя у этой области есть важные исторические предшественники (Holmes, 1994, 1998), в последние десятилетия наблюдается явный рост научного внимания к этой проблеме. Из -за своего нового характера воспроизводимость и сравнительный анализ по-прежнему представляют значительные проблемы в этой области (Стамататос, 2009). Общедоступные наборы эталонных данных трудно найти, главным образом из -за проблем с авторским правом и конфиденциальностью, и существует лишь несколько стабильных кроссплатформенных программных пакетов, которые широко используются в сообществе. К счастью, ряд недавних инициатив способствуют этому, например, недавнее участие в конкурсе на авторство в PAN (http://pan.webis.de), где, например, осуществляется эффективный обмен соответствующими наборами данных.**

**Обычно стилометрический анализ включает в себя сложный, многоэтапный процесс (i) предварительной обработки, (ii) выделения признаков, (iii) статистического анализа и, наконец, (iv) представления результатов, например, с помощью визуализации. С этой целью исследователям в настоящее время приходится прибегать к специальной комбинации проприетарных, зависящих от языка инструментов, которые нелегко переносить на разные платформы. Такие решения трудно поддерживать и обмениваться ими между отдельными исследователями (группами), что препятствует прямому воспроизведению результатов исследований и повторному использованию существующего кода. представленный пакет stylo предлагает богатый и удобный набор функциональных возможностей, который идеально подходит для быстрого исследовательского анализа текстовых массивов , а также для задач классификации, которые необходимы для установления авторства. Пакет предлагает реализацию основных методов, которые в настоящее время используются в данной области. Таким образом, его основное преимущество заключается в интеграции типичных процедур (например, предварительной обработки)**

**Метрика Уникальной особенностью stylo является то, что он предлагает эталонные реализации для ряда установленных показателей расстояния, полученных на основе многомерного статистического анализа, которые популярны в стилометрии, но редко встречаются за ее пределами. Лучшим примером здесь является дельта Берроуза (Burrows, 2002); это интуитивно понятный показатель расстояния, который привлек большое внимание сообщества, в том числе и с теоретической точки зрения (Hoover, 2004a, b; Argamon, 2011). Графический интерфейс пользователя Высокоуровневые функции пакета предоставляют ряд графических интерфейсов пользователя (GUI), которые могут быть использованы для интуитивной настройки ряда установленных экспериментальных рабочих процессов несколькими щелчками мыши (например , визуализация текстов без контроля на основе частотности слов). Эти интерфейсы могут быть легко вызваны из командной строки в R и предоставляют привлекательный обзор различных доступных экспериментальных параметров, позволяя пользователям быстро изучить основную стилистическую структуру корпусов. Эта особенность является особенной**

**Программа предварительной обработки stylo предлагает богатый набор опций для загрузки текстов в различных форматах из файловой системы (предпочтительно в кодировке UTF-8 Unicode, но она также поддерживает другие кодировки, например, в Windows). Помимо необработанного текста, stylo позволяет загружать тексты, закодированные в соответствии с рекомендациями Text Encoding Initiative, которая является относительно популярной в сообществе исследователей текстового анализа**

**Stylo упрощает дальнейшую обработку объектов, полученных в результате анализа. Чтобы удовлетворить потребности пользователей, не обладающих техническими знаниями, результаты, полученные в результате анализа, по умолчанию сохраняются в нескольких стандартных файлах и выводятся на экран. Опытные пользователи могут легко использовать возвращенные объекты при последующей обработке**

**8.** **В RussNet есть несколько различных значений для глагола "гордиться", и каждое из них имеет свои собственные отношения с другими словами в словаре. Рассмотрим некоторые из них:**

**1. Гордиться (чувство удовлетворения от достижений или качеств):**

**- Синонимы: тщеславие, самодовольство, самолюбие**

**- Антонимы: сомневаться, стыдиться**

**- Гипонимы: восхищаться, восхвалять, восхваление**

**2. Гордиться (принадлежность к чему-либо):**

**- Синонимы: иметь, обладать**

**- Антонимы: лишаться, утрачивать**

**- Меронимы: достижения, качества**

**3. Гордиться (проявление гордости, высокомерия):**

**- Синонимы: хвастаться, высокомерничать**

**- Антонимы: скромничать, унижаться**

**- Ассоциации: высокомерие, самовлюбленность**

**Каждое из этих значений глагола "гордиться" имеет свои особенности и связанные с ними слова в словаре RussNet. Например, в первом случае речь идет о чувстве удовлетворения от своих достижений или качеств, во втором - о принадлежности к чему-либо, а в третьем - о проявлении высокомерия или гордости.**