**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.

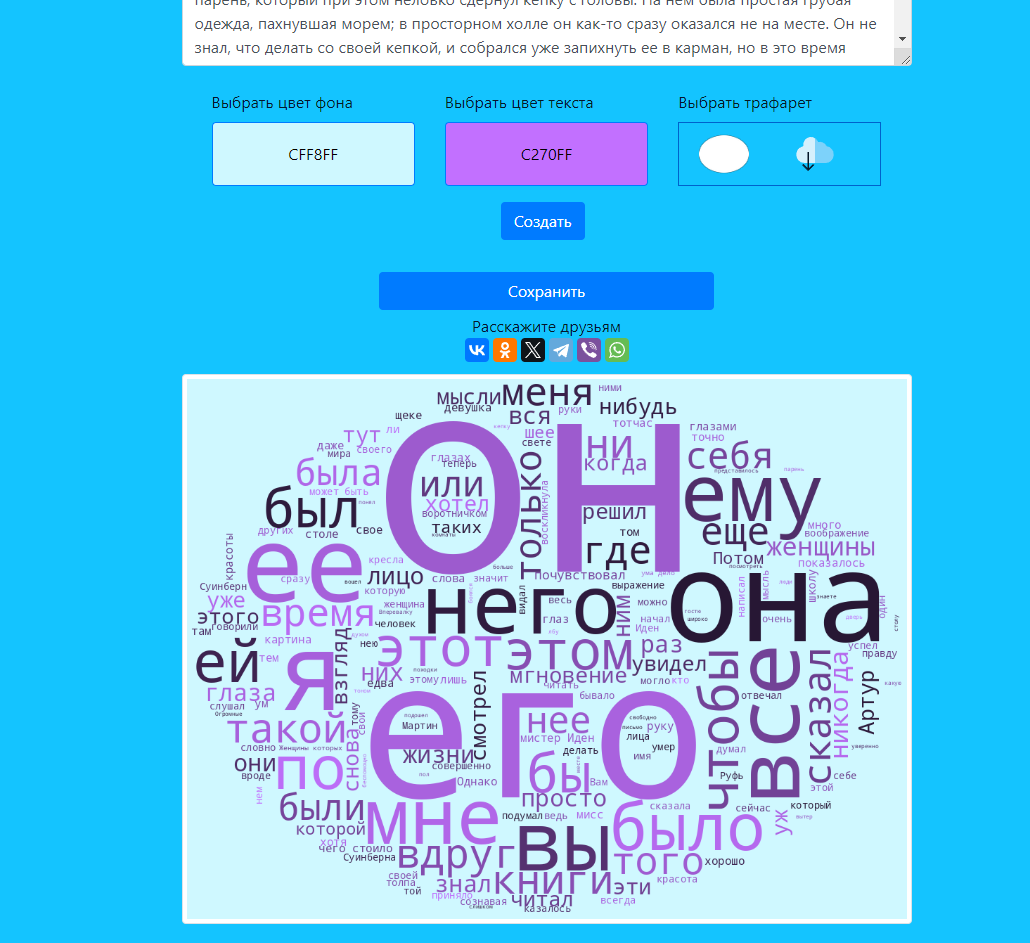
6. Протестировать визуализацию: перед окончательным представлением визуализации протестируйте ее на различных аудиториях, чтобы убедиться, что она эффективно передает нужную информацию.

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.

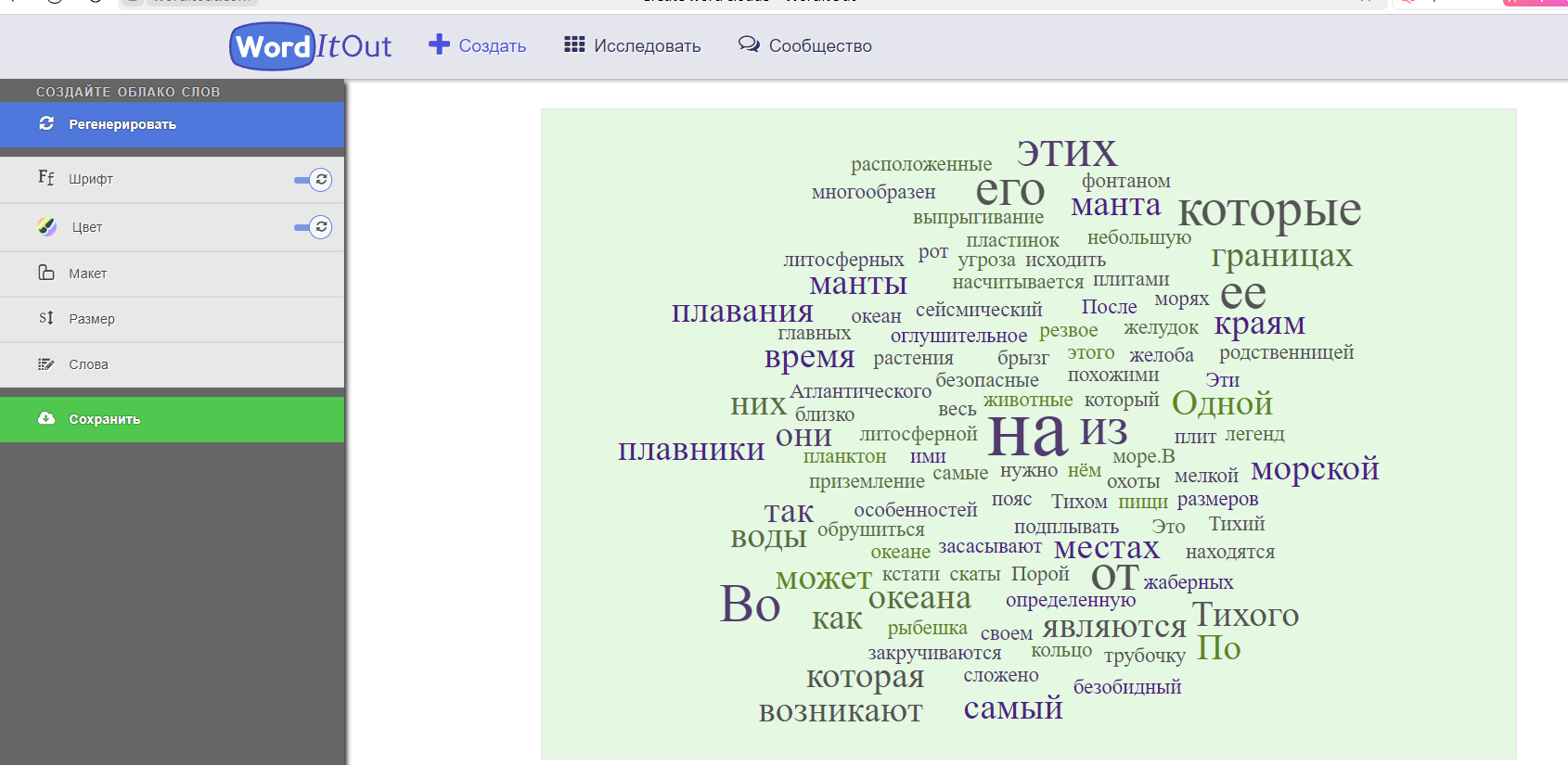
Word Clouds.

Первая глава произведения Джека Лондона «Мартин Иден»



Word it out

Статья про скатов



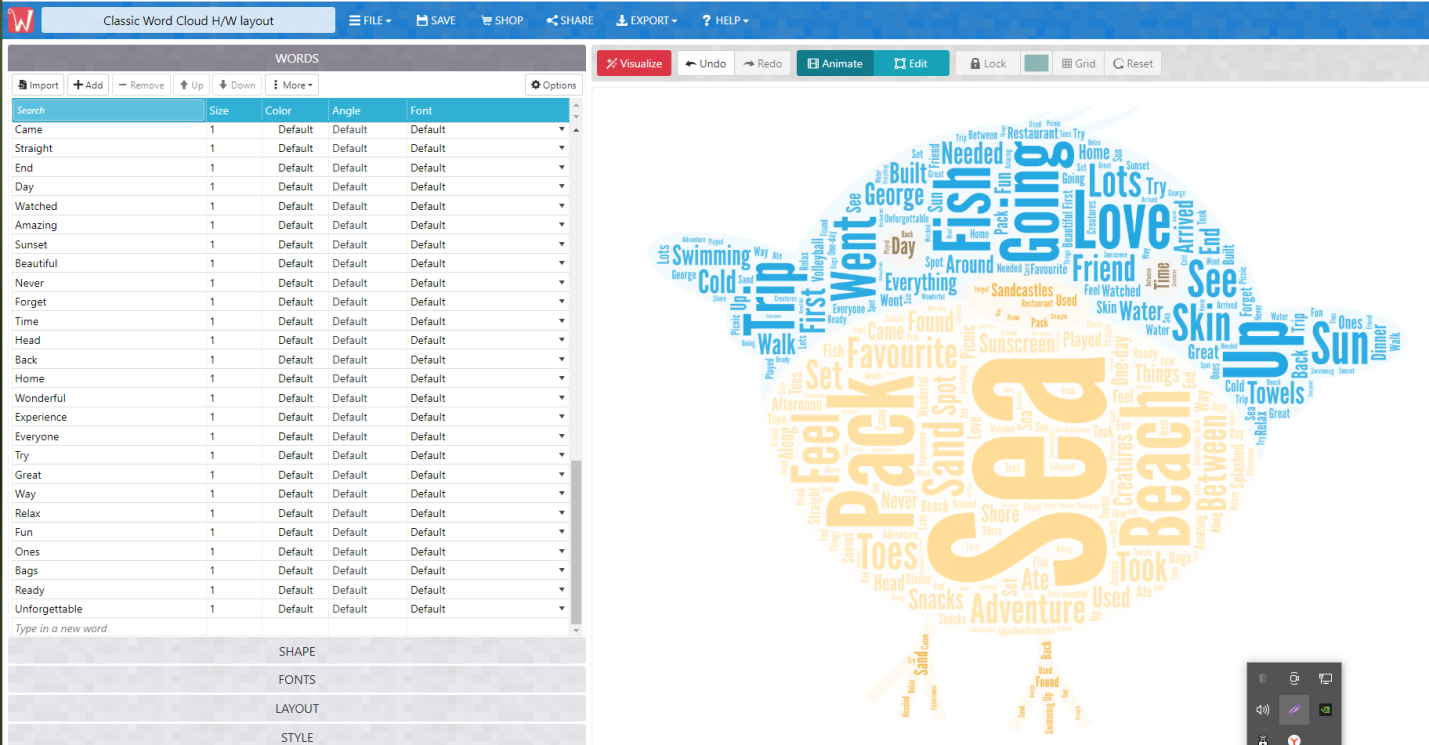
Word Cloud by Jason Davies

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.

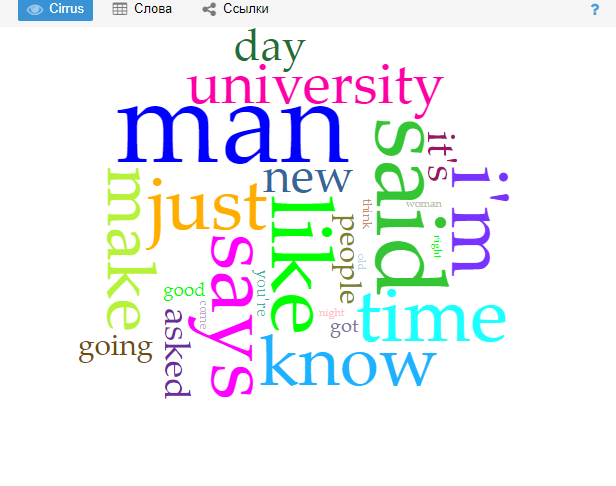
**VocabGrabber** (visualthesaurus.com/vocabgrabber): VocabGrabber - это онлайн-инструмент, который помогает анализировать текст и создавать облака слов на основе ключевых терминов. Пользователи могут увидеть частоту использования слов и их связи в тексте.

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



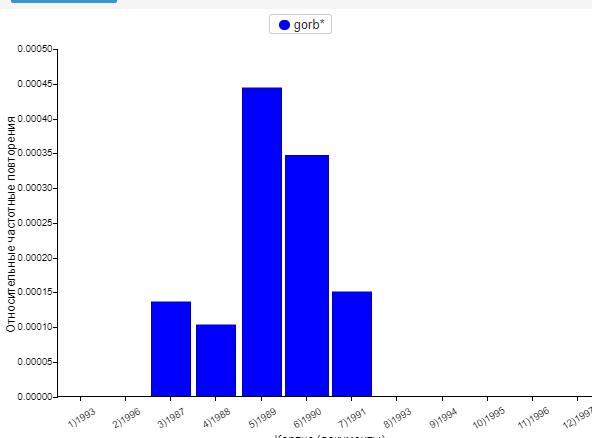
3.5

Значимые слова

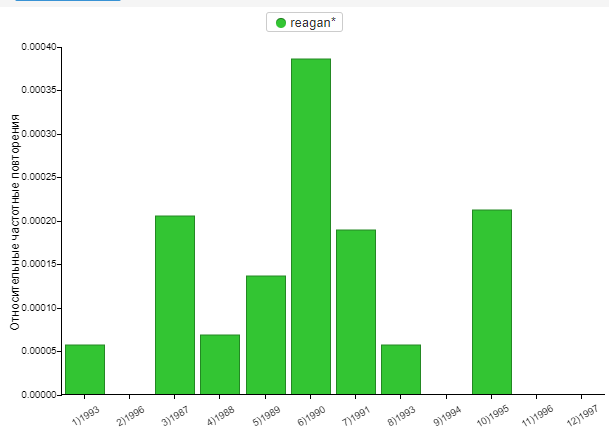




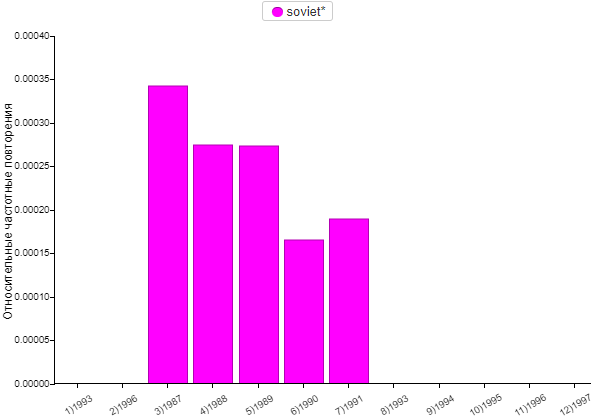
3.6



Постепенно прекратило употребляться

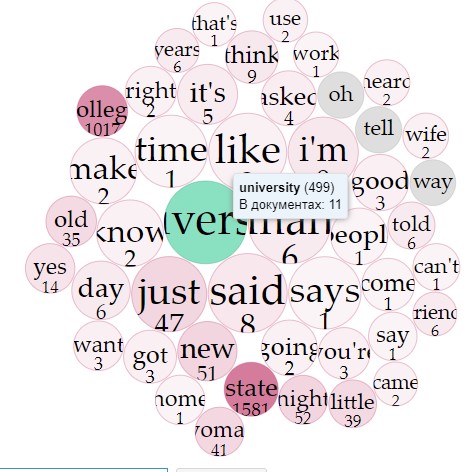


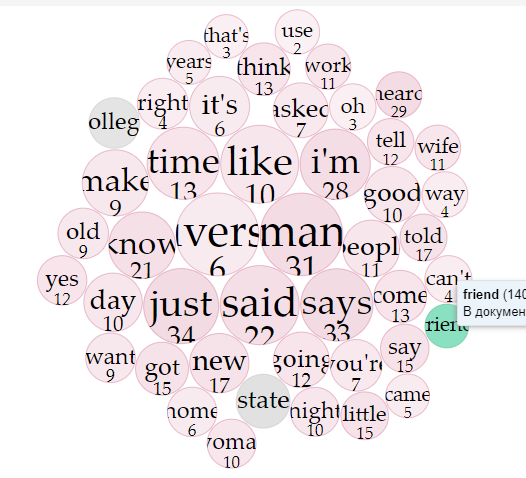
В 1990 употреблялось больше всего



Активно употреблялось несколько лет подряд.

3.7







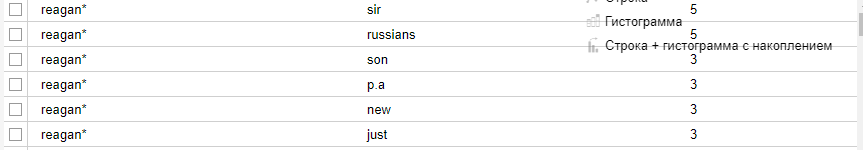
3.8



3.9

Walks into the bar чаще всего употребляется с man

3.10



**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.

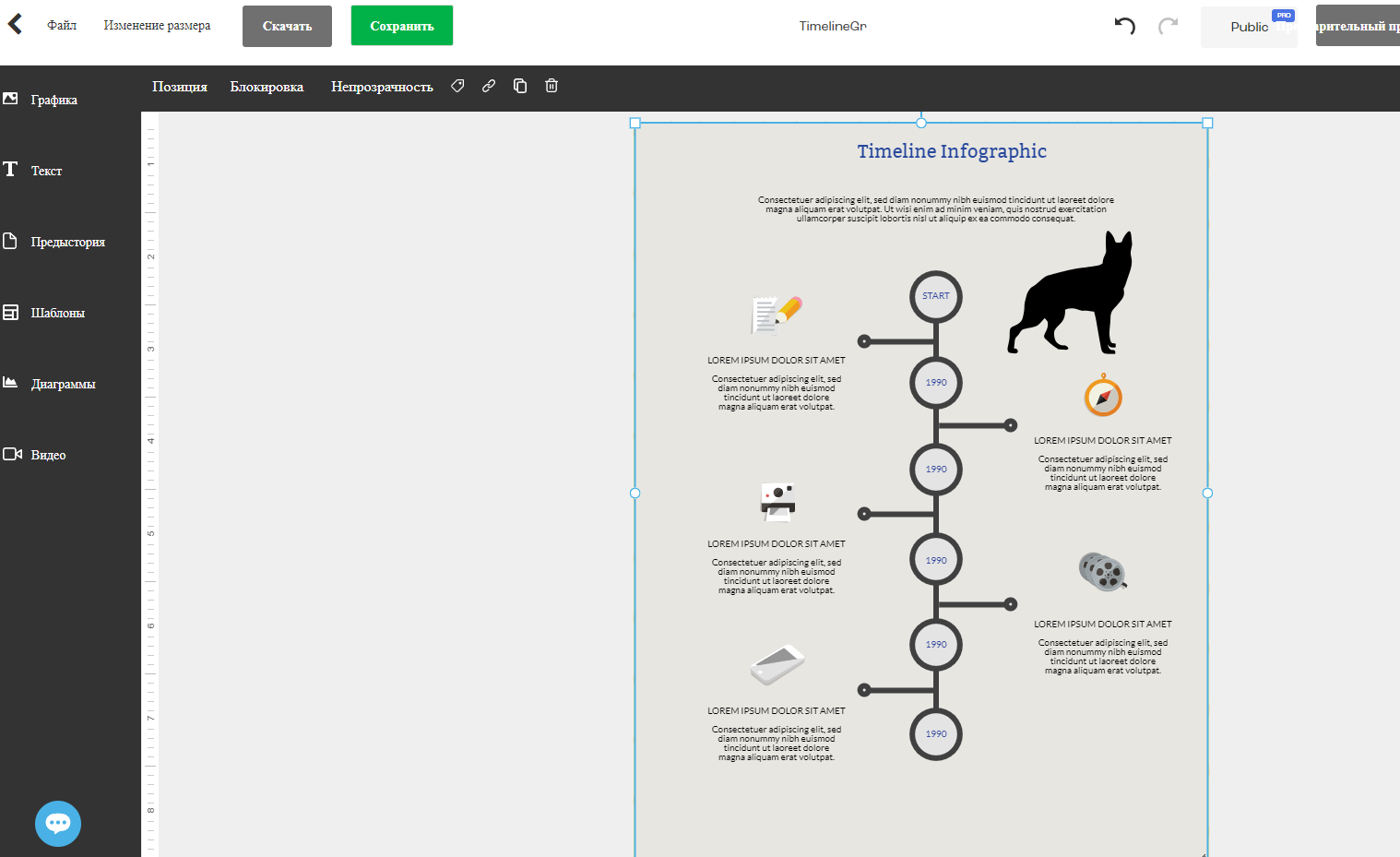
**Задание 2**

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.

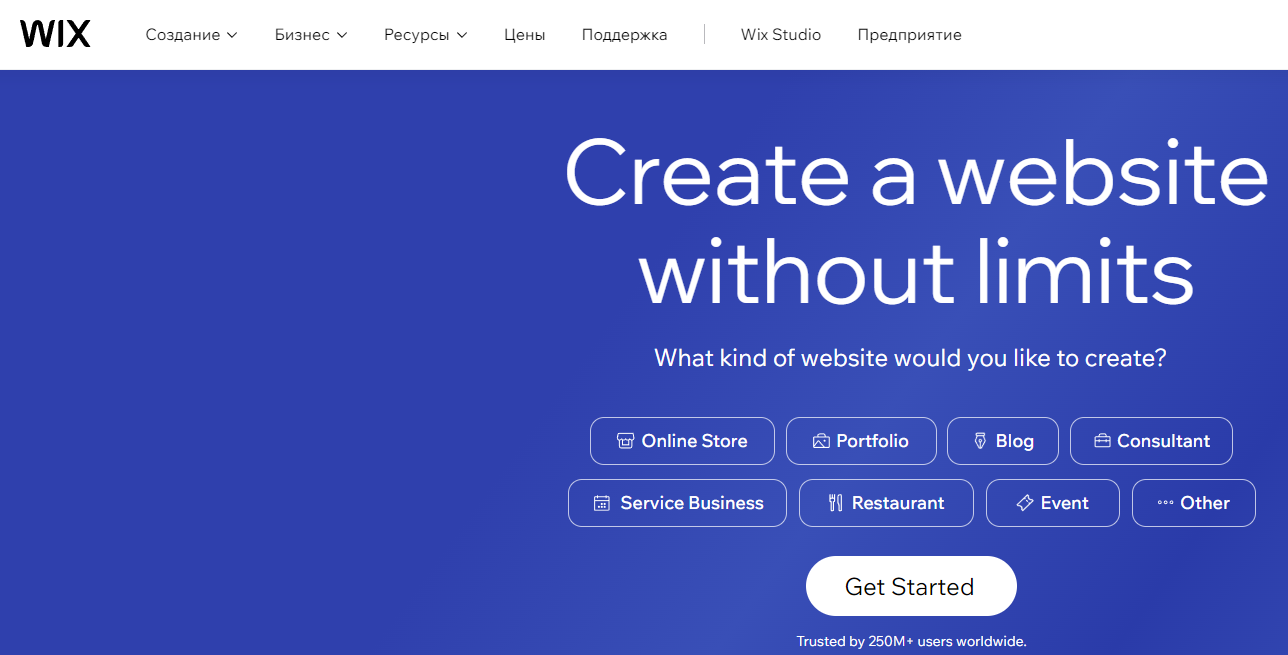
2. Canva - это популярный графический дизайнерский инструмент, который позволяет пользователям создавать профессионально выглядящие дизайны без необходимости обладать специальными навыками дизайна. Canva предлагает широкий выбор шаблонов, изображений, шрифтов и элементов дизайна, чтобы помочь пользователям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

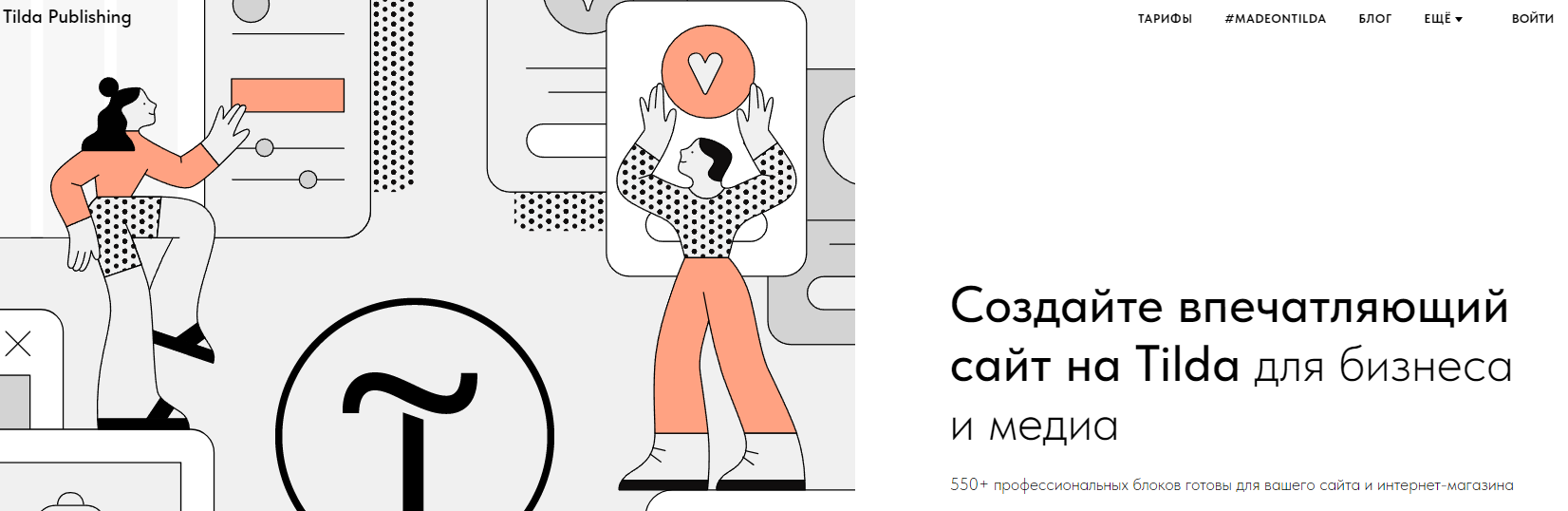
3. Easel.ly - это онлайн-инструмент для создания информационных графиков и инфографики. С его помощью пользователи могут легко и быстро создавать привлекательные и информативные графические изображения для презентаций, отчетов, блогов и других целей. Easel.ly предлагает различные шаблоны, графические 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.