Разработанное приложение предназначено для оценки изменений в обсуждаемых в новостных документах темах. Функционал приложения представлен на рисунке 2. Одной из основных целей является создание удобного интерфейса, позволяющего пользователю, не используя программный код, обработать текстовые данные, получить списки ключевых слов и облака слов, а также оценить динамику изменений в новостных сообщениях, собранных за разные периоды времени.

Приложение позволяет:

- производить загрузку срезов в виде файлов;
- выбирать срез, для которого будет сформирован список ключевых слов и построено облако слов;
- выбирать метод извлечения ключевых слов, с помощью которого будет производиться извлечение ключевых слов;
- строить тренд-карту с помощью выбранного метода для проанализированных срезов;
- строить матрицу косинусного подобия с помощью выбранного метода для проанализированных срезов;
- просматривать результаты вычислений;
- сохранить результаты вычислений.

При разработке приложения был сформирован список стоп-слов, который включал в себя русскоязычные союзы, частицы, предлоги, имена, фамилии, названия регионов, для которых применялась программа, и слова, не несущие смысловой нагрузки в контексте экономической тематики. Список стоп-слов был настроен для конкретных регионов.

На рисунке 1 представлен интерфейс приложения с выбранной вкладкой «Период 1». Интерфейс содержит следующие вкладки: «Период 1», «Период 2», «Период 3», «Период 4», «Период 5», «Оценка динамики». Пользователь может переключаться между вкладками. Находясь на вкладке периода, пользователь может загрузить файл, содержащий срез с новостными документами. Также он может выбрать метод извлечения ключевых слов, который будет применяться при обработке срезов. Выбор стоит между методами TF и RAKE. По умолчанию выбран метод TF. При вычислении индекса RAKE определяется минимальная частота встречаемости ККФ min_freq, равная 10. Максимальное количество слов в ККФ $ngram\_max$ определяется равным 3.

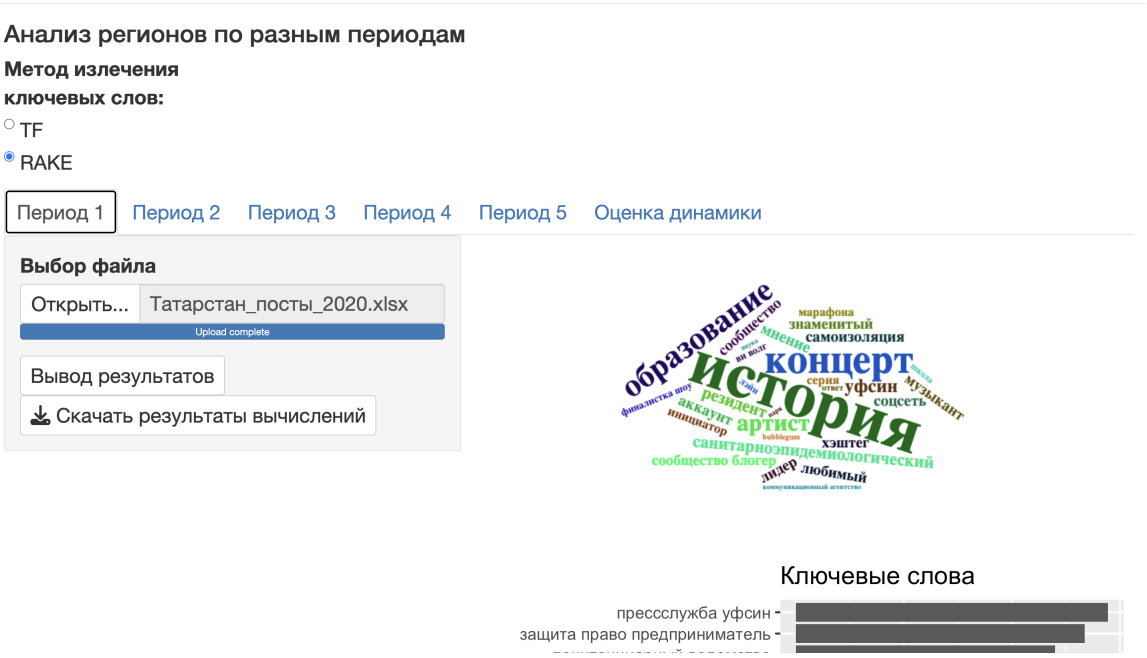

Рисунок 1 – Интерфейс приложения. Вкладка «Период 1»

После загрузки файла при нажатии на кнопку «Вывод результатов» пользователь видит оповещение о том, что ведутся вычисления, в правом нижнем углу экрана. После завершения вычислений выводится оповещение об их окончании, а также выводятся результаты работы алгоритма. В них входят:

— диаграмма с десятью ключевыми словами, с наибольшой характеристикой RAKE или TF, в зависимости от выбранного метода;
— таблица с десятью ключевыми словами с характеристиками RAKE или TF;
— облако слов с 30 наиболее часто встречающимися словами или фразами.

В программе существуют два списка, в которых хранятся результаты вычислений. Один список для метода извлечения ключевых слов TF, второй — для RAKE. По окончании обработки среза результаты вычислений, полезные для вычисления оценки динамики, сохраняются в соответствующие списки.

Оценка динамики может высчитываться как минимум для двух файлов. Подразумевается, что на вкладке «Период 1» будет находиться наиболее ранний период, а на вкладке «Период 5» наиболее поздний период. Если количество проанализированных файлов с помощью выбранного метода меньше двух, то при нажатии на кнопку «Сравнить проанализированные файлы» на вкладке «Оценка динамики» пользователь увидит оповещение с текстом: «Для анализа должно быть обработано не менее двух файлов с помощью выбранного метода.» Иначе будут выведены результаты, показанные на рисунке 3. Результатами оцен-

Рисунок 2 – Диаграмма прецедентов

ки динамики являются матрица косинусного подобия тем новостного контента и тренд-карта для 10 слов с наибольшей суммой нормализованных показателей динамики и значимости для определенного региона. От выбранного метода извлечения ключевых слов зависит то, как будут высчитываться тренд-карта и матрица косинусного подобия. Оба варианта вычислений представлены во втором разделе бакалаврской работы.

Рисунок 3 – Интерфейс приложения. Вкладка «Оценка динамики»