ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
“ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”»

Факультет гуманитарных наук
Образовательная программа
«Компьютерная лингвистика»

Кулакова Елена Петровна

СОСТАВЛЕНИЕ СЛОВАРЯ ЭМОЦИОНАЛЬНО-ОКРАШЕННЫХ ПРИЛАГАТЕЛЬНЫХ РУССКОГО ЯЗЫКА

Выпускная квалификационная работа студента 2 курса магистратуры группы 72/л

Академический руководитель
образовательной программы
канд. филологических наук, доц.
А.А. Бонч-Осмоловская

_____

« » июня 2015 г.

Научный руководитель

Д. А. Алексеевский

Москва 2015

ОГЛАВЛЕНИЕ

Введение

Задача автоматического определения тональности текста - одна из до сих пор полностью нерешенных задач компьютерной лингвистики. Спектр подзадач в этой области довольно широкий: от определения субъективности текста и выявления сарказма до выделения эмоциональной составляющей текста в виде бинарной полярной оценки или сложного комплекса эмоций.

Задача анализа тональности активно развивается в последнее время за счет возможности собирать большие корпуса эмоционально-окрашенных текстов из интернета и благодаря коммерческому применению в различных сферах деятельности, таких как: маркетинг, политика, психология, социология и других. Кроме того, анализ тональности находится прямо на стыке лингвистики и компьютерных наук. Самые различные подходы оказываются в той или иной степени успешными: как сугубо программное машинное обучение, так и ручные правила и словари.

Использование оценочных словарей существенно улучшает разметку при практически любом выбранном методе, поэтому создание качественного по содержанию и разметке словаря эмоционально-окрашенной лексики является востребованной задачей. Для английского языка существует открытый ресурс SentiWordNet [ES 2006], созданный несколько лет назад на основе англоязычного ворднета. Этот ресурс представляет собой базу данных прилагательных из ворднета. Каждому значению прилагательного методом машинного обучения приписана тональная оценка. Для русского языка такого словаря в открытом доступе пока что не существует. Повторить алгоритм создания английского ресурса на данный момент не представляется возможным по причине отсутствия русского ворднета. Тем не менее, возможны и другие способы создания тонального словаря.

В данной работе описывается составление и процесс разметки словаря эмоционально-окрашенных прилагательных русского языка «Сентинет». В первой части рассматриваются уже существующие подходы решения задачи на примере научно-исследовательских работ и программных ресурсов в этой области. Во второй части описывается процесс создания и пополнения списка словосочетаний оценочных прилагательных с существительными. В третьей части рассказывается об этапах создания игры для разметки словаря и анализируются полученные данные.

Часть 1. Задачи и методы анализа тональности

1.1 Предпосылки к развитию области анализа тональности

В сфере естественной обработки языка задача определения тональности является относительно новой. Например, первые работы по машинному переводу появляются еще в 1950-е годы, тогда как анализом тональности стали активно заниматься начиная с 2000-ых. Бурному развитию поспособствовали несколько факторов.

Во-первых, коммерческое применение. Для крупных компаний очень важно проанализировать как рынок воспринимает уже выпущенные продукты или спрогнозировать как воспримет новые. Для этого нужно оценить какие отзывы люди оставляют о продукте. Иногда отзывы сопровождаются численной оценкой, что способно облегчить задачу, но одна лишь оценка не позволяет узнать каким именно аспектом товара не удовлетворен покупатель, или наоборот, что именно в товаре ему понравилось. Слишком большое количество текстов отзывов не позволяет просматривать каждый вручную, поэтому даже самый простой бинарный классификатор, различающий положительные и отрицательные отзывы, существенно упрощает анализ.

Вторым фактором активного развития области анализа мнений является социализация интернета. На данный момент существует множество ресурсов, где пользователи могут высказать свое мнение относительно чего угодно: прочитанной книги, просмотренного фильма, купленного телефона. Причем оставить мнение можно в профиле в социальной сети или же на специализированном сайте, где помимо общей оценки предлагается также оценить некоторые аспекты объекта. Например, ресурс посвященный только

отзывам о гостиницах или ресторанах, предлагает отдельно оценить такие параметры как качество питания, размещение, сервис и другие.

Третьим фактором можно выделить сложную, но в тоже время интересную научно-исследовательскую составляющую задачи: помимо множества трудностей с лемматизацией, дизамбигуацией, кореференцией и других, характерных для задач обработки естественного языка в целом, возникают специфические только для этой области проблемы. Хорошим примером является наличие сарказма в текстах. Предложение может иметь явно положительную оценку, содержать исключительно позитивную тональную лексику и при этом нести негативный посыл. Такие случаи очень трудно классифицировать автоматически, даже человек не всегда может распознать сарказм. Удачный подход к решению задачи распознавания сарказма должен комбинировать методы машинного обучения и лингвистические наблюдения [BS 2015], что требует от исследователей владения самыми разнообразными навыками.

Современный интернет позволяет собрать огромную коллекцию текстов с явно выраженной эмоциональной оценкой, причем иногда эта оценка даже имеет численный эквивалент. На таких коллекциях текстов определенной тематики можно довольно успешно обучать классификаторы. В исследовательских работах по анализу тональности последних лет очень популярным источником данных является Твиттер. Посты в Твиттере ограничены по длине, что повышает информативность текста, и довольно часто содержат оценочное мнение пользователя по поводу чего-либо. Кроме того, они могут сопровождаться смайлами и хэштегами, которые могут нести оценку, указывать объект оценки или его аспекты, а также предупреждать об ироничности. Например, по хэштегу «сарказм» можно собрать целый корпус саркастических высказываний и использовать его как обучающую выборку, что и широко применяется на практике [DSR 2010, BS 2015].

1.2 Предмет исследований

В англоязычной литературе существует два термина, возникшие из отдельных подзадач, которые сейчас применяются для обозначения области в целом: *opinion mining* (извлечение мнений) и *sentiment analysis* (анализ тональности) [PL 2008]. Анализ тональности как подзадача подразумевает лишь присваивание положительной или отрицательной оценки тексту, предложению или набору слов. Тогда как извлечение мнений состоит из нескольких отличающихся друг от друга подзадач, часть которых совершенно нетривиальна и заслуживает отдельной работы [KH 2006].

Мнение или *оценочную ситуацию* можно определить как кортеж, состоящий из следующих компонент:

- субъект высказывания, который чаще всего совпадает с автором текста
- объект высказывания или цель – то, о чем высказывается субъект
- параметр объекта высказывания, если оценка дается не объекту в целом, а лишь какому-то его аспекту, зачастую параметр может быть скрытым
- собственно мнение - тональная оценка, положительный или негативный отзыв и возможная градация
- время высказывания
- адресат мнения, если таковой имеется

Не все эти компоненты являются обязательными при анализе. Некоторые компоненты могут быть явно выражены в данных изначально. Например, при анализе отзывов на кинофильмы, очевидно, что автор оценки – это тот, кто написал отзыв, объект оценки – кинофильм, время оценки – время публикации. Необходимо только определить тональность, которая в данном примере априори должна быть. Несет ли текст оценочное суждение или является нейтральным, то есть разграничение объективности и

субъективности, – это уже отдельная задача [WR 2005]. Полярность оценки тоже может быть определа по-разному. Иногда выделяют просто позитивную и негативную оценку, иногда добавляют в набор нейтральную оценку, иногда используют сложную шкалу с “ярко-выраженной негативной”, “слабой негативной” и другими оценками.

1.3 Основные подходы к задаче автоматического извлечения мнений

Можно выделить два основных метода анализа тональности текста: правиловый подход, опирающийся на вручную написанные правила и тональные словари, и статистический подход, представляющий собой машинное обучение на больших коллекциях текстов, иногда тоже с использованием тональных словарей. Также весьма успешно работают комбинированные методы.

Машинное обучение стало довольно классическим приемом при решении задачи анализа тональности. При наличии обучающей выборки алгоритм действий довольно стандартный. Рассмотрим его на примере работы [КК 2012а], выполненной в рамках семинара РОМИП 2011 года. На вход дана обучающая коллекция тонально размеченных текстов, содержащих оценочное мнение. В данном случае это были отзывы на книги и фильмы. После лемматизации текстов авторы составляют матрицу объект-признак из TF-IDF, то есть встречаемости термина в документе, рассчитаной по известной формуле. Далее сравнивается обучение с учителем, то есть когда алгоритм знает тональную оценку текстов обучающей выборки, и обучение без учителя, то есть кластеризация на два или несколько классов. Авторы сравнивают точность, аккуратность и F-меру метода опорных векторов (SVM), наивного Байеса и метода ключевых слов. Метод опорных векторов оказывается лучше Байса, ключевые слова слабо работают сами по себе, но

улучшают работу метода опорных векторов, поэтому самым лучшим подходом признана комбинация SVM и ключевых слов.

Поле экспериментов с методами машинного обучения довольно широко. Помимо выбора конкретных алгоритмов и их параметров, можно использовать самые разные признаки: N-граммы, символьные N-граммы, частоты, POS-метки, только слова определенных частей речи, слова из определенных наборов. К этому можно добавлять структуру предложения, контекст, эмотиконы, простые логические правила. Поэтому работ в области автоматического анализа тональности методом машинного обучения очень много, чему еще и способствуют регулярно проводимые соревнования, например, в рамках семинара РОМИП в 2011 году или конференции «Диалог» в 2015 году. Несмотря на то, что машинное обучение показывает весьма эффективные результаты на текстах одной предметной области, этот метод не является универсальным решением задачи анализа тональности.

Правиловый подход обладает хорошей точностью, но требует определенной и иногда трудоемкой подготовительной работы. Например, на одних и тех же данных правиловый подход, основанный на шаблонах с применением относительно небольшого словаря тональной лексики и некоторых слов-модификаторов, работает лучше, чем метод опорных векторов [КК 2012b]. В работе [ПС 2011] подробно описываются механизмы правилового метода. Заметим, что авторы отмечают необходимость использования тональных словарей практически при любом методе решения задачи анализа тональности.

1.4 Составление словарей тональной лексики

Существует несколько различных способов классификации тональной лексики, в зависимости от конкретной задачи. Например, в работе [ПС 2011] ставилась цель разметки новостей, поэтому авторы уделяют большое внимания коллокациям с глаголами и выделяют целых 9 различных типов

глаголов. В более общем виде, в качестве тональной лексики можно использовать:

- Списки слов, явно выражающие позитивную или негативную оценку (оценочные прилагательные и существительные типа “умница”)
- Интенсификаторы (например, “очень”)
- Инверторы (слова, инвертирующие оценку – “не”, “без” и другие)
- Амбивалентную лексику (оценочные лексемы, ориентация которых зависит от контекста)
- Лексику, обладающую положительной или отрицательной коннотацией (например, «коррупция»)

Для конкретного исследования такие списки можно составлять вручную, брать уже готовые или составлять их полуавтоматически. Составление списков вручную самый трудозатратный, но при этом наиболее эффективный способ, поэтому он часто применяется на практике, особенно при работе с русскоязычными текстами, потому что готовых тональных словарей для русского языка в открытом доступе пока что не существует. Есть списки тональной, но не размеченной по оценке лексики с того же семинара РОМИП [CL 2012], есть существительные, с размеченной положительной или негативной коннотацией из “Семантического словаря русского языка” под редакцией Н.Ю. Шведовой, есть несколько размеченных тонально слов в Национальном Корпусе Русского Языка, но единого открытого ресурса нет. К полуавтоматическим методам можно отнести машинное обучение, выявление устойчивых шаблонов, бутстрэпинг и другие.

1.5 Контекстно-независимый словарь тональных прилагательных английского языка SentiWordNet

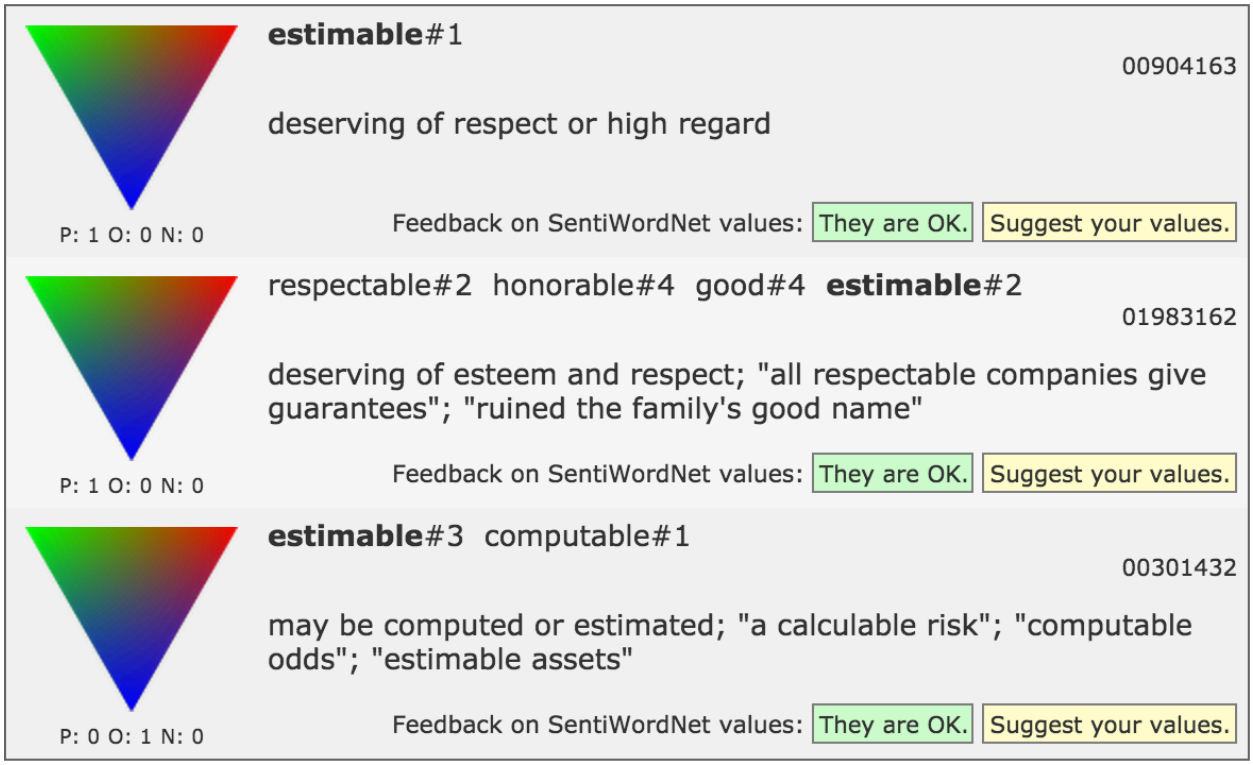


Рисунок 1 Пример выдачи SentiWordNet для прилагательного estimable

Для английского языка была разработана открытая база данных тональных прилагательных на основе связей понятий (*synset*) из сети WordNet. Основной алгоритм реализации описан в работе [ES 2006], особенности новой версии, вышедшей несколько лет спустя описаны в работе [BES 2010]. Для каждого значения эмоционально-окрашенного прилагательного приписывается оценка от нуля до единицы по трем шкалам: Pos (позитивность), Neg (негативность) и Obj (объективность), так что сумма всех трех оценок равна единице. Полученная оценка визуализирована в виде треугольника с разноцветными вершинами. На рисунке 1 приводится пример выдачи SentiWordNet по прилагательному “estimable”. Как мы видим, в зависимости от значения прилагательному может приписываться как полностью объективная оценка, так и чисто положительная. На самом деле, оценка приписывается целому синсету, в котором оказываются сразу несколько прилагательных со сходным значением. Подход классификации синсэтов позволяет эффективно разрешить проблему дизамбигуации по смыслу. Чтобы использовать такой словарь на реальных текстах,

дизамбигуировать значения прилагательных придется самостоятельно. Тем не менее, далеко не все прилагательные обладают большим разбросом тональной оценки в зависимости от смысла, поэтому такой словарь может оказаться эффективным инструментом при решении задач анализа тональности. Поэтому SentiWordNet активно применяется в работе с англоязычными текстами, о чем говорит цитируемость работы [ES 2006] – более 1380 упоминаний.

Техника создания англоязычного ресурса, описанная в работе [ES 2006] основывается на *semi-supervised learning* – фактически бутсрэпинге по связям синсэтов в ворднэте. В начале берется небольшое количество семян (*seeds*) прилагательных с явно выраженной позитивной или негативной оценкой. Затем эти множества пополняются засчет добавления синонимичных синсэтов той же категории или антонимичным синсэтов из противоположной категории. Оставшиеся синсэты попадают в класс Obj. На полученной выборке методами опорных векторов и Наивным Байесом обучаются три классификатора, которые определяют позитив, негатив и объективность соответственно. Затем каждому классификатору с наилучшими параметрами предлагается разметить не попавшие в обучающую выборку синсэты. Оценка синсэту выставляется как взвешенное оценок всех трех классификаторов. В последней версии SentiWordNet алгоритм значительно усложнен для повышения качества разметки.

В рамках научно-исследовательской группы «Сентинет» на факультете гуманитарных наук ВШЭ планировалось сделать русскоязычный аналог SentiWordNet'a. К сожалению, нам не удалось повторить опыт зарубежных коллег в связи с тем, что русский ворднэт сейчас находится на ранних этапах развития и прилагательные в нем окажутся еще не скоро. Тем не менее, нашей учебной группе в составе Анастасии Данейко, Елены Кулаковой и Екатерины Лыточкиной удалось сделать первые шаги по созданию собственного ресурса, который, как мы надеемся, в скором времени станет

доступен для шикорого применения. О создании нашего “Сентинета” и будет рассказано дальше. Пользуясь случаем, хотелось бы выразить благодарность коллегам за сотрудничество.

Часть 2. Составление списка пар слов

2. 1 Первоначальный список и его пополнение

Основная задача проекта «Сентинет» - составить словарь эмоционально-окрашенной лексики для русского языка. Мы начали с небольшого списка тональных прилагательных взятых из примеров в русскоязычных работах по анализу тональности, например из работы [ДА 2013]. Затем на основе “Словаря русской идиоматики” Г.И. Кустовой, доступном на dict.ruslang.ru, были выбраны прилагательные, сочетающиеся с магнификантами, то есть словами со значением высокой степени. Как и следовало ожидать, большинство полученных таким образом прилагательных оказались эмоционально-окрашенными. В итоге в первоначальном списке оказалось 128 прилагательных.

2. 2 Разметка первоначального списка

Разметка первоначального списка производилась двенадцатью экспертами, которым было предложено проставить следующие оценки: «-1» для отрицательно ориентированных слов, «+1» для положительно ориентированных и «0» для амбивалентных прилагательных. За счет большого числа разметчиков нам сразу же удалось приписать тональную оценку прилагательным из списка как среднее арифметическое всех оценок. Ручная разметка выявила следующее: несмотря на наличие слов, которые были единогласно признаны всеми экспертами как позитивные (например, потрясающий, ладный, приятный, наилучший) и как негативные (например, возмутительный, грубый, жестокий), большинство слов получило не полную оценку. Это можно объяснить тем, что оценивая слово разметчик представляет себе какой-то его контекст или думает о возможных

контекстах. В зависимости от контекста даже сильно положительное слово, (например, «красивый») может менять оценку на противоположную (например, в словосочетании «красивая физиономия»). Более того, само существительное может нести оценку или коннотацию, поэтому оценка пары прилагательное с существительным может никак не следовать из отдельных оценок каждого слова. В зависимости от контекста прилагательное может приобрести самую разную оценку. Например, прилагательное «большой» само по себе является объективным и может порождать безоценочные словосочетания типа «большой палец». Еще оно может выступать магнификантом и усиливать отрицательную тональность существительного, например, «большая проблема». Может в сочетании с нейтральным существительным образовывать явно положительную пару, например, «большое сердце». А может в сочетании с таким же нейтральным существительным образовать пару, меняющую оценку в зависимости от контекста, например «большой экран». Неужели так много разных контекстов должен прокрутить в голове асессор при разметке прилагательного «большой»? И какой же из них он выберет? Чтобы использовать экспертную разметку и получить качественную выдачу, необходимо было дополнить словарь хотя бы минимальным контекстом.

2. 3 Расширение списка оценочных прилагательных

Для расширения словаря оценочных прилагательных было выбрано несколько методов: добавление антонимов из словаря антонимов; анализ встречаемости прилагательных основного списка в словарных статьях толкового словаря, а также анализ толкований прилагательных основного списка.

Использование антонимов для расширения списка оценочных прилагательных основывается на мнении, что антонимы оценочных прилагательных также будут оценочными. Базовый список оценочных

прилагательных количеством 128 лексем был расширен путем добавления антонимов входящих в него прилагательных. Антонимы были взяты из ресурса http://antonymonline.ru , который представляет собой единый список антонимов русского языка, при составлении которого использовался ряд словарей антонимов таких, как «Словарь антонимов русского языка» М. Р. Львова, «Словарь антонимов русского языка» Л. А. Введенской, «Словарь антонимов русского языка: Сложные слова» Н. М. Меркуловой и ряд других. В ходе работы с данным словарем было найдено 196 антонимов для 76 из 128 прилагательных, причем некоторым прилагательным соответствует сразу несколько антонимов. После того, как из объединенного списка прилагательных были удалены «двойники», список прилагательных составил 233 лексемы, т. е. увеличился на 105 прилагательных.

В качестве одного из методов пополнения словаря оценочных прилагательных был использован метод, основанный на встречаемости прилагательных основного списка в толкованиях, для этого был выбран «Толковый словарь русского языка» С.И. Ожегова. Были проанализированы словарные статьи прилагательных основного списка, а также толкования, их содержащие. Данный метод основан на допущении, что словарные статьи оценочных прилагательных будут содержать оценочные прилагательные. В результате проделанной работы было найдено 781 прилагательное, а после исключения дубликатов основной список был пополнен 469 прилагательными.

В результате список оценочных прилагательных пополнился на 574 прилагательных и на данный момент составляет 702 прилагательных.

2. 4 Поиск коллокаций по корпусу Leeds

В качестве решения проблемы зависимости от контекста было принято решение собрать не только тональные прилагательные, а словосочетания тональное прилагательное с существительным и просить разметчиков

присваивать оценку паре. Для существительных было решено сделать семантическую классификацию наподобие той, что уже существует в Национальном Корпусе Русского Языка. Таким образом оценку прилагательного можно получить из присвоенных ему оценок в сочетании с разными классами существительных.

Для сбора коллокаций был использован ресурс Corpus Leeds: http://corpus.leeds.ac.uk/ruscorpora.html. Для каждого прилагательного были найдены все коллокации с существительными из текстов Национального Корпуса Русского Языка, в итоговый список вошли только пары с наибольшим показателем loglikelihood score (LL score). Затем список был просмотрен вручную и оттуда были удалены явно ошибочные пары. В спорных случаях пара оставлялась в списке. Таким образом был получен список из более чем 7000 пар вида тональное прилагательное с существительным. Давать такой большой список на разметку асессорам было бы слишком трудоемко, к тому же материал для разметки довольно монотонный, что может плохо сказаться на качестве. Поэтому было принято решение придумать и реализовать простую краудсорсинговую игру, чтобы как можно больше носителей языка смогло побыть нашими разметчиками. Даже если каждый разметит небольшое количество слов, за счет большого количества игроков нам удастся покрыть все пары и собрать статистику ответов. В следующей части будет полностью описан процесс создания игры: от выбора модели до технической реализации.

Часть 3. Создание игры и анализ результатов

3. 1 Ограничения и возможные модели реализации

При выборе игровой модели нужно было учитывать следующие аспекты. Во-первых, игра должна полностью отвечать постановке задачи. То есть в концепт игры должна вписываться пара слов тональное прилагательное с существительным и должна быть возможность эту пару оценить. Во-вторых, игра должна быть достаточно простой и интересной, у пользователя должна быть какая-то мотивация разметить как можно больше пар. В третьих, игра не должна сказываться на результатах разметки. Например, в условиях ограничения по времени качество разметки могло бы значительно снизиться. В-четвертых, игра должна мотивировать пользователя размечать правильно и думать о том, какую оценку присвоить. Причем контролировать это мы никак не можем, пока не соберем достаточно данных для записи «правильных ответов».

Какими способами можно поддержать интерес игрока? Ответ на этот вопрос можно найти проанализировав уже существующие игры. Обычно это продуманная система уровней, условия ограничения по времени, система подсчета очков, система наград, набор «жизней». При имеющемся у нас материале из всего этого разнообразия можно использовать только систему подсчета очков и достижений, так как мы не можем ограничивать пользователя по времени и отбирать жизни или делать уровни, потому что никак не можем проверить насколько грамотно пользователь размечает пары. Зато система поощрений как раз работает так, как нужно – чем больше пар пользователь разметил, тем больше он молодец.

3. 2 Визуальное решение

Нами было рассмотрено несколько моделей реализации, вот лучшие из них:

- Система трех ведер: сверху падает словосочетание, которое пользователь должен отправить в одну из трех корзинок – позитивную, негативную, нейтральную
- Съедобное-несъедобное: на экране появляются пары, которые нужно откинуть влево, если это негатив, и вправо, если позитив

По сути эти модели очень похожи, только вторая скорее подходит для приложения на телефон, а первую легче реализовать как браузерную игру. Фактически, это даже и не игра, а скорее просто сбор данных, когда пользователю дается пара, которой нужно поставить оценку, только в облегченном варианте.

3. 3 Первая реализация

В самом первом варианте игра представляла собой код, написанный на языке Python с использованием библиотеки Flask для сбора веб-страниц. На вход подавался csv-файл с неразмеченными парами, откуда для пользователя случайным образом выбиралась текущая и показывалась на экране, под парами располагались три кнопки: “positive”, “negative” и «?». При нажатии пользователем одно из кнопок в файл выдачи добавлялась строка с user_id пользователя, завязанного на сессии, текущей парой и ответом. На отдельной страничке можно было прочитать краткое описание и оставить свой отзыв, который тоже записывался в файл. Эта версия была протестирована в узком кругу и тщательно доработана. В качестве фонового изображения был выбран рисунок медведя, лежащего на льдине, который находил

положительный отклик у асессоров. Поэтому к следующей версии было решено добавить больше приятных картинок и милых мелочей. На рисунке 2 представлен скриншот самого первого прототипа игры.

Рисунок 2 Первый прототип игры

3. 4 Устройство игры и формат фидбэка

В доработанной версии мы заменили кнопки “positive” и “negative” на изображения доброго и злого медведя, а вместо кнопки «?» появилась новая кнопка с нейтральным медведем и корзина, куда пользователь мог отправить пару, которую посчитал бы некорректной или нехарактерной для русского языка. В дальнейшем, при автоматическом расширении списка пар мы сможем выкидывать некорректные словосочетания, как те, которые были отправлены в корзину несколькими пользователями.

Так же появилась кнопочка «назад», позволяющая вернуться к предыдущей паре. Игра обзавелась страничкой с правилами, был доработан

раздел «о Сентинете», появилась секретная страничка для выгрузки фидбэка: комментариев и размеченных пар.

Немного изменился формат фидбэка: теперь в выдачу записывается время, user_id, текущая пара и оценка. Если пользователь возвращался к предыдущей паре, новая строка ответа пишется следующей, таким образом мы не теряем никакой информации и можем узнать, сомневался ли пользователь при разметке пары.

За каждую размеченную пару увеличивается счетчик рыбок. На определенных значениях счетчика, выбранных случайным образом, появляется милая картинка с мотивирующей надписью – таким образом реализована система награждений.

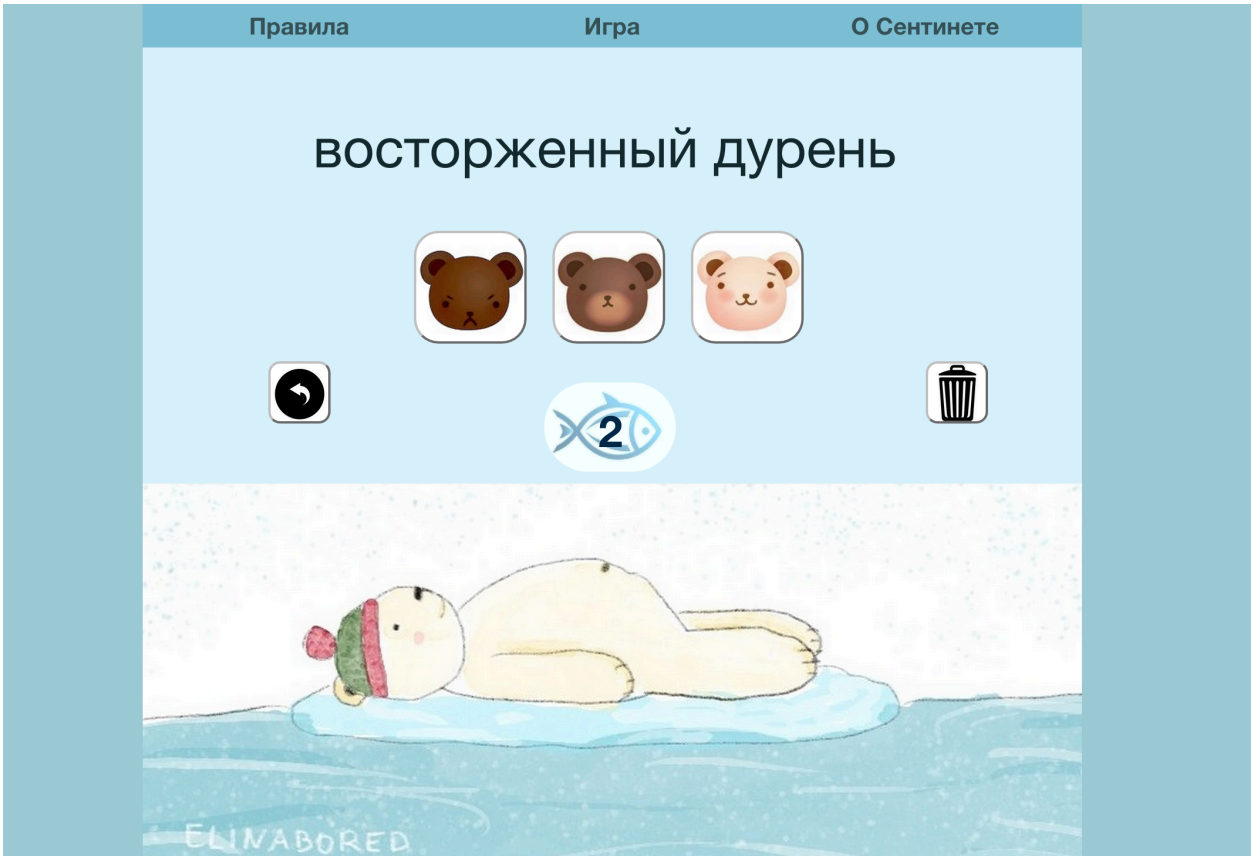
3. 5 Связь с базой данных

Фидбэк игры после небольшой обработки попадает в базу данных, где хранятся пары, прилагательные и существительные с классами. Сейчас в базе данных работает поиск по прилагательному – когда можно посмотреть, какие оценки это прилагательное получило в словосочетаниях со всеми существительными. В дальнейшем мы хотим добавить в базу данных возможность поиска по прилагательному с заданным классом существительного.

3. 5 Восприятие игры асессорами

В первый день после официального запуска пользователи проставили более 25000 оценок. В целом игра была воспринята положительно, кто-то даже отозвался, что она затягивает. Неожиданным плюсом оказались сами словосочетания, некоторые из них были довольно смешные или необычные, что мотивировало людей играть дальше. При простоте дизайна и реализации, многие очень хорошо отзывались о визуальном решении.

Люди с лингвистическим образованием сразу поняли, что это какая-то разметка, но отнеслись снисходительно и с понимаем. Некоторые не знают, что мы никак не проверяем «правильность» ответов и играют честно. Но есть



в выдаче и явные «спамеры», которые просто жмут одну кнопку очень

Рисунок 3 Последняя рабочая версия игры

быстро. Таких пользователей мы можем вычислить по времени и user_id в выдаче и в дальнейшем нужно будет от «спамеров» выдачу очищать. Так же поступило очень много предложений и замечаний о том, как еще можно доработать игру, которые мы надеемся в будущем учесть.

На рисунке 3 представлен скриншот игры в ее последней версии.

Игра доступна по ссылке: http://web-corpora.net/wsgi/senti_game.wsgi/

Заключение

В данной работе были проанализированы задачи и проблемы области анализа тональностей. Одним из необходимым ресурсов для успешного развития области на русском языке является тональный словарь. Именно созданию эмоционально-окрашенного словаря словосочетаний посвящено данное исследование. Изначально словарь возник из небольшого списка прилагательных и был дополнен методами бутстрэпинга. Чтобы разграничивать различные значения одного прилагательного, по корпусу Leeds был собран список коллокаций с существительными. Для разметки было решено использовать краудсорсинг посредством реализации игры. Простой, но симпатичный дизайн игры оказался очень привлекательным для пользователей.

В дальнейшем, мы надеемся собрать оценки и сделать базу данных общедоступной. Она может стать не только полезным инструментом для анализа тональности текстов, но и представить интересные данные для теоретических исследований.

Список литературы

[BES 2010] Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining //LREC. – 2010. – Т. 10. – С. 2200-2204.

[BS 2015] Bamman D., Smith N. A. Contextualized Sarcasm Detection on Twitter //Ninth International AAAI Conference on Web and Social Media. – 2015.

[CL 2012] Chetviorkin I. I. , Loukachevitch N. V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of COLING 2012: Technical Papers , pages 593–610

[DSR 2010] Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semisupervised recognition of sarcastic sentences in Twitter and Amazon. In CoNLL, 107–116.

[ES 2006] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: a publicly available lexical resource for opinion mining. In Proceeding of the International Conference on Language Resources and Evaluation, pages 417–422, 2006.

[KH 2006] Kim, S.-M. And E. Hovy: 2006, 'Extracting Opinions, Opinion Holders, And Topics Expressed In Online News Media Text'. In: Proceedings Of The Acl/Coling Workshop On Sentiment And Subjectivity In Text. Sydney, Australia.

[PL 2008] Pang B., Lee L. Opinion mining and sentiment analysis //Foundations and trends in information retrieval. – 2008. – Т. 2. – №. 1-2. – С. 1-135.

[WR 2005] Wiebe, J. And E. Riloff: 2005, 'Creating Subjective And Objective Sentence Classifiers From Unannotated Texts'. In: Proceeding Of Cicling-05, International Conference On Intelligent Text Processing And Computational Linguistics, Vol. 3406 Of Lecture Notes In Computer Science. Mexico City, Mx, Pp. 475–486.

[ДА 2013] Дегтева А.В., Азарова И.В. (2013) Структура эмоционально-экспрессивного компонента в тезаурусе русского языка RussNet. По

материалам международной конференции «Диалог-2013»

[КК 2012a] Котельников Е. В., Клековкина М. В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая - 3 июня 2012г.). Вып. 11 (18).- М.: Изд-во РГГУ, 2012.

[КК 2012b] Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики //Труды. – 2012. – С. 118-123.

[ПС 2011] Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»(Бекасово, 25–29 мая 2011 г.). М.: Изд-во РГГУ. – 2011. – №. 10. – С. 17.

Приложение 1. Фрагмент выдачи игры.

03 Jun 2015 22:25:36;139743012;отличный профессионал;pos

03 Jun 2015 22:25:37;4121318936;неповторимый ансамбль;pos

03 Jun 2015 22:25:38;3949448346;плохое снабжение;neg

03 Jun 2015 22:25:38;139743012;нежная благодарность;pos

03 Jun 2015 22:25:38;2884548617;величайший представитель;pos

03 Jun 2015 22:25:39;3191389908;жестокий опыт;neg

03 Jun 2015 22:25:39;3084589522;умопомрачительное па;trash

03 Jun 2015 22:25:39;4121318936;светлый глаз;neut

03 Jun 2015 22:25:39;3949448346;преступная мать;neg

03 Jun 2015 22:25:41;3191389908;превосходное состояние;pos

03 Jun 2015 22:25:41;2884548617;беспощадное истребление;neg

03 Jun 2015 22:25:41;4121318936;правильная ориентировка;pos

03 Jun 2015 22:25:41;3949448346;горькое похмелье;neg

03 Jun 2015 22:25:43;3949448346;ужасное открытие;neg

03 Jun 2015 22:25:43;139743012;совершенный атом;neut

03 Jun 2015 22:25:43;4121318936;сногсшибательное предложение;pos

03 Jun 2015 22:25:43;3191389908;беспросветная злоба;neg

03 Jun 2015 22:25:43;3084589522;злой окрик;neg

03 Jun 2015 22:25:44;3949448346;надежная работа;pos

03 Jun 2015 22:25:44;139743012;удобное местечко;pos

03 Jun 2015 22:25:45;3949448346;добрый приятель;pos

03 Jun 2015 22:25:46;139743012;надежное средство;pos

03 Jun 2015 22:25:46;3191389908;непроходимый дурак;neg

03 Jun 2015 22:25:47;139743012;удобная обувь;pos

03 Jun 2015 22:25:47;3084589522;замечательная пьеса;pos

03 Jun 2015 22:25:47;3949448346;дьявольский труд;neg

03 Jun 2015 22:25:48;139743012;необыкновенный блеск;pos

03 Jun 2015 22:25:48;4121318936;шумный карнавал;neut

03 Jun 2015 22:25:48;3191389908;великий бог;pos

03 Jun 2015 22:25:49;3949448346;лучшее достижение;pos

03 Jun 2015 22:25:49;2884548617;беспощадный приговор;trash

03 Jun 2015 22:25:50;139743012;непроходимый тупица;neg

03 Jun 2015 22:25:50;4121318936;омерзительная подробность;neg

03 Jun 2015 22:25:51;139743012;приятный разговор;pos

03 Jun 2015 22:25:52;2884548617;выдающийся деятель;pos

03 Jun 2015 22:25:52;139743012;дивная вещь;pos

03 Jun 2015 22:25:53;3949448346;добрый кусок;pos

03 Jun 2015 22:25:53;3191389908;яростный гудок;neg

03 Jun 2015 22:25:53;139743012;ладная вещь;pos

03 Jun 2015 22:25:54;4121318936;наглый человек;neg

03 Jun 2015 22:25:55;3191389908;страшный взрыв;neg

03 Jun 2015 22:25:55;3949448346;добрый кусок;trash

03 Jun 2015 22:25:56;139743012;жесточайшее испытание;neg

03 Jun 2015 22:25:56;2884548617;безжалостное время;neg

03 Jun 2015 22:25:56;4121318936;красивые губы;pos

03 Jun 2015 22:25:57;139743012;злостное хулиганство;neg

03 Jun 2015 22:25:59;139743012;замечательный сорт;pos

03 Jun 2015 22:26:00;2884548617;ослепительная вспышка;neut

03 Jun 2015 22:26:00;3949448346;феноменальный случай;neut

03 Jun 2015 22:26:01;3191389908;хороший мужик;pos

03 Jun 2015 22:26:02;4121318936;красивые губы;neut

03 Jun 2015 22:26:02;2884548617;любимая работа;pos

03 Jun 2015 22:26:02;3949448346;феноменальный случай;pos

03 Jun 2015 22:26:03;3191389908;преступная война;pos

03 Jun 2015 22:26:03;139743012;жестокий метод;neg

03 Jun 2015 22:26:04;3949448346;холодный туман;neut

03 Jun 2015 22:26:05;139743012;неприятное зрелище;neg

03 Jun 2015 22:26:06;3949448346;холодный голос;neg

03 Jun 2015 22:26:06;139743012;добрый пастырь;pos

03 Jun 2015 22:26:07;2884548617;восторженный адепт;trash

03 Jun 2015 22:26:08;3949448346;неповторимый день;pos

03 Jun 2015 22:26:09;139743012;тяжкое раздумье;neg

03 Jun 2015 22:26:09;2884548617;великий соотечественник;pos

03 Jun 2015 22:26:10;3949448346;грубый стук;trash

03 Jun 2015 22:26:11;1844275499;неприятное открытие;neg

03 Jun 2015 22:26:12;3949448346;любимая тема;pos

03 Jun 2015 22:26:14;2884548617;беспардонный винегрет;trash

03 Jun 2015 22:26:15;139743012;совершенный Ренуар;neut

03 Jun 2015 22:26:17;139743012;блестящий шелк;neut

03 Jun 2015 22:26:17;3949448346;бессовестный любитель;neg

03 Jun 2015 22:26:18;139743012;добрая охота;pos

03 Jun 2015 22:26:18;3949448346;сладкое яблоко;pos

03 Jun 2015 22:26:19;2884548617;мрачный лес;neut

03 Jun 2015 22:26:20;139743012;великий соотечественник;pos

03 Jun 2015 22:26:21;139743012;неприятная весть;neg

03 Jun 2015 22:26:22;1844275499;мягкий шаг;neut

03 Jun 2015 22:26:23;3949448346;сладкое яблоко;neut

03 Jun 2015 22:26:24;2884548617;блестящий диск;trash

03 Jun 2015 22:26:24;139743012;яростный крик;neg

03 Jun 2015 22:26:24;3949448346;добрая надежда;pos

03 Jun 2015 22:26:25;1844275499;чрезвычайный успех;pos

03 Jun 2015 22:26:26;2884548617;всесильный Вяхирев;trash

03 Jun 2015 22:26:28;3949448346;лучшее доказательство;pos

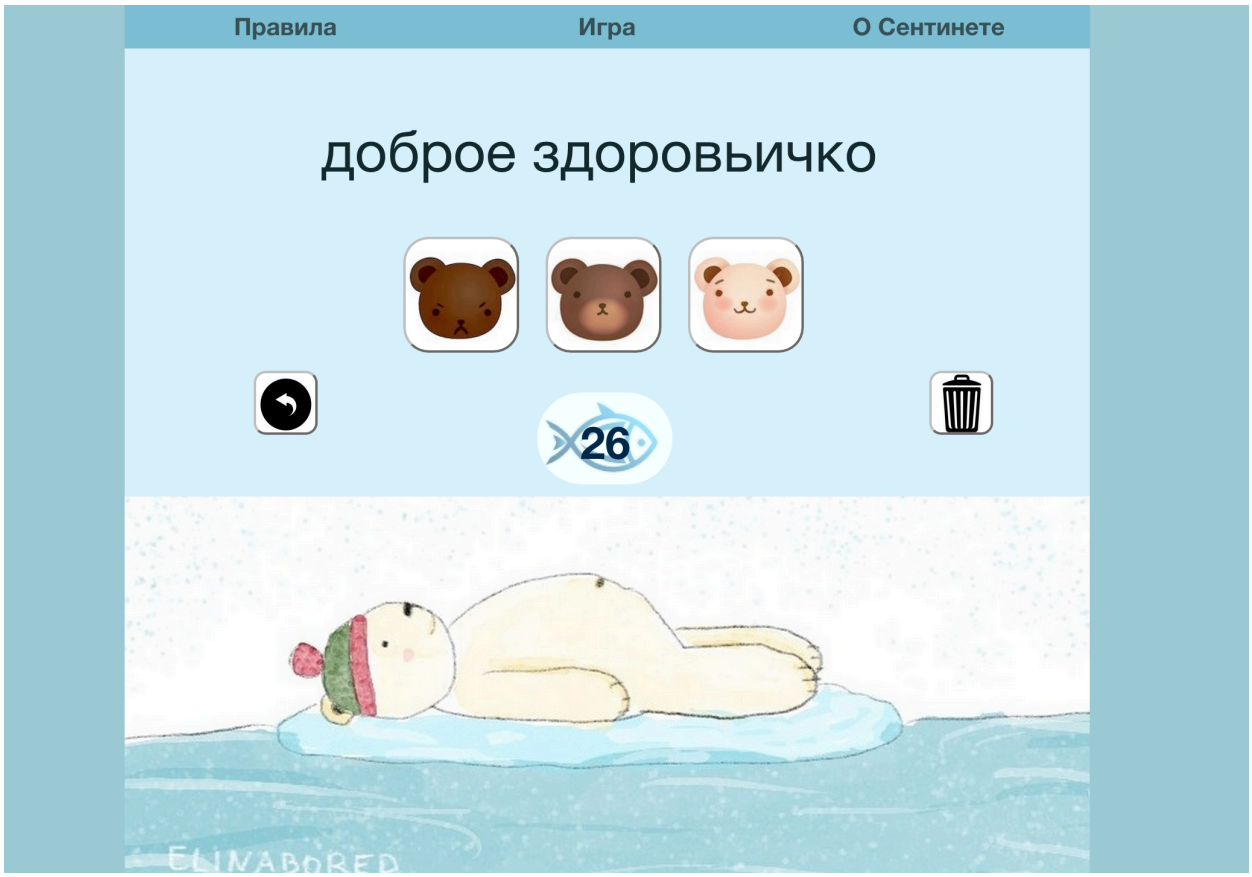
03 Jun 2015 22:26:29;139743012;беспросветная нужда;neg

03 Jun 2015 22:26:29;2884548617;умопомрачительный халат;trash

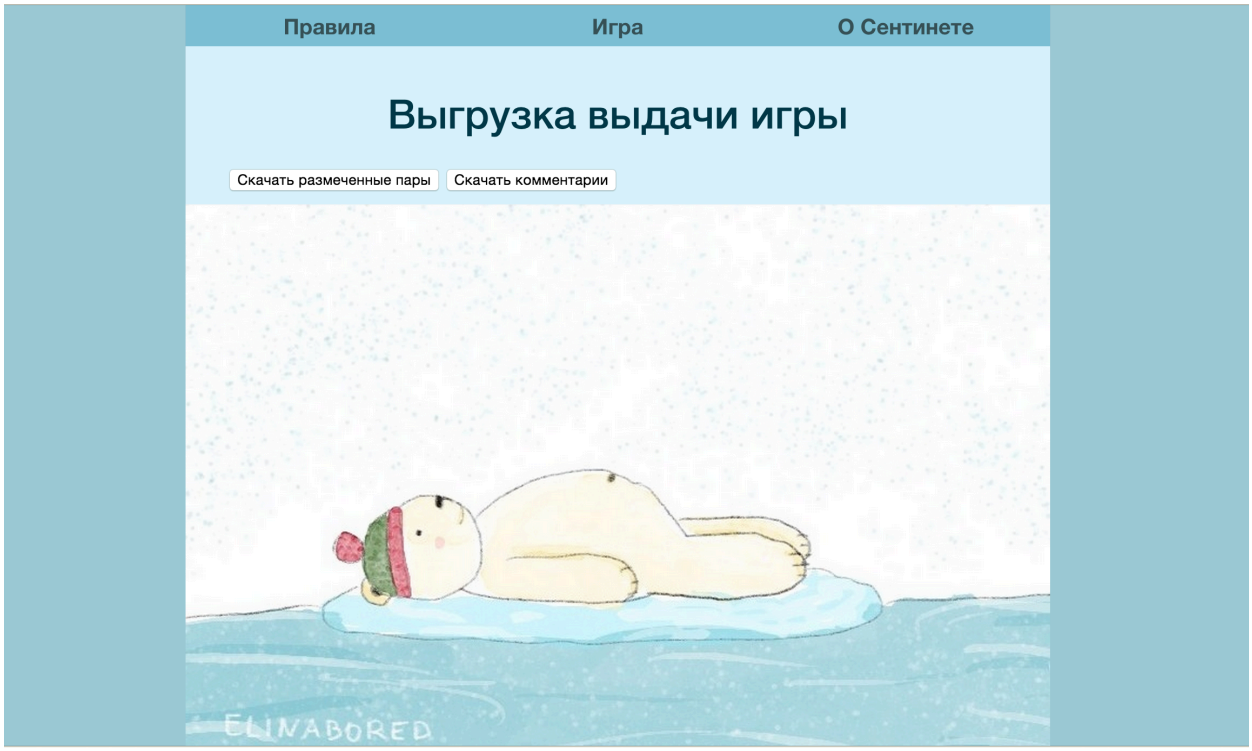
03 Jun 2015 22:26:30;3949448346;тошнотворная детальность;neg
03 Jun 2015 22:26:31;139743012;жесточайшая экономия;neg
03 Jun 2015 22:26:32;3949448346;адский ужас;neg
03 Jun 2015 22:26:33;139743012;грубая шерсть;neut
03 Jun 2015 22:26:35;3949448346;незаурядный успех;pos
03 Jun 2015 22:26:35;1844275499;добрая слободка;trash
03 Jun 2015 22:26:35;2884548617;красивая внешность;pos
03 Jun 2015 22:26:36;4121318936;красивые губы;neut
03 Jun 2015 22:26:37;3949448346;восторженный гам;pos
03 Jun 2015 22:26:38;139743012;блестящая иллюстрация;pos
03 Jun 2015 22:26:38;3949448346;мощный стимул;pos
03 Jun 2015 22:26:39;139743012;тяжкий вред;neg
03 Jun 2015 22:26:41;139743012;фантастический реализм;neut
03 Jun 2015 22:26:42;3949448346;сладкое вино;neut
03 Jun 2015 22:26:42;139743012;хорошая репутация;pos
03 Jun 2015 22:26:43;139743012;отвратительный вкус;neg
03 Jun 2015 22:26:44;139743012;беспросветная дура;neg
03 Jun 2015 22:26:45;3540383610;незаурядный актер;pos
03 Jun 2015 22:26:46;3949448346;надежная оказия;pos
03 Jun 2015 22:26:46;139743012;выдающийся труд;pos
03 Jun 2015 22:26:47;3540383610;неприятное место;neg
03 Jun 2015 22:26:47;139743012;лучшая книга;pos
03 Jun 2015 22:26:47;3949448346;омерзительные духи;neg
03 Jun 2015 22:26:49;139743012;мучительный кашель;neg
03 Jun 2015 22:26:49;3949448346;потрясающая новость;pos
03 Jun 2015 22:26:50;139743012;неприятный хруст;neg
03 Jun 2015 22:26:51;4121318936;сердитая Марья;neut
03 Jun 2015 22:26:51;3949448346;отвратительное качество;neg
03 Jun 2015 22:26:52;139743012;неотразимое впечатление;neut

03 Jun 2015 22:26:53;3540383610;несусветная неразбериха;neg
03 Jun 2015 22:26:53;139743012;ужасный случай;neg
03 Jun 2015 22:26:53;3949448346;сердитый прапорщик;neg
03 Jun 2015 22:26:56;1844275499;хороший пинок;neg
03 Jun 2015 22:26:56;3540383610;ослепительный снег;pos
03 Jun 2015 22:26:57;139743012;сногсшибательная дама;pos
03 Jun 2015 22:26:58;139743012;блестящая перспектива;pos
03 Jun 2015 22:26:58;3540383610;хорошее впечатление;pos
03 Jun 2015 22:27:01;3949448346;бессовестный бездельник;neg
03 Jun 2015 22:27:01;1844275499;хороший пинок;pos
03 Jun 2015 22:27:02;139743012;суровый человек;neut
03 Jun 2015 22:27:03;3540383610;правильный ход;neut
03 Jun 2015 22:27:03;139743012;неприятная слабость;neg
03 Jun 2015 22:27:04;3949448346;горькая неудача;neg
03 Jun 2015 22:27:06;3949448346;возмутительный лозунг;neg
03 Jun 2015 22:27:06;4121318936;наглый образ;trash
03 Jun 2015 22:27:07;3949448346;омерзительный персонаж;neg
03 Jun 2015 22:27:09;3949448346;замечательный врач;pos
03 Jun 2015 22:27:10;3540383610;наихудший элемент;neut
03 Jun 2015 22:27:11;3949448346;беспросветное унижение;neg
03 Jun 2015 22:27:12;4121318936;бессовестная фортуна;neg
03 Jun 2015 22:27:12;1844275499;замечательный приз;pos
03 Jun 2015 22:27:13;3949448346;изумительное произведение;pos
03 Jun 2015 22:27:13;3540383610;сногсшибательный туалет;pos
03 Jun 2015 22:27:14;3949448346;замечательная идея;pos
03 Jun 2015 22:27:15;4121318936;необыкновенный падеж;trash
03 Jun 2015 22:27:15;3949448346;приятное впечатление;pos
03 Jun 2015 22:27:17;4121318936;грандиозный процесс;pos
03 Jun 2015 22:27:18;3540383610;сладкие духи;neut

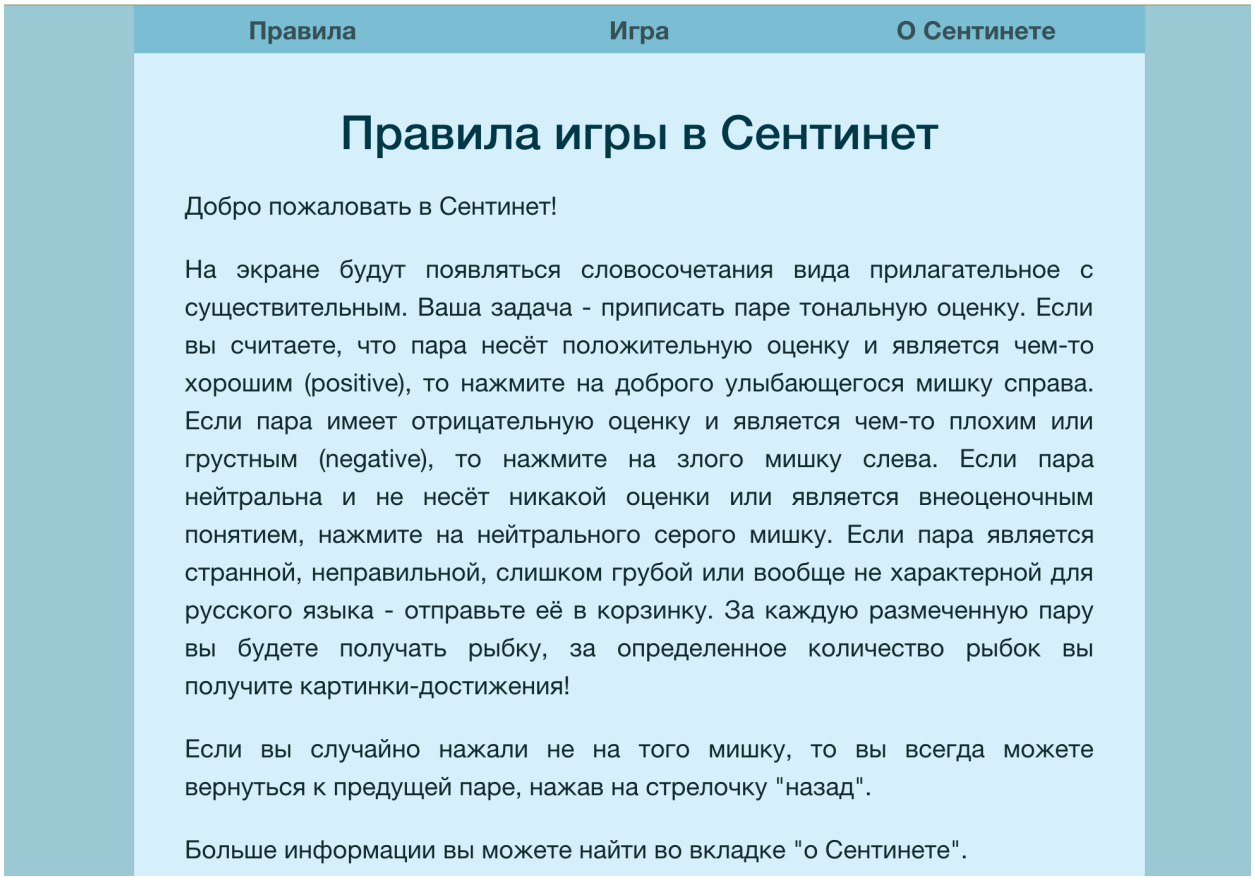
Приложение 2. Скриншоты игры.



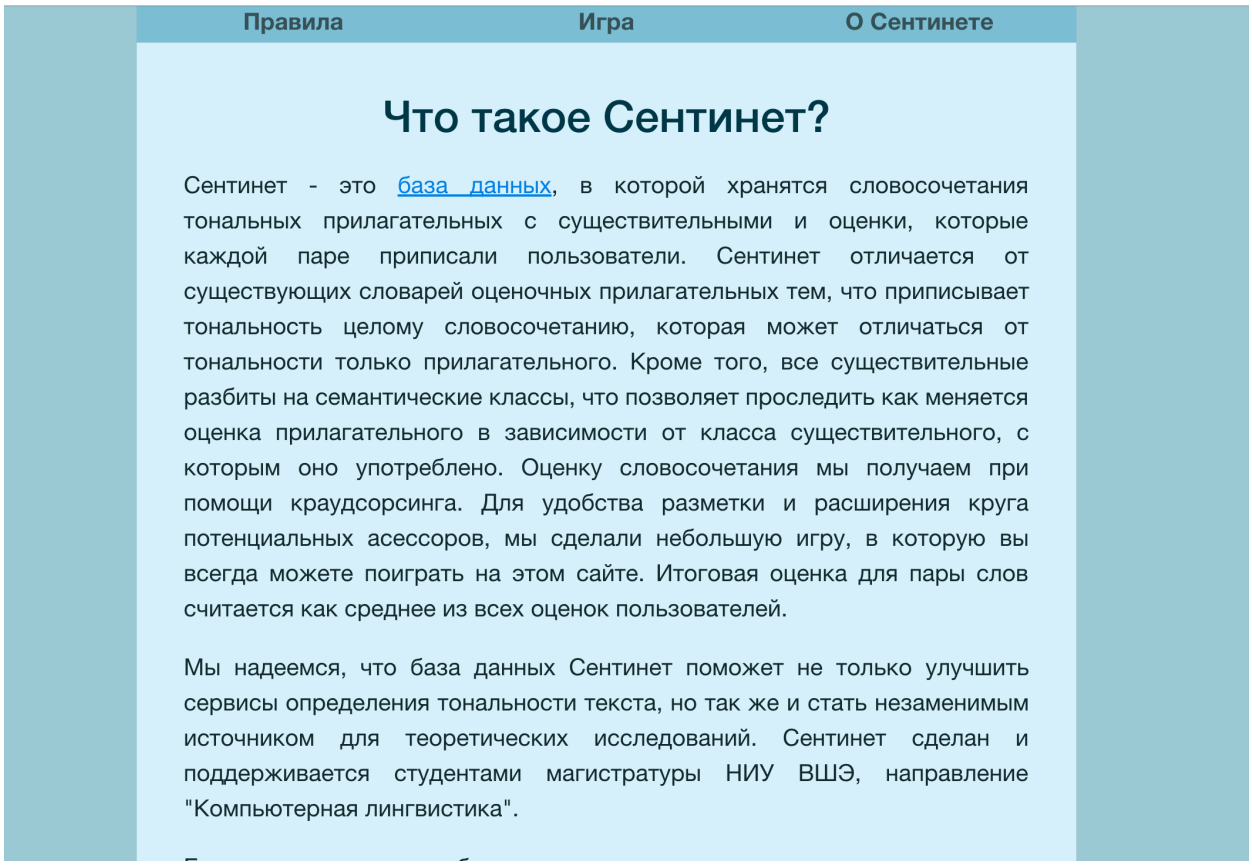
Секретная страничка с выгрузкой результатов



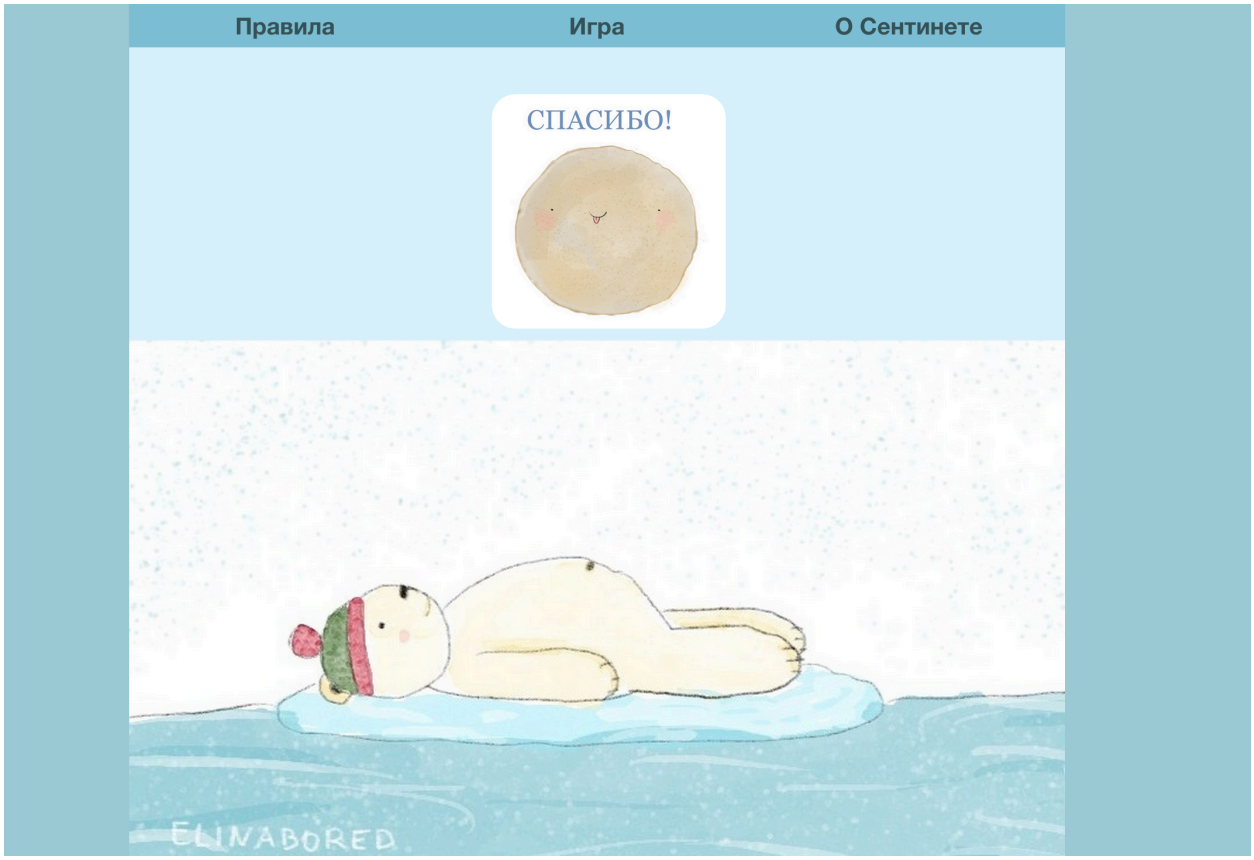
Правила игры



Вкладка «О Сентинете»



За оставленный отзыв блинчик говорит вам «Спасибо!»



Неуловимая «ачивка»