
TIMA

Abschlussarbeit

TIMA

Nathanael Philipp, Felix Rauchfuß, Kai Trott

15. Januar 2016

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Motivation . . . 1

2 TIMA als Citizen Science Projekt **3**
2.1 Citizen Science . . . 3
2.1.1 TIMA im Citizen Science Rahmen . . . 4
2.2 Grundlegende Dateneingabe . . . 4
2.2.1 Wortauswahlalgorithmus . . . 5
2.2.2 ExcludeWords . . . 5
2.2.3 Punkte . . . 6
2.3 Spiele . . . 6
2.3.1 Assoziationskette . . . 6
2.4 Newsletter . . . 7
2.5 Datenqualität . . . 7
2.6 Lizenz . . . 8

3 Assoziationsdatenbank und API **9**
3.1 Backend und Datenbank . . . 9
3.1.1 Datenmodell . . . 9
3.2 API . . . 11
3.2.1 Nicht authentifizierte Anfragen . . . 11
3.2.2 Authentifizierte Anfragen . . . 12
3.2.3 OAI-PMH . . . 14

4 Applikationen und Webseite **15**
4.1 Webseite . . . 15
4.2 Applikationen . . . 15
4.2.1 Bibliothek . . . 15
4.2.2 Aufbau . . . 16
4.3 Sicherheit . . . 17

5 Ausblick **18**

6 Zusammenfassung **20**

7 Literaturverzeichnis **21**

1 Einleitung

In Zeiten von schnellen Prozessoren und riesigen Speichermedien sind mithilfe von Text Mining und automatischer Sprachverarbeitung eine Vielzahl von Datenbanken entstanden, die ganze Sprachen aufgrund von Satzbau, Wortkookurrenzen und Wortarten analysieren und speichern. In dieser Menge der Daten fehlt jedoch eine sehr wichtige Eigenschaft von Worten - die Assoziation. Bisher war es nicht möglich, diese bedeutende menschliche Fähigkeit maschinell zu simulieren. TIMA, rekursives Palindrom für "TIMA is my association", oder auf Deutsch: "TIMA ist meine Assoziation", setzt es sich zum Ziel eine Datenbank zu schaffen, bei der diese Verbindungen zwischen Worten abgerufen werden können. Da wie beschrieben bisher keine automatische Methode dazu existiert, setzt TIMA auf eine Menge freiwilliger Nutzer, die ihre Assoziationen zu Worten eingeben.

Um eine relevante Menge an Daten sammeln zu können, wird TIMA als Citizen Science Projekt aufgezogen. Ein derartiges Projekt hat eine Reihe besonderer Ansprüche und um ihnen gerecht zu werden, sind eine ganze Reihe Vorkehrungen zu treffen, die in dieser Arbeit betrachtet werden sollen.

Zuerst wird näher betrachtet, warum das Erstellen einer Assoziationsdatenbank überhaupt sinnvoll ist und welche Funktionen für Nutzer ansprechend wären. Danach werden einige Gedanken zur Gestaltung als Citizen Science Projekt geäußert. Besonders wird dabei auf den Schutz der gesammelten Daten eingegangen und der wichtigen Frage, wie man Nutzer motiviert, am Projekt teilzunehmen. Nach den technischen Details zur Implementierung einzelner Bestandteile: der Datenbank, der API, der Webseite und einer ersten Applikation, die die Möglichkeiten der API anschaulich demonstriert, folgt ein Abschnitt zum Ausblick auf zukünftige Projekte, die entweder zur Datenbank beitragen können oder sie nutzen.

1.1 Motivation

Die Möglichkeiten einer Assoziationsdatenbank sind vermutlich in erster Linie in Bereichen der automatischen Sprachverarbeitung angesiedelt, dort jedoch beinahe in jedem Teilgebiet nutzbar.

Ein großes Problem aller Automaten ist ihre Reaktivität. Sie sind stets auf Schlüsselbegriffe angewiesen, die der Nutzer eingibt, beziehungsweise spricht. So ist für Suchanfragen jeglicher Art, ob nun Internetsuche, Eingabe in das Navigationsge-

rät oder Sprachbefehle neuartiger Steuerungen für mobile Endgeräte, wie zum Beispiel Siri, sehr schwer die korrekte Reaktion zu liefern, wenn der Nutzer von fest programmierter Terminologie abweicht. Sucht ein Autofahrer statt einer Tankstelle nach Benzin, wird er eventuell keine Antwort bekommen, obwohl einem Menschen intuitiv klar ist, wonach der Fahrer sucht. Eine Maschine kann diese Schlüsse jedoch nicht ziehen und daher müsste man als Programmierer jede einzelne dieser Möglichkeiten bedenken und implementieren. Selbst für ein Navigationsgerät mit relativ eingeschränktem Handlungsspielraum ist dies schon sehr aufwendig, für eine Internetsuchmaschine oder Anwendungen im Bereich des Computational Advertisings jedoch quasi unmöglich. Die Varianz an Suchbegriffen ist einfach zu groß.

Die Problematik einer solchen Datenbank ist jedoch, dass sie sich nicht automatisch erstellen lässt. Schon per Definition ist eine Assoziation eine vom Menschen gezogene Verbindung zwischen zwei Sachverhalten. Über Kookurrenzen lassen sich über Umwege ähnliche Ergebnisse erzielen. Echte Assoziationen, wie sie Menschen ziehen, werden jedoch nur ein Bruchteil der Ergebnisse darstellen. Will man schlechte Ergebnisse vermeiden, ist es unumgänglich, die Assoziationsdatenbank per Hand von Menschen füllen zu lassen. Dass dies auf gewöhnlichem Weg ein sehr großer, auch finanzieller, Aufwand wäre, zeigt sich alleine daran, dass es bisher keine derartige Datenbank gibt, obwohl ein Nutzen, vor allem im Bereich Internetwerbung, nicht von der Hand zu weisen ist. Daher wollen wir hier den Citizen Science Ansatz benutzen, um eine derartige Datenbank zu realisieren.

Ein weiterer Vorteil, das Projekt mit Citizen Science Ansatz zu bearbeiten, bietet die größere Streuung von Assoziationen. Wenn ein einzelner Nutzer eine Assoziation zu einem bestimmten Wort eingeben soll, wird diese sehr oft die gleiche, oder zumindest eine sehr ähnliche sein. Wenn eine große Menge Personen Assoziationen eingibt, wird die Datenbank mit einer größeren Auswahl von Zusammenhängen gefüllt. Stammen diese unterschiedlichen Personen auch noch aus sehr differenzierten Hintergründen, lokal und mit verschiedenen Interessen, so werden die Assoziationen sehr vielfältig. Ein Elektrotechniker wird sicherlich mit dem Begriff Halbleiter etwas anderes verbinden als ein Grundschullehrer. Ein Jugendlicher, der in einer Dorf am Meer groß geworden ist, wird vermutlich einen anderen Bezug zu Fisch haben, als ein Gleichaltriger aus einer Gebirgsstadt.

In den nachfolgenden Kapiteln wird erklärt, wie wir TIMA als Citizen Science Projekt umgesetzt haben.

2 TIMA als Citizen Science Projekt

Im folgenden Kapitel wird der Rahmen von Citizen Science definiert und die Einordnung von TIMA darin. Anschließend erfolgt eine Erläuterung der Grundlagen von TIMA beginnend mit der Dateneingabe. Im weiteren Verlauf des Kapitels, folgt die Sicherung der Datenqualität und die Lizenz.

2.1 Citizen Science

Für Citizen Science gibt es keine klare Definition. Es gibt verschiedene Grenzen die um den Begriff Citizen Science gezogen werden. Unterschieden wird dabei z.B. zwischen Citizen Science und Crowdsourcing, jedoch sind die Übergänge meist fließend.

Als ein Maßstab der zur Abtrennung zwischen Citizen Science und Crowdsourcing benutzt wird M. Haklays Stufen der Beteiligung genutzt. Dabei werden vier Stufen unterschieden. Am unteren Ende der Skala steht dabei das Crowdsourcing auf Stufe eins, bei dem die Teilnehmer als Sensoren oder Freiwillige mitarbeiten. Am oberen Ende, auf Stufe vier, steht das „extreme“ Citizen Science, bei dem sowohl die Problemdefinition, als auch die Datensammlung und -auswertung von den Teilnehmern gemacht wird.[WC11]

Eine weiter Abgrenzung erfolgt über eine Einordnung des Citizen Science Projektes selber in fünf verschiedene Typen:

Aktion (Action) bezeichnet meist das Involieren von Teilnehmern in Lokale Anliegen, die nicht von Wissenschaftlern organisiert sind, aber oft eine beratende Funktion dabei einnehmen. Es handelt sich in der Regel um langfristige Bewegungen mit gesellschaftlichen Zielen.

Erhaltung (Conservation) umfasst das Verwalten mit eindeutigen Lehrzielen. Diese sind meist jedoch nicht in einem leicht zugänglichen Format und stark abhängig von staatlichen Mitteln.

Die Untersuchung (Investigation) beschäftigt sich mit dem Sammeln von Daten die für ein spezifisches Forschungsziel von Nöten sind. Die Projekte müssen in wissenschaftlichen Rahmenbedingungen geführt werden.

Virtuell (Virtual) ist mit dem vorherigen Punkt „Untersuchung“ zu vergleichen. Der Unterschied besteht darin, das keine physikalische Datensammlung notwendig ist.

Lehre (Education) hat zum Hauptziel, die Verbreitung von Wissen. Gültige Resultate haben hierbei weniger Wert, als der Lernprozess an sich.

Für uns muss ein Projekt, damit es als Citizen Science angesehen werden kann, zu mindestens einem der fünf Typen zugeordnet werden können. Des weiteren muss es in M. Haklays Stufen der Beteiligung mindestens auf Stufe zwei einsortiert werden können, da bei Stufe eins die Teilnehmer nicht ausreichend konstruktiv am Problem mitarbeiten.

2.1.1 TIMA im Citizen Science Rahmen

In einem Citizen Science Projekt sind weitere Problemstellungen zu beachten, als in einem gewöhnlichen Informatikprojekt. Die Nutzer sind keine bezahlten oder interessierten Angestellten, sondern gehören der breite Masse an. So nutzen wir die Intelligenz der Masse, die wichtigsten Assoziationen zu einem bestimmten Wort zu finden. Eine sinnvolle Assoziationsdatenbank lässt sich nur mit mithilfe einer großen Anzahl Nutzer erstellen.

In einem Citizen Science Projekt muss die **Datenqualität** überprüft und gesichert werden. Da fast jeder Nutzer ein Laie ist, gibt er eventuell nicht immer ideale Werte ein. Dadurch, dass Nutzer freie Werte eintragen können, ist das Projekt auch gegen Vandalismus besonders anfällig. Zuletzt muss bedacht werden, das die Nutzer in keiner Form bezahlt werden. Daher muss sich TIMA damit auseinander setzen, wie es Menschen die nötige **Motivation** liefert, etwas zum Projekt beizutragen. Dabei fokussieren wir uns auf Spiele, um einen Anreiz zum Mitwirken zu geben.

Wir ordnen TIMA dem Typ „Virtuell“ zu und setzen es auf Stufe zwei von M. Haklays Stufen der Beteiligung.

2.2 Grundlegende Dateneingabe

Die Dateneingabe erfolgt bei TIMA im einfachsten Fall entweder über die Webseite oder die bereits entstanden Applikationen. Dabei werden nacheinander, dem Nutzer Wörter gegeben, zu denen er eine Assoziation eingeben soll. Der dabei gewählte Algorithmus zur Wortauswahl wird im nächsten Abschnitt beschrieben.

2.2.1 Wortauswahlalgorithmus

In diesem Abschnitt wird die Funktionsweise des Wortauswahlalgorithmus genauer beschrieben. In der Datenbank ist es wünschenswert, dass jedes Wort mindestens eine Assoziation hat und zusätzlich ein ungefähres Gleichgewicht an Assoziationen verteilt zwischen allen Worten herrscht. Der Wortauswahlalgorithmus ist zuständig für das Auswählen der Worte, für die eine Assoziation gegeben werden soll, damit diese Eigenschaften erreicht werden.

Daraus ergeben sich die folgenden Anforderungen an den Algorithmus:

- Jeder Nutzer soll zu möglichst vielen verschiedenen Worten eine Assoziation geben.
- Wörter, die wenig assoziiert wurden, entweder insgesamt oder von einem einzelnen Nutzer, sollten für diesen Nutzer bevorzugt werden.
- Wörter sollen ausgeschlossen werden können, um z.B. zu verhindern, dass das selbe Wort mehrmals hintereinander kommt.

Der Algorithmus, der in TIMA umgesetzt ist und diese Anforderungen erfüllt, ist in Listing 2.1 aufgeführt. Zunächst werden alle Wörter der entsprechenden Sprache ausgewählt, davon werden anschließend alle Wörter, die nicht vorkommen sollen gefiltert. Danach wird in Abhängigkeit vom Status des Nutzers, die Liste der möglichen Wörter sortiert und auf 15 Wörter beschränkt, aus denen zufällig ein Wort gewählt wird.

```
1 w = [alle Worter einer Sprache]
2 w = w - [alle auszuschließenden Wörter]
3 if ( Annonymer Nutzer )
4   w = [15 Wörter mit niedrigstem Häufigkeit aus w]
5 else
6   w = w - [alle auszuschließenden Wörter des Nutzers]
7   w = [15 Wörter mit niedrigstem Auftreten in der
       AssociationHistory des Buntzers1 aus w]
8 return [zufälliges Wort aus w]
```

Listing 2.1: Wortauswahlalgorithmus

2.2.2 ExcludeWords

Falls einem Nutzer zu einem Wort keine Assoziation einfällt, hat er die Möglichkeit das Wort zu überspringen. Damit der Nutzer nicht in kurzer Zeit wiederholt danach

gefragt wird (vgl. Abschnitt 2.2.1), wird das Wort auf eine Ausschlussliste gesetzt. Einträge die älter als 7 Tage sind, werden automatisch von der Liste gelöscht. Danach wird der Nutzer erneut nach seiner Assoziation zu diesem Wort gefragt.

2.2.3 Punkte

Um einen grundlegenden Anreiz für die Mithilfe an TIMA zu geben, wird für jede Assoziation Punkte vergeben. Die Vergabe der Punkte erfolgt dabei nach Formel 2.1. A_w steht dabei für die durchschnittliche Häufigkeit des Wortes, A_a die durchschnittliche Häufigkeit der Assoziation, x für die Häufigkeit des Wortes und y für die Häufigkeit der Assoziation.

$$p = \frac{A_w}{x} * \frac{y}{A_a} * 2 \tag{2.1}$$

2.3 Spiele

Darüber hinaus benutzen wir einen spielerischen Ansatz für das Eintragen von Assoziationen, um Menschen zur Mitarbeit bei TIMA zu motivieren. Derzeit ist lediglich das Spiel Assoziationskette umgesetzt, welches im nächsten Abschnitt ausführlich erläutert wird. Die Planung für das zweite Spiel Familienduell (vgl. Kapitel 5) ist abgeschlossen, allerdings aus Zeitgründen noch nicht umgesetzt.

2.3.1 Assoziationskette

Dieses Spiel sollte für einen Nutzer deutlich interessanter sein, im Gegensatz zur Standardeingabe. Voraussetzung für dieses Spiel ist jedoch eine bereits gefüllte Assoziationsdatenbank. Das Spiel Assoziationskette funktioniert folgendermaßen:

Die Spieler assoziieren nacheinander auf die jeweilige Assoziation ihres Gegenübers. Dadurch bildet sich eine Kette mehrerer Assoziationen bis zu bestimmten Abbruchbedingungen. Ein mögliches Ende stellt das Assoziieren eines bereits verwendeten Wortes dar und die damit verbundene Schließung eines Kreises in der Kette. Das Spiel ist ebenfalls beendet, sollte einem Spieler keine weitere Assoziation zeitnah einfallen.

Das Spiel Assoziationskette steht derzeit nur als Variante mit einem Computergegner auf der Webseite zur Verfügung. Der Computer wählt dabei seine Assoziation zufällig aus den Assoziationen der Datenbank zu dem gegebenen Wort aus, wobei sichergestellt wird, das er keinen Kreis bildet.

Alle gegebenen Assoziationen des Spielers werden hierbei mit Punkten belohnt und für die Datenbank verwendet, die Punkte berechnen sich dabei wie im Abschnitt

2.2.3 beschrieben. Die Gesamtlänge der Kette wird zusätzlich Bewertet und dem Sieger gutgeschrieben.

2.4 Newsletter

Ein weiteres Feature über das TIMA verfügt, ist der Newsletter. Falls ein Nutzer möchte, kann er auf der Webseite Wörter auswählen, die ihn interessieren und wöchentlich darüber einen Newsletter erhalten, in dem zu jedem Wort die Assoziationen mit deren Häufigkeit enthalten sind. Durch die regelmäßig Erinnerung könnte sich ein Nutzer motiviert fühlen, erneut an einem Spiel für TIMA teilzunehmen. Dies erfordert eine vorherige Eintragung der E-Mail-Adresse, die für die Anmeldung nicht erforderlich ist.

2.5 Datenqualität

Um die Datenqualität in der Datenbank zu sichern, setzen wir auf drei Dinge:

1. Rechtschreibprüfung
2. Mengenrelevanz
3. Nutzermanagement

Rechtschreibung

Die naheliegendste Überprüfung um die Datenqualität zu gewährleisten, ist eine Rechtschreibkontrolle. Dabei wird nur auf Korrektheit eines Wortes geachtet und dem Nutzer werden keine Vorschläge zur Rechtschreibung unterbreitet.

Ein eingegebenes Wort wird zuerst mit der Datenbank abgeglichen, um zu schauen ob das Wort bereits bekannt ist. Wenn dies der Fall ist wird es als richtig angesehen ansonsten wird das Wort mit Wiktionary[1] abgeglichen.

Sollte das Wort nicht gefunden werden wird der Nutzer darauf hingewiesen, damit er ggf. Änderungen vornehmen kann.

Mengenrelevanz

Mit einer Rechtschreibprüfung lässt sich natürlich nicht verhindern, dass ein Nutzer falsche oder irrelevante Daten eingibt. Ob eine Assoziation falsch ist, kann man

[1] https://www.wiktionary.org/

natürlich von außerhalb nicht entscheiden. Jeder Mensch hat eigene Assoziationen zu bestimmten Worten.

Wir betrachten eine Assoziation als relevant, wenn sie im Allgemeinen nachvollziehbar ist. Gibt ein Nutzer eine Assoziation ein, die diesem Anspruch nicht genügt, ob mutwillig oder nicht, kann man davon ausgehen, dass andere Menschen diese Assoziation nicht geben. In der Datenbank wird eine irrelevante Assoziation also statistisch vernachlässigbar.

Nutzermanagement

Ein weiteres mächtiges Werkzeug zur Vermeidung von Vandalismus ist die Nutzerkontrolle. In der Datenbank werden alle Assoziationen gespeichert, die ein Nutzer jemals gegeben hat. Sollte ein Nutzer auffällig werden, zum Beispiel, weil er innerhalb sehr kurzer Zeit sehr viele Punkte in Spielen erreicht, kann er überprüft werden. Sollte bei einer solchen Überprüfung auffallen, dass seine Assoziationen sehr oft nicht dem allgemeinen Verständnis entsprechen, können diese Eintragungen aus der Datenbank gelöscht werden. Dieser Vorgang ist momentan nicht automatisiert.

2.6 Lizenz

Wir haben uns dazu entschieden, die Wörter und ihre Assoziationen unter die Creative Commons Attribution 4.0 International[2] Lizenz zu stellen, damit jeder die Möglichkeit hat die Daten frei zu verwenden.

Den Quellcode hingegen haben wir unter eine angepasste LGPLv3[3] gesetzt. Zusätzlich wurden Einschränkungen gemacht, dass TIMA nicht in Bereichen eingesetzt werden darf, bei denen Menschenleben in irgendeiner Weise in Gefahr gebracht werden können.

[2] http://creativecommons.org/licenses/by/4.0/

[3] https://github.com/Tima-Is-My-Association/TIMA/blob/master/LICENSE

3 Assoziationsdatenbank und API

Das Ziel von TIMA ist das Erstellen einer Datenbank, in denen Assoziation gespeichert werden. Daher liegt ein Schwerpunkt unserer Arbeit darin, diese zu Erstellen und zu Befüllen. Die Datenbank ist direkt verknüpft mit einem Webfrontend, das durch die Bereitstellung einer umfassenden API der Hauptanlaufpunkt für die Nutzer und die Applikationen ist.

In diesem Kapitel werden das Backend der Webseite und die Datenbank näher beschrieben. Dabei wird genauer auf Designentscheidungen eingegangen, die zum Aufbau der einzelnen Datenbankbestandteile geführt haben.

3.1 Backend und Datenbank

Für das Backend der Website haben wir uns für Django als grundlegende Bibliothek entschieden. Bei Django handelt es sich um ein in Python geschriebenes Webframework, das dem Model-View-Controller-Schema folgt. Django bietet unter anderem einen sehr komplexen objektrelationalen Mapper, der es ermöglicht auch komplexe Objektstrukturen abzubilden ohne die verwendete Datenbank explizit zu kennen. Neben allen notwendigen Funktionen gewährleistet Django zusätzlich also gute Wiederverwendbarkeit und wurde deshalb für das Erstellen des Backends genutzt.

3.1.1 Datenmodell

In Abbildung 3.1 ist das komplette Datenmodell von TIMA dargestellt. Das Modul `associations.models` spielt dabei die Schlüsselrolle. Hier werden die grundlegenden Daten für die Assoziationsdatenbank gespeichert: die Worte und deren Verknüpfungen.

Das Modell `Word` speichert einzelne Wörter und das Modell `Association` die Assoziationen zwischen diesen Wörtern. Für jedes Wort wird gespeichert, wie oft für dieses nach einer Assoziation gefragt wurde. In ähnlicher Weise besitzt auch jede Assoziation in der Datenbank eine Häufigkeit, die angibt wie oft die Assoziation von Nutzern gegeben wurde.

Um eine Unterscheidung zwischen verschiedenen Sprachen zu ermöglichen, repräsentiert das Modell `Language` die verfügbaren Sprachen. Existiert ein Wort in

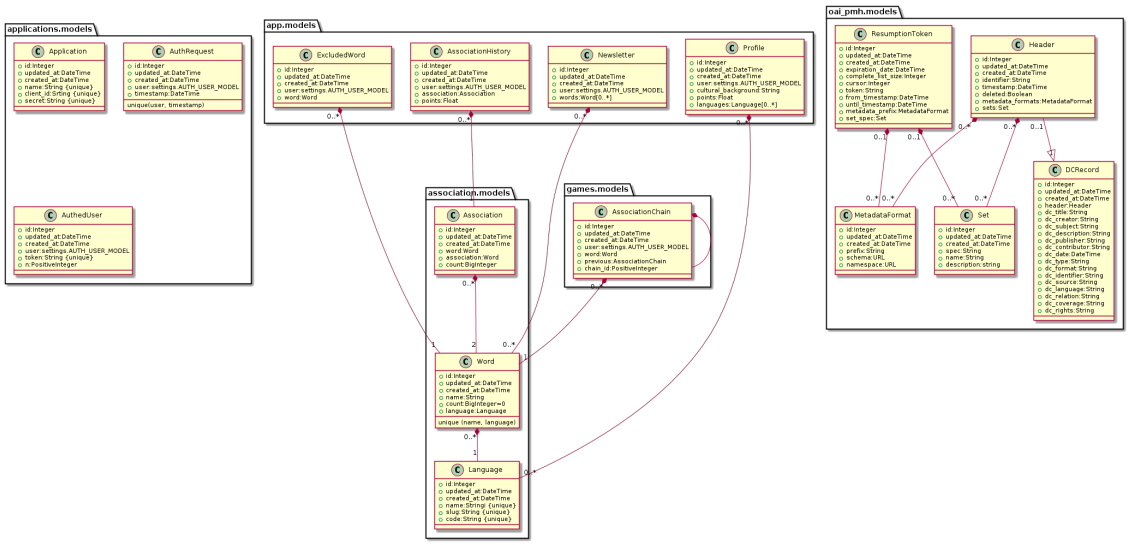

Abbildung 3.1: UML des TIMA Datenmodells

mehreren Sprachen oder wird in verschiedenen Sprachen genutzt, hat es für jede Sprache einen eigenen Eintrag.

Das Modul `games.models` enthält Modelle die für die verschiedenen Spiele wichtig sind. Dies ist im Moment nur das Spiel AssoziationsKette (vgl. Abschnitt 2.3), hierfür werden in dem Modell `AssociationChain` die letzte beziehungsweise aktuelle Assoziationskette eines Nutzers gespeichert. Diese wird beim Start eines neuen Spieles gelöscht.

Um grundlegende Funktionen des Nutzermanagements zu ermöglichen, wurden das Modul `app.models` eingeführt. Das Modell `Profile` speichert grundlegende Informationen zu jedem Nutzer, zum Beispiel die Punktzahl und die Sprachen, für die ein Nutzer assoziiert hat. Diese Daten werden in einem Ranglistensystem genutzt, das die Nutzer motivieren soll, sich gegenseitig zu messen. In dem Modell `AssociationHistory` wird die gesamte Assoziationsgeschichte eines Nutzers gespeichert, mit den jeweils für eine Assoziation erhaltenen Punkte. Somit können im Falle eines Missbrauchs die gegebenen Assoziationen aus der Datenbank gelöscht werden und dem Nutzer die Punkte entzogen werden. Das Modell `ExcludeWord` enthält für jeden Nutzer die Wörter, die er innerhalb der letzten sieben Tage übersprungen hat (vgl. Abschnitt 2.2.2). Das letzte Modell in diesem Modul speichert für jeden Nutzer welche Worte er in seinem Newsletter empfangen möchte.

Für die Kommunikation zwischen der Applikation und dem Backend, insbesondere der Authentifizierung der Schreibzugriffe auf die Datenbank (vgl. Abschnitt 3.2) dient das Modul `applications.models`. Das Modell `Applikation` speichert die Applikationen, mit denen es möglich ist sich zu authentifiziert, mit den nötigen Daten für die Authentifizierung (vgl. Abschnitt 3.2.2). Die beiden anderen Modelle dieses Moduls `AuthRequest` und `AuthedUser` speichern die nötigen Information für einen Nutzer der sich authentifizieren möchte oder sich bereits authentifiziert hat. Durch dieses Modul wird also gewährleistet, dass nur Nutzer auf die Datenbank zugreifen können, die eine von uns erlaubte Anwendung nutzen.

Das letzte Modul und die beinhalteten Modell sind für OAI-PMH erforderlich, näheres dazu in Abschnitt 3.2.3.

3.2 API

Damit verschiedene Applikationen mit der TIMA Datenbank kommunizieren können, haben wir uns entschieden eine umfangreiche API zu implementieren. Diese lässt sich grob in drei Teile gliedern. Erstens gibt es die Anfragen, die keiner Authentifizierung bedürfen, zweitens jene die eine Authentifizierung erfordern und drittens eine OAI-PMH Schnittstelle.

Bis auf die OAI-PMH Schnittstelle, die XML zurück gibt, werden immer JSON-Objekte zurückgegeben. Jedes dieser JSON-Objekte enthält immer den aktuellen Zeitstempel.

Eine vollständige Dokumentation der API ist im Git-Repository in der Datei API.md[4] zu finden.

Im folgenden Abschnitt werden die einzelnen API Anfragen erläutert, zuerst die keiner Authentifizierung benötigen, dann die die eine erfordern und zum Schluss wird dann noch ein Abschnitt zu OAI-PMH folgen.

3.2.1 Nicht authentifizierte Anfragen

Die API-Anfragen, die keine Authentifizierung bedürfen sind allgemeine Anfragen, an die Assoziationsdatenbank, die auch über die Webseite ohne eine Anmeldung erfolgen können.

Rangliste Die Antwort enthält eine Liste der Nutzer, mit den gleichen Daten wie sie über die Webseite einsehbar sind, also Rank, Benutzernamen, Punkte, Sprachen und kultureller Hintergrund.

[4] `https://github.com/Tima-Is-My-Association/TIMA/blob/master/API.md`

Statistik Es sind die aktuelle Daten über Nutzerzahl, Wortmenge und Assoziationen enthalten.

Sprachen Es kann eine Liste aller Sprachen in TIMA angefordert werden, hier ist neben dem Namen, der Sprach-Code in der Antwort enthalten, der bei vielen anderen Anfragen als Parameter angegeben werden muss.

Wörter Um entweder ein einzelnes Wort oder eine Liste von Wörtern abzufragen, ist diese Anfrage bestimmt. Es können optional Wort-IDs, Sprache oder ein Limit für die Anzahl der Assoziationen pro Wort angegeben werden. Das JSON-Objekt der Antwort enthält unter anderem zu jedem Wort einen Link zur Website des Wortes, einen Link zu dieser Anfrage mit der Auswahl auf das einzelne Wort und den OAI-PMH-Identifier des Wortes.

3.2.2 Authentifizierte Anfragen

Da nicht jeder Nutzer Schreibzugriff auf die Datenbank erhalten soll, ist eine Authentifizierung notwendig. Außerdem ist eine Authentisierung notwendig, um Nutzer der Applikation oder der Webseite eindeutig identifizieren zu können. Aus diesem Grund war es erforderlich, dass einige API Anfragen eine Authentifizierung benötigen.

Um dies zu realisieren haben wir uns zunächst bestehende Bibliotheken wie zum Beispiel OAuth2 angeschaut und getestet in wie weit diese unseren Anforderungen genügen. Dies hat allerdings zu keinen zufriedenstellendem Ergebnis geführt, weswegen wir entschieden haben dies selbstständig zu implementieren.

Die grundlegenden Anforderungen die wir dabei hatten sind wie folgt:

1. Sichere Authentisierung einer Applikation
2. Sichere Authentisierung eines Nutzers
3. Sicherstellen, das spätere Anfragen von einem authentifizierten Nutzer kommen

In Abbildung 3.2 ist der Authentisierungsprozess schematisch dargestellt. Der Client ist dabei eine Applikation, über die sich ein Nutzer authentisieren möchte. Die Applikation verfügt zum einen über eine `client_id` und über ein `secret`, beides von TIMA vergebene eindeutige zufällige Strings. Der Authentisierungsprozess läuft wie folgt ab:

1. Eine Applikation sendet eine Anfrage an TIMA mit dem `username` des Nutzers und der `client_id`. TIMA prüft diese beiden Werte auf Existenz und antwortet entweder mit **200** (HTTP Response Code) und dem aktuellen Zeitstempel oder mit **404**.

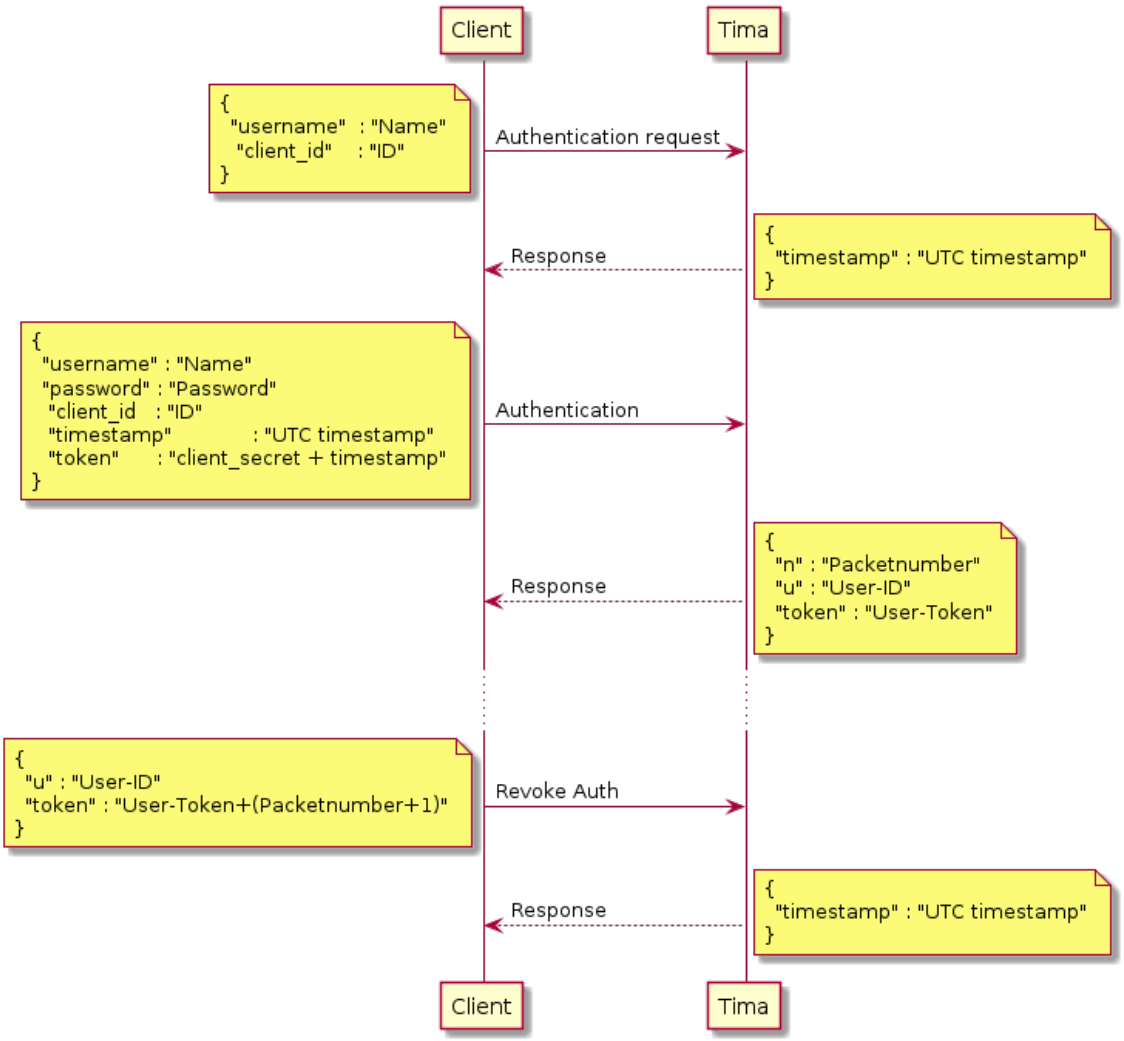

Abbildung 3.2: Authentisierungsprozess

2. Als nächstes sendet die Applikation die eigentliche Authentisierungsanfrage. Mit `username` und `password` des Nutzers, `client_id` der Applikation, dem Zeitstempel der Antwort der letzten Anfrage und einem `token` das aus dem `secret` der Applikation und dem Zeitstempel geniert wird (SHA512).

3. TIMA antwortet wenn die Authentisierung erfolgreich war mit **200** und den folgenden drei Werten:

   n Paketnummer - jede Anfrage einer Applikation muss diese um eins nach oben zählen. Als Wertebereich ist `uint32` zu benutzen.

   u eine eindeutige Nutzer-ID, die bei jeder Anfrage mit zusenden ist

token ein zufälliger String, der bei jeder Anfrage zusammen mit der Paketnummer `n` in einem SHA512 Hash zu senden ist

Aus Kompatibilitätsgründen läuft die Paketnummer im Wertebereich `uint32`, also bis maximal 2147483646, und fängt nach erreichen der Maximalzahl wieder bei 0 an.

3.2.3 OAI-PMH

Bei OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) handelt es sich um ein auf XML basierendes Protokoll zum Sammeln von Metadaten. Es wird dabei unterschieden zwischen Data Providern und Service Providern. Ein Data Provider betreibt ein oder mehrere Repositories, die OAI-PMH unterstützen um Metadaten bereitzustellen. Service Provider sammeln die Metadaten der Data Providern und bieten Mehrwertdienste an. Da TIMA ein offenes Projekt ist, liegt es nahe, derartige Projekte ebenfalls zu unterstützen.

Wir haben das OAI-PMH Protokoll implementiert um Metadaten zu den gesammelten Assoziationen bereitzustellen. Wir stellen dabei als Data Provider Metadaten zu den Wörtern und im geringen Maße zu den Nutzern bereit.

4 Applikationen und Webseite

Die Applikationen und die Webseite sollen die Verwendung vom TIMA für möglichst viele Endnutzer möglich machen.

4.1 Webseite

Die Webseite bzw. das Webfrontend ist die Hauptanlaufstelle für Nutzer. Hierüber kann er sowohl anonym als auch angemeldet Assoziationen eingegeben, Wörter und deren Assoziationen ansehen und weitere Funktionen wie die Rangliste und andere Statistiken aufrufen.

Das Webfrontend basiert auf Django, wurde zusätzlich zu HTML mit Bootstrap und JQuery erstellt. Zur Visualisierung von Assoziationsgraphen wurde D3.js verwendet.

4.2 Applikationen

Die Applikationen sollen die Verwendung von TIMA ohne Webbrowser ermöglichen. Durch Applikationen wird besonders für mobile Geräte die Nutzung vereinfacht. In einer erste Variante der Applikation wurden grundlegende Funktionen der Webseite nachgebaut. Dies diente unter anderem der Entwicklung der API, um etwaige Fehler im Übertraungsprotokoll aufzuzeigen und zu beheben beziehungsweise Verbesserungsmöglichkeiten zu erhalten.

4.2.1 Bibliothek

Als Bibliothek wurde sich für Qt5 entschieden Aufgrund der weitreichenden Unterstützung der Bibliothek auf verschiedenen Endgeräten. Hier ist besonders darauf hinzuweisen, das Qt5 sowohl auf Andriod als auch auf iOS läuft und so nicht die gleiche Applikation für beide Betriebssysteme geschrieben werden muss. Es wurde zudem Java in betracht bezogen, da dies vorrangig auf Mobilen Geräten Verwendung findet. Allerdings wurde sich aus den eben genannten, sowie Sympathiegründen des Entwicklers dagegen entschieden.

4.2.2 Aufbau

Die innere Logik wird durch einen Zustandsautomaten dargestellt um Mehrfachanfragen zu vermeiden und eine einfache Fehlerkorrektur zu ermöglichen. In Abbildung 4.1 wird der Zusammenhang der einzelnen Zustände angezeigt. Das Wechseln der Zustände wird ausschließlich über die Signale geregelt, die mit Qt implementiert sind.

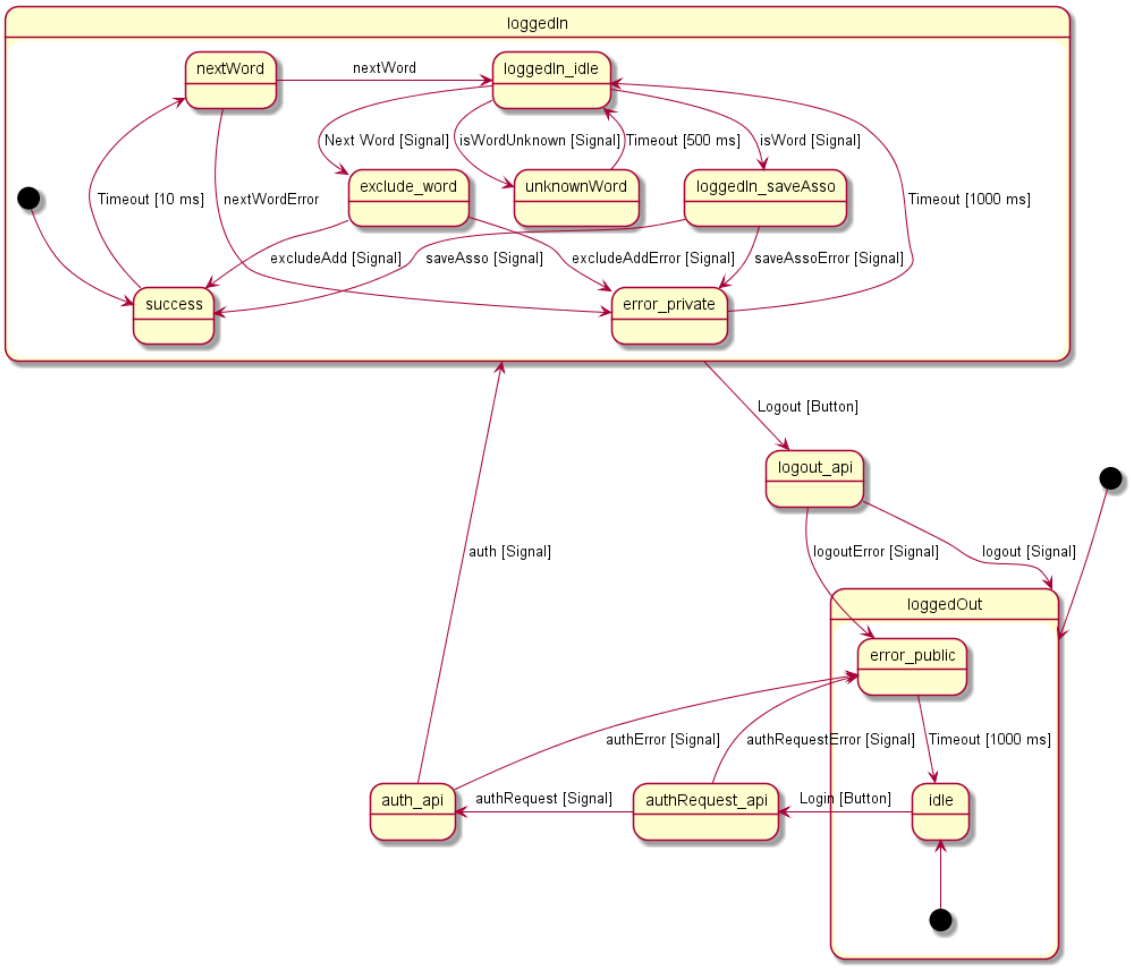

Abbildung 4.1: UML State Diagramm des Applikationszustandsautomaten

4.3 Sicherheit

Die Sicherheit hat bei der Entwicklung eine große Rolle gespielt. Jede Applikation nutzt die in Abschnitt 3.2.2 vorgestellte Autorisierungsmethode um mit der API zu kommunizieren. Dies dient dazu, dass lediglich von TIMA akzeptierte Applikationen Schreibrechte auf der Datenbank haben. Das Auslesen der Informationen bleibt davon jedoch unangetastet, sofern es sich nicht um benutzerspezifische Daten handelt, und ist nach wie vor für jeden offen.

5 Ausblick

Die bisher verrichtete Arbeit bietet sehr gute Vorraussetzungen für zukünftige Projekte, die entweder mit den Assoziationen arbeiten oder die Assoziationsdatenbank erweitern möchten. Durch strikte Trennung von Front- und Backend ist ein einfaches Austauschen beider Bestandteile jederzeit möglich. Um die Datenbank weiter zu füllen, waren einige weitere Spiele beziehungsweise Funktionen für bestehende Spiele geplant, die jedoch aus Zeitgründen nicht weiter verfolgt wurden.

Assoziationskette Momentan fehlt eine Implementierung als Applikation, in der dann ebenfalls auch kooperativ mit anderen menschlichen Gegnern gespielt werden kann. Eine interessante Spielart könnte dabei sein, zufälligen einen Mitspielern zugewiesen zu bekommen. Dies kann auch ein Computergegner sein. Nach Beendigung des Spiels den Spieler zu fragen, ob er mit einem Menschen oder dem Computer gespielt hat, ist eine interessante Abwandlung des Turingtests, der auch über die Qualität unserer Assoziationsdatenbank entscheiden kann.

Familienduell Das Spiel Familienduell ist momentan nicht implementiert, die Planung aber abgeschlossen, bietet jedoch großen Unterhaltungswert. Der Name soll hierbei nur die Ausrichtung des Spieles näher bringen, denn wie bei der Fernsehserie[5], werden dem Nutzer verschiedene verdeckte Antworten auf eine Frage gezeigt und für jede richtig gegebene Antwort erhält der Spieler Punkte. Die Fragen sind jedoch im Unterschied zum Fernsehen, ausschließlich die meist genannten Assoziationen zu einem bestimmten Wort. Da alle korrekten Antworten auf TIMA nachgeschaut werden können, sollte zusätzlich eine Zeitbegrenzung für das Eingeben implementiert werden, um so einen zusätzlichen Anreiz zu schaffen selbständig zu antworten. Je nach Schwierigkeitsgrad kann ein Zeitbonus für korrekte Assoziationen gegeben werden und ein Malus bei falschen Antworten. Die Einwirkung auf das Spielvergnügen müsste entsprechend getestet werden.

Jede gegebene Antwort sollte für das Füllen der Datenbank verwendet werden, obgleich es eine gesuchte Lösung war oder nicht. Punkte sollte ein Spieler jedoch nur für richtige Lösungen und einen Bonus für alle Lösungen erhalten.

[5] `https://de.wikipedia.org/wiki/Familien-Duell`

Andere Sprachen Wünschenswert wäre auch eine bessere Umsetzung aller Komponenten in andere Sprachen. Zum einen sollen weitere Sprachen für die Assoziationen hinzugefügt werden, so dass eine noch größere Nutzerschaft angesprochen werden kann. Derzeit werden hier die Sprachen Deutsch, Englisch, Persisch und Spanisch unterstützt.

Zum anderen soll das Webfrontend und die Applikation mehrsprachig sein. Hier ist momentan lediglich das Webfrontend in Deutsch und Englisch verfügbar.

Analsye der Daten Mit den vorhandenen Daten lassen sich viele Analysen durchführen. Zum Beispiel wäre ein Vergleich zwischen Kookurrenzen und Assoziationen möglich. Doch nicht nur im Bereich der automatischen Sprachverarbeitung lassen sich sicherlich interessante Beobachtungen feststellen. Wie sich der Hintergrund des Nutzers auf seine Assoziationen auswirken, ist ein spannender sozialwissenschaftlicher Aspekt.

Wir wünschen uns in der Zukunft viele Projekte, die TIMA unterstützen, weiter aufbauen und natürlich nutzen.

6 Zusammenfassung

TIMA ist eine Datenbank zum sammeln von Assoziationen. Den Hauptbestandteil bildet die Webseite bestehend aus Front- und Backend. Das Backend besteht aus der eigentlichen Datenbank und einer umfangreichen API. Diese setzt für Authentifizierung auf ein eigenes Protokoll. Um Nutzer für das Projekt zu gewinnen, setzen wir vor allem auf Spiele und Wettweberbcharakter durch Ranglisten. Zusätzlich zum Webfrontend haben wir eine Applikation implementiert, die das Spiel Assoziationskette auf mobilen Geräten ermöglicht.

Momentan befinden sich über 3000 Worte mit mehr als 4000 Assoziationen in der Datenbank. Verteilt sind diese auf vier verschiedene Sprachen: Deutsch, Englisch, Persisch und Spanisch. Die am stärksten vertretene Sprache ist mit großem Abstand Deutsch. Dies rührt sicherlich daher, dass TIMA bisher keine große Verbreitung gefunden hat. Die bisherigen zwölf angemeldeten Nutzer stammen vor allem aus dem Bekanntenkreis der Programmierer. Für die Zukunft gibt es viele verschiedene Anwendungsmöglichkeiten für TIMA, die den Nutzerkreis deutlich erweitern können.

7 Literaturverzeichnis

[WC11] Andrea Wiggins und Kevin Crowston. „From conservation to crowdsourcing: A typology of citizen science“. In: *System Sciences (HICSS), 2011 44th Hawaii international conference on*. IEEE. 2011, S. 1–10.