FACULTEIT
WETENSCHAPPEN

TESTed: one judge to rule them all

Een universele judge voor het beoordelen van software in een educative context

Niko Strijbol

Studentennummer: 01404620

Promotoren: prof. dr. Peter Dawyndt, dr. ir. Bart Mesuere
Begeleiding: Charlotte Van Petegem

Masterproef ingediend tot het behalen van de academische graad van
Master of Science in de informatica

Academiejaar: 2019 – 2020

UNIVERSITEIT
GENT

Inhoudsopgave

Dankwoord

Dank aan iedereen!

1. Educational software testing

1.1. Inleiding

TODO Programmeren -> steeds belangrijker en nuttigere kennis om te hebben Goed programmeren -> vereist veel oefening Zeker in cursussen met meer mensen -> goede ondersteuning lesgever -> veel tijd Automatiseren van beoordeling programmeeroefeningen -> Dodona

1.2. Wat is Dodona?

TODO Intro over Dodona: korte geschiedenis, terminologie, hoe Dodona werkt (oefeningen, judges, enz.) -> Over de judge wordt in het deel hierna meer verteld.

Dodona:

- Opgaves (of oefeningen), opgesteld door lesgevers - Oplossingen, opgesteld door studenten - Judge beoordeelt oplossing van een opgave, door Dodona-team

1.3. Beoordelen van oplossingen

1.3.1. De judge

In Dodona wordt elke ingediende oplossing beoordeeld door een evaluatieprogramma, de *judge*. In wezen is dit een eenvoudig programma: via de standaardinvoerstroom (`stdin`) krijgt het programma een configuratie binnen van Dodona. Deze configuratie bevat de invoer, bestaande uit onder andere de programmeertaal van de oplossing, het pad naar het oplossingsbestand en geheugen- en tijdslimieten. Het resultaat van de beoordeling wordt uitgeschreven naar de standaarduitvoerstroom (`stdout`). Zowel de invoer als de uitvoer van de judge zijn json, waarvan het formaat vastgelegd is in een json-schema.[1]

Concreet wordt elke beoordeling uitgevoerd in een Docker-container. Deze Docker-container wordt gemaakt op basis van een Docker-image die bij de judge hoort, en alle dependencies bevat die de judge in kwestie nodig heeft. Bij het uitvoeren van de beoordeling zal Dodona een *bind mount*[2] voorzien, zodat de code van de judge zelf, de code van de oefening en de code van de student beschikbaar zijn in de container. Via de configuratie geeft Dodona aan de judge aan waar deze bestanden zich bevinden.

Samenvattend bestaat interface tussen de judge en Dodona uit drie onderdelen:

[1] Dit schema en een tekstuele beschrijving ervan is te vinden in de handleiding op https://dodona-edu.github.io/en/guides/creating-a-judge/.

[2] Informatie over deze term vindt u op https://docs.docker.com/storage/bind-mounts/

1. De judge zal uitgevoerd worden in een Docker-container, dus een Docker-image met alle dependencies moet voorzien worden. Deze Docker-image moet ook de judge opstarten.
2. De judge stelt de invoer van een beoordeling ter beschikking voor de judge. Bestanden worden via een bind mount aan de Docker-container gekoppeld. De paden naar deze bestanden binnen de container en andere informatie (zoals programmeertaal van de oplossing of natuurlijke taal van de gebruiker) worden via de configuratie aan de judge gegeven (via standaardinvoer).
3. De judge moet het resultaat van zijn beoordeling uitschrijven naar standaarduitvoer, in een vastgelegd formaat.

Buiten deze interface legt Dodona geen vereisten op aan de werking van judge. Door deze vrijheid lopen de manieren waarop de bestaande judges geïmplementeerd zijn uiteen. Sommige judges beoordelen oplossingen in dezelfde programmeertaal als de taal waarin ze geschreven zijn. Zo is de judge voor Python-oplossingen geschreven in Python en de judge voor Java-oplossingen in Java. Bij andere judges is dat niet het geval: de judges voor Bash en Prolog zijn bijvoorbeeld ook in Python geschreven. Ook heeft elke judge een eigen manier waarop de testen voor een oplossing opgesteld moeten worden. Zo worden in de Java-judge jUnit-testen gebruikt, terwijl de Python-judge doctests en een eigen formaat ondersteunt.

1.3.2. De beoordeling zelf

De beoordeling van een oplossing van een student laat zich beschreven als het volgende stappenplan:

1. De student dient de oplossing in via de webinterface van Dodona.
2. Dodona start een Docker-container voor de judge.
3. Dodona voorziet de container van de bestanden van de judge, de oefening en de ingediende oplossing.
4. De judge wordt uitgevoerd met de configuratie als invoer.
5. De judge beoordeelt de oplossing aan de beoordelingsmethodes opgesteld door de lesgever (d.w.z. de jUnit-test, de doctests, …). Sommige judges voeren ook bijkomende taken, zoals linting, beoordeling van de performantie of *grading* van de code van de oplossing.
6. De judge vertaalt zijn beoordeling naar het Dodona-formaat en schrijft het resultaat naar het standaarduitvoerkanaal.
7. Dodona slaat dat resultaat op in de databank.
8. Op de webinterface krijgt de student het resultaat te zien als feedback op de ingediende oplossing.

1.4. Probleemstelling

De manier waarop de huidige judges werken, resulteert in twee belangrijke nadelen. Bij het bespreken hiervan is het nuttig een voorbeeld in het achterhoofd te houden, teneinde de nadelen te kunnen concretiseren. Als voorbeeld gebruiken we de „Lotto"-oefening[3], met volgende opgave:

> De **lotto** is een vorm van loterij die voornamelijk bekend is vanwege de genummerde balletjes, waarvan er een aantal getrokken worden. Deelnemers mogen zelf hun eigen nummers aankruisen op een lottoformulier. Hoe groter het aantal overeenkomstige nummers tussen het formulier en de getrokken balletjes, hoe groter de geldprijs.
>
> ### Opgave
>
> Schrijf een functie `loterij` waarmee een lottotrekking kan gesimuleerd worden. De functie moet twee parameters `aantal` en `maximum` hebben. Aan de parameter `aantal` (`int`) kan doorgegeven worden hoeveel balletjes a er moeten getrokken worden (standaardwaarde 6). Aan de parameter `maximum` (`int`) kan doorgegeven worden uit hoeveel balletjes m er moet getrokken worden (standaardwaarde 42). Beide parameters kunnen ook weggelaten worden, waarbij dan de standaardwaarde gebruikt moet worden. De balletjes zijn daarbij dus genummerd van 1 tot en met m. Je mag ervan uitgaan dat $1 \leq a \leq m$. De functie moet een string (`str`) teruggeven die een strikt stijgende lijst (`list`) van a natuurlijke getallen (`int`) beschrijft, waarbij de getallen van elkaar gescheiden zijn door een spatie, een koppelteken (-) en nog een spatie. Voor elk getal n moet gelden dat $1 \leq n \leq m$.
>
> ### Voorbeeld
>
> ```
> > loterij()
> '2 - 17 - 22 - 27 - 35 - 40'
> > loterij(aantal=8)
> '5 - 13 - 15 - 31 - 34 - 36 - 39 - 40'
> > loterij(aantal=4, maximum=38)
> '16 - 20 - 35 - 37'
> ```

Oplossingen voor deze oefening staan in codefragmenten 1.1 en 1.2, voor respectievelijk Python en Java.

Implementatie van oefeningen

Het eerste en belangrijkste nadeel aan de werking van de huidige judges heeft betrekking op de lesgevers en komt voor als zij een oefening willen aanbieden in meerdere programmeertalen. Enerzijds is dit een zware werklast: de oefening, en vooral de code voor de beoordeling, moet voor elke judge opnieuw geschreven worden. Voor de Python-judge zullen doctests nodig zijn, terwijl de

[3] Vrij naar een oefening van prof. Dawyndt. De originele oefening is beschikbaar op https://dodona.ugent.be/nl/exercises/2025591548/

```java
import java.util.HashSet;
import java.util.Set;
import java.util.concurrent.ThreadLocalRandom;
import java.util.stream.Collectors;

class Main {

    public static String loterij(int aantal, int maximum) {
        var r = ThreadLocalRandom.current();
        var result = new HashSet<Integer>();
        while (result.size() < aantal) {
            result.add(r.nextInt(1, maximum + 1));
        }
        return result.stream()
                .sorted()
                .map(Object::toString)
                .collect(Collectors.joining(" - "));
    }

    public static String loterij(int aantal) {
        return loterij(aantal, 42);
    }

    public static String loterij() {
        return loterij(6, 42);
    }
}
```

Codefragment 1.1.: Voorbeeldoplossing in Java.

```python
from random import randint


def loterij(aantal=6, maximum=42):
    getallen = set()
    while len(getallen) < aantal:
        getallen.add(randint(1, maximum))

    return " - ".join(str(x) for x in sorted(getallen))
```

Codefragment 1.2.: Voorbeeldoplossing in Python.

Java-judge jUnit-testen vereist. Anderzijds lijdt dit ook tot verschillende versies van dezelfde oefening, wat het onderhouden van de oefeningen moelijker maakt. Als er bijvoorbeeld een fout sluipt in de beoordelingscode, zal de lesgever er aan moeten denken om de fout te verhelpen in alle varianten van de oefening. Bovendien geeft elke nieuwe versie van de oefening een nieuwe mogelijkheid voor het introduceren van fouten.

Kijken we naar onze Lotto-oefening, merken we dat het gaat om een eenvoudige opgave en een eenvoudige oplossing. Bovendien zijn de verschillen tussen oplossingen in verschillende programmeertalen niet zo groot. In de voorbeeldoplossingen in Python en Java zijn de verschillen minimaal, zij het dat de Java-oplossing wat langer is. De Lotto-oefening zou zonder problemen in nog vele andere programmeertalen opgelost kunnen worden. Eenvoudige programmeeroefeningen, zoals de Lotto-oefening, zijn voornamelijk nuttig in twee gevallen: studenten die voor het eerst leren programmeren en studenten die een nieuwe programmeertaal leren. In het eerste geval is de eigenlijke programmeertaal minder relevant: het zijn vooral de concepten die belangrijk zijn. In het tweede geval is de programmeertaal wel van belang, maar moeten soortgelijke oefeningen gemaakt worden voor elke programmeertaal die aangeleerd moet worden. In beide gevallen is het dus een meerwaarde om de oefening in meerdere programmeertalen aan te bieden.

We kunnen tot eenzelfde constatatie komen bij meer complexe oefeningen die zich concentreren op algoritmen: ook daar zijn de concepten belangrijker dan in welke programmeertaal een algoritme uiteindelijk geïmplementeerd wordt. Een voorbeeld hiervan is het vak „Algoritmen en Datastructuren" dat gegeven wordt door prof. Fack binnen de opleiding wiskunde[4]. Daar zijn de meeste opgaven vandaag al beschikbaar in Java en Python op Dodona, maar dan als afzonderlijke oefeningen.

Een ander aspect is de beoordeling van een oefening. Voor de Lotto-oefening is de beoordeling niet triviaal, door het gebruik van niet-deterministische functies. Het volstaat voor dit soort oefeningen niet om de uitvoer gegenereerd door de oplossing te vergelijken met een op voorhand vastgelegde verwachte uitvoer. De geproduceerde uitvoer zal moeten gecontroleerd worden met code, specifiek gericht op deze oefening, die de verwachte vereisten van de oplossing controleert. Deze evaluatiecode moet momenteel voor elke programmeertaal en dus elke judge opnieuw geschreven worden. In de context van de Lotto-oefening controleert deze code bijvoorbeeld of de gegeven getallen binnen het bereik liggen en of ze gesorteerd zijn.

Implementatie van judges

Een tweede nadeel aan de werking zijn de judges zelf: voor elke programmeertaal die men wil aanbieden in Dodona moet een nieuwe judge ontwikkeld worden. Ook hier is er dubbel werk: dezelfde concepten en features, die eigenlijk programmeertaalonafhankelijk zijn, moeten in elke judge opnieuw geïmplementeerd worden. Hierbij denken we aan bijvoorbeeld de logica om te bepalen wanneer een beoordeling positief of negatief moet zijn.

Onderzoeksvraag

We beschouwen het eerste nadeel als het belangrijkste nadeel, en vatten het samen als de onderzoeksvraag waarop deze thesis een antwoord wil bieden:

[4]De studiefiche is beschikbaar op https://studiegids.ugent.be/2019/NL/studiefiches/C002794.pdf

> Is het mogelijk om een judge zo te implementeren dat de opgave en beoordelingsmethoden van een oefening slechts eenmaal opgesteld dienen te worden, waarna de oefening beschikbaar is in alle programmeertalen die de judge ondersteunt? Hierbij willen we dat eens een oefening opgesteld is, deze niet meer gewijzigd moet worden wanneer talen toegevoegd worden aan de judge.

Als bijzaak zijn we ook geïnteresseerd of de judge uit de onderzoeksvraag een voordeel kan bieden voor het implementeren van judges zelf (het tweede nadeel).

De aandachtige lezer zal opmerken dat de opgave voor de Lotto-oefening programmeertaalspecifieke en taalspecifieke elementen bevat. Zo zijn de voorbeelden in Python en zijn de namen van functies en argumenten in het Nederlands. Beide zaken worden voor deze thesis expliciet als *out-of-scope* gezien en zullen niet behandeld worden.

1.5. Opbouw van de thesis

Hoofdstuk 2 handelt over het antwoord op bovenstaande vraag, waar een prototype van een dergelijke judge wordt voorgesteld. Daarna volgt ter illustratie een gedetailleerde beschrijving van hoe een oefening opgesteld moet worden voor deze judge. Nadien volgt een beschrijving van hoe een nieuwe programmeertaal moet toegevoegd worden. Daar deze twee hoofdstukken voornamelijk ten doel hebben zij die met de judge moeten werken te informeren, nemen deze hoofdstukken de vorm aan van meer traditionele softwarehandleidingen. Tot slot wordt afgesloten met een hoofdstuk over beperkingen van de huidige implementaties, en waar er verbeteringen mogelijk zijn (het „toekomstige werk”).

2. TESTed

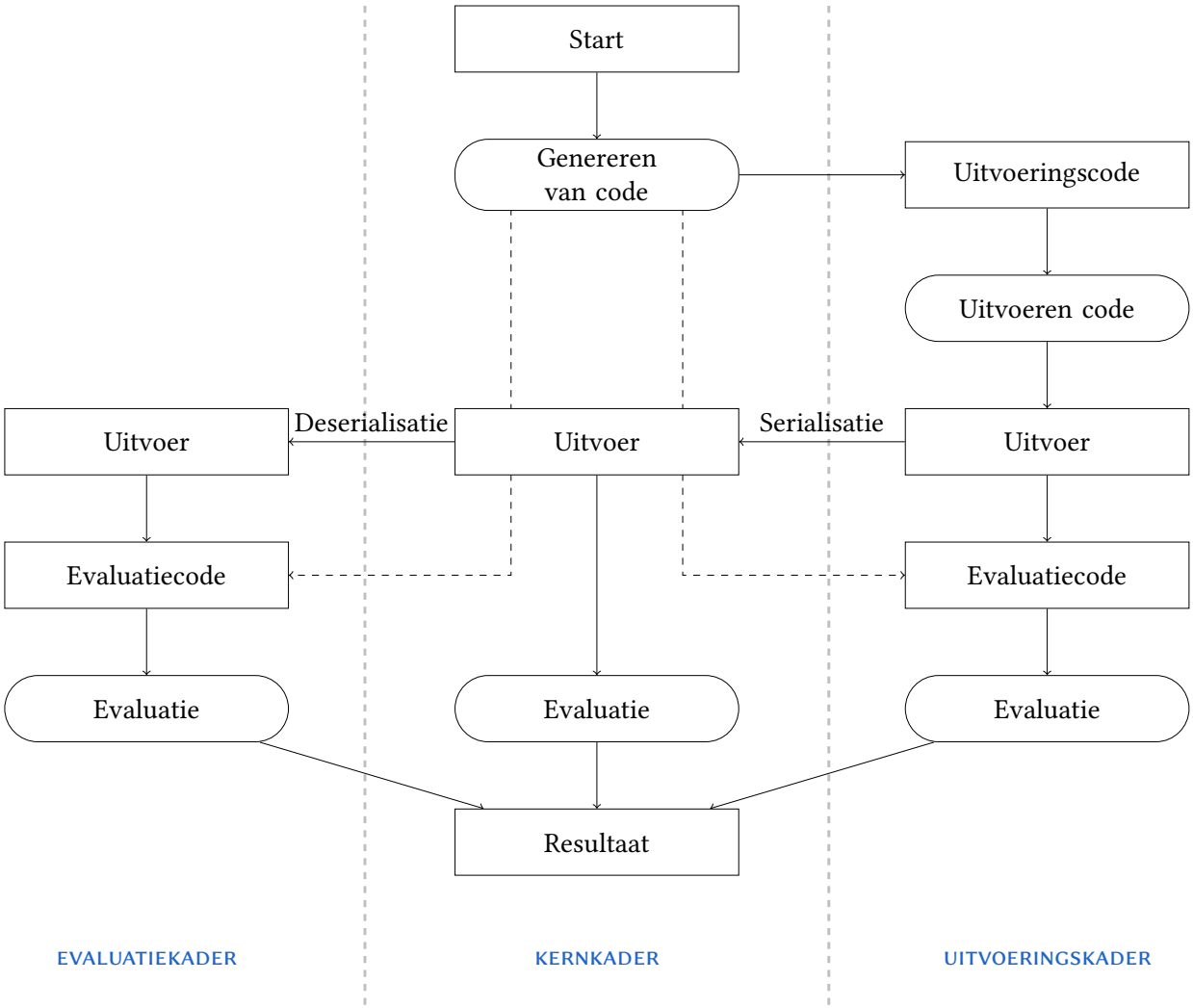
Het antwoord op de onderzoeksvraag uit het vorige hoofdstuk neemt de vorm aan van een nieuwe judge voor het Dodona-platform: de *universele judge*. Bij deze judge is de opgave programmeertaalonafhankelijk en kan de judge oplossingen in verschillende programmeertalen evalueren. Dit hoofdstuk licht de werking en implementatie van deze judge toe, beginnend met een algemeen overzicht, waarna elk onderdeel in meer detail besproken wordt.

2.1. Overzicht

Figuur 2.1 toont de opbouw van de judge op schematische wijze. De twee dikkere stippellijnen geven de programmeertaalbarrières aan, die de judge ook logisch onderverdelen in drie contexten of processen, die we kaders noemen (om verwarring te vermijden met een computerproces en contexten uit het testplan, waarover later meer). De judge zelf is geschreven in Python, wat ook betekent dat in het *kernkader* enkel Python gebruikt wordt. De oplossing wordt uitgevoerd in het *uitvoeringskader*, waar de programmeertaal de programmeertaal van de oplossing is. Tot slot is er nog het *evaluatiekader*, waar door de lesgever geschreven evaluatiecode wordt uitgevoerd. Deze moet niet in dezelfde programmeertaal als de oplossing zijn.

Het schema van hierboven kunnen we ook uitschrijven als een stappenplan, wat hieronder gebeurt.

1. De judge wordt opgestart en het testplan wordt geladen. Het gestarte proces noemen we het *kernkader*. Het is vooral hier dat de code van de judge uitgevoerd wordt.
2. Er wordt gecontroleerd of het testplan de programmeertaal van de oplossing ondersteunt. Als het testplan bijvoorbeeld programmeertaalspecifieke code bevat die enkel voor Java gegeven is, zal een oplossing in Python niet geëvalueerd kunnen worden. Bevat het testplan een functie die een verzameling moet teruggeven, dan zullen talen als Bash niet in aanmerking komen.
3. De code voor het evalueren van de oplossing wordt gegenereerd en eventueel gecompileerd.
4. Nu kan het *uitvoeringskader* beginnen: de gegenereerde code wordt uitgevoerd. Elke context uit he testplan wordt in een afzonderlijk subproces uitgevoerd, teneinde het delen van informatie tegen te gaan. De onafhankelijkheid van de contexten laat ons ook toe om contexten in parallel uit te voeren. Toch moeten we opmerken dat het parallel uitvoeren vooral een theoretische mogelijkheid is: in werkelijkheid gebruiken er meerdere gebruikers tegelijk Dodona, waardoor er ook meerdere oefeningen tegelijk beoordeeld worden, wat op zijn beurt ervoor zorgt dat er geen grote ruimte voor snelheidswinst door parallellisatie binnen de judge is.
5. De uitvoering in de vorige stap resulteert in enkele resultaten. Deze worden verzameld en beoordeeld. Hiervoor zijn drie mogelijke manieren:



Figuur 2.1.: Schematische voorstelling van de opbouw van de universele judge.

a) Programmeertaalspecifieke evaluatie. Hierbij wordt de evaluatie gedaan na de uitvoering gedaan in hetzelfde proces als de uitvoering. In het schema zitten we nog steeds in het uitvoeringskader.

b) Aangepaste evaluator. Hierbij is er evaluatiecode geschreven die los staat van de oplossing. De evaluatiecode kan in een andere programmeertaal geschreven zijn dan de oplossing. De aangepaste evaluator wordt gegenereerd, gecompileerd en uitgevoerd na het uitvoeren van de oplossing, in het *evaluatiekader*.

c) Ingebouwde evaluatie. Hierbij is het de judge zelf die evalueert, waardoor dit vooral eenvoudige evaluaties betreft, zoals het vergelijken van geproduceerde uitvoer en verwachte uitvoer. Dit gebeurt in het kernproces.

6. Tot slot verzamelt de judge alle evaluatieresultaten en stuurt ze door naar Dodona, waarna ze getoond worden aan de gebruiker.

2.2. Beschrijven van een oefening

Elke evaluatie begint met het *testplan*, een document dat beschrijft hoe een oplossing voor een oefening geëvalueerd moet worden. Het vervangt de taalspecifieke testen van de bestaande judges (ie. de jUnit-tests of de doctests in respectievelijk Java en Python). Het testplan sensu lato wordt opgedeeld in verschillende onderdelen, die hierna besproken worden.

2.2.1. Het testplan

Het testplan sensu stricto beschrijft de structuur van een evaluatie van een oplossing voor een oefening. Deze structuur lijkt qua opbouw sterk op de structuur van de feedback zoals gebruikt door Dodona. Dat de structuur van de oplossing in Dodona en van het testplan op elkaar lijken heeft als voordeel dat er geen mentale afbeelding tussen de structuur van het testplan en dat van Dodona moet bijgehouden worden. Concreet ziet de structuur er als volgt uit:

Tab Een testplan bestaat uit verschillende *tab*s of tabbladen. Deze komen overeen met de tabbladen in de gebruikersinterface van Dodona. Een tabblad kan een naam hebben, die zichtbaar is voor de gebruikers.

Context Elk tabblad bestaat uit een of meerdere *context*en. Een context is een onafhankelijke uitvoering van een evaluatie. De nadruk ligt op de „onafhankelijkheid", zoals al vermeld. Elke context wordt in een nieuw proces en in een eigen map uitgevoerd, zodat de kans op het delen van informatie klein is. Hierbij willen we vooral onbedoeld delen van informatie (zoals statische variabelen of het overschrijven van bestanden) vermijden. De gemotiveerde student zal nog steeds informatie kunnen delen tussen de uitvoeringen, door bv. in een andere locatie een bestand aan te maken en later te lezen.

Testcase Een context bestaat uit een of meerdere *testcase*s of testgevallen. Een testgeval bestaat uit invoer en een aantal tests. De testgevallen kunnen onderverdeeld worden in twee soorten:

Main testcase of hoofdtestgeval. Van deze soort is er maximaal een per context. Dit testgeval heeft als doel het uitvoeren van de main-functie (of de code zelf als het gaat om een scripttaal zoals Bash of Python). Als invoer voor dit testgeval kunnen enkel het standaardinvoerkanaal en de programma-argumenten meegegeven worden.

Normal testcase of normaal testgeval. Hiervan kunnen er nul of meer zijn per context. Deze testgevallen zijn voor andere aspecten te testen, nadat de code van de gebruiker met success ingeladen is. De invoer is dan ook uitgebreider: het kan gaan om het standaardinvoerkanaal, functieoproepen en variabeletoekenningen. Een functieoproep of variabeletoekenning is verplicht (zonder functieoproep of toekenning aan een variabele is er geen code om te testen).

Test De uitvoer van een testgeval bestaat uit meerdere *test*s, die elk een aspect van een dat testgeval controleren. Met aspect bedoelen we het standaarduitvoerkanaal, het standaardfoutkanaal, opgevangen uitzonderingen (*exceptions*), de teruggegeven waarden van een functieoproep (returnwaarde) of de inhoud van een bestand. Elke test bevat instructies om het resultaat te beoordelen, door een van de drie eerder beschreven evaluatiemogelijkheden te gebruiken.

Bij de keuze voor een formaat voor het testplan (json, xml, ...), hebben we vooraf enkele vereisten geformuleerd waaraan het gekozen formaat moet voldoen. Het moet:

- leesbaar zijn voor mensen,
- geschreven kunnen worden met minimale inspanning, met andere woorden de syntaxis dient eenvoudig te zijn, en
- programmeertaalonafhankelijk zijn.

Uiteindelijk is gekozen om het op te stellen in json. Niet alleen voldoet json aan de vooropgestelde voorwaarden, het wordt ook door veel talen ondersteund.

Toch zijn er ook enkele nadelen aan het gebruik van json. Zo is json geen beknopte of compacte taal om met de hand te schrijven. Een oplossing hiervoor gebruikt de eigenschap dat veel talen json kunnen produceren: andere programma's kunnen desgewenst het testplan in het json-formaat genereren, waardoor het niet met de hand geschreven moet worden. Hiervoor denken we aan een *DSL* (*domain specific language*), maar dit valt buiten de thesis en wordt verder besproken in hoofdstuk 5.

Een tweede nadeel is dat json geen programmeertaal is. Terwijl dit de implementatie van de judge bij het interpreteren van het testplan weliswaar eenvoudiger maakt, is het tevens beperkend: beslissen of een testgeval moet uitgevoerd worden op basis van het resultaat van een vorig testgeval is bij wijze als voorbeeld niet mogelijk. Ook deze beperking wordt uitgebreider besproken in hoofdstuk 5.

2.2.2. Dataserialisatie

In het testplan, zoals beschreven in de paragraaf hierboven, wordt gewag gemaakt van returnwaarden en variabeletoekenningen. Aangezien het testplan programmeertaalonafhankelijk is, moet er dus een manier zijn om data uit de verschillende programmeertalen voor te stellen en te vertalen: het *serialisatieformaat*.

Keuze van het formaat

Zoals bij het testplan is een keuze voor een bepaald formaat gemaakt. Daarvoor zijn er opnieuw enkele voorwaarden vooropgesteld, waaraan het serialisatieformaat moet voldoen. Het formaat moet:

- door mensen geschreven kunnen worden (*human writable*),
- onderdeel van het testplan kunnen zijn,
- in meerdere programmeertalen bruikbaar zijn, en
- de types ondersteunen die we willen aanbieden in het programmeertaalonafhankelijke deel van het testplan.

Een voor de hand liggende oplossing is ook hiervoor json gebruiken, en zelf in json een structuur op te stellen voor de waarden. In tegenstelling tot het testplan bestaan er al een resem aan dataserialisatieformaten, waardoor het de moeite is om na te gaan of er geen bestaand formaat voldoet aan de vereisten. Hiervoor is gestart van een overzicht op Wikipedia, (Wikipedia-bijdragers 2020). Uiteindelijk is niet gekozen voor een bestaand formaat, maar voor de json-oplossing. De redenen hiervoor zijn samen te vatten als:

- Het gaat om een binair formaat. Binaire formaten zijn uitgesloten op basis van de eerste twee voorwaarden die we opgesteld hebben: mensen kunnen het niet schrijven zonder hulp van bijkomende tools en het is moeilijk in te bedden in een json-bestand (zonder gebruik te maken van encoderingen zoals base64). Bovendien zijn binaire formaten moeilijker te implementeren in sommige talen.
- Het formaat ondersteunt niet alle gewenste types. Sommige formaten hebben ondersteuning voor complexere datatypes, maar niet voor alle complexere datatypes die wij nodig hebben. Uiteraard kunnen de eigen types samengesteld worden uit basistypes, maar dan biedt de ondersteuning voor de complexere types weinig voordeel, aangezien er toch een eigen dataschema voor die complexere types opgesteld zal moeten worden.
- Sommige formaten zijn omslachtig in gebruik. Vaak ondersteunen dit soort formaten meer dan wat wij nodig hebben.
- Het formaat is niet eenvoudig te implementeren in een programmeertaal waarvoor geen ondersteuning is. Sommige dezer formaten ondersteunen weliswaar veel talen, maar we willen niet dat het serialisatieformaat een beperkende factor wordt in welke talen door de judge ondersteund worden. Het mag niet de bedoeling zijn dat het implementeren van het serialisatieformaat het meeste tijd in beslag neemt.

Een lijst van de overwogen formaten met een korte beschrijving volgt:

Apache Avro Een volledig „systeem voor dataserialisatie”. De specificatie van het formaat gebeurt in json, terwijl de eigenlijke data binair geëncodeerd wordt. Heeft uitbreidbare types, met veel ingebouwde types (*Apache Avro™ 1.9.1 Documentation* 2019).

Apache Parquet Minder relevant, dit is een bestandsformaat voor Hadoop (*Apache Parquet* 2020).

ASN.1 Staat voor *Abstract Syntax Notation One*, een formaat uit de telecommunicatie. De hoofdstandaard beschrijft enkel de notatie voor een dataformaat. Andere standaarden beschrijven dan de serialisatie, bv. een binair formaat, json of xml. De meerdere serialisatievormen zijn in theorie aantrekkelijk: elke taal moet er slechts een ondersteunen, terwijl de judge ze allemaal kan ondersteunen. In de praktijk blijkt echter dat voor veel talen er slechts één serialisatieformaat is, en dat dit vaak het binaire formaat is (*Information technology – Abstract Syntax Notation One (ASN.1): Specification of basic notation* 2015).

Bencode Schema gebruikt in BitTorrent. Het is gedeeltelijk binair, gedeeltelijk in text (Cohen 2017).

Binn Binair dataformaat (Ramos 2019).

BSON Een binaire variant op json, geschreven voor en door MongoDB (*BSON 1.1* 2019).

CBOR Een lichtjes op json gebaseerd formaat, ook binair. Heeft een goede standaard, ondersteunt redelijk wat talen (Bormann en Hoffman 2013).

FlatBuffers Lijkt op ProtocolBuffers, allebei geschreven door Google, maar verschilt wat in implementatie van ProtocolBuffers. De encodering is binair (Oortmerssen 2019).

Fast Infoset Is eigenlijk een manier om xml binair te encoderen (te beschouwen als een soort compressie voor xml), waardoor het minder geschikt voor ons gebruik wordt (*Information technology – Generic applications of ASN.1: Fast infoset* 2005).

Ion Een superset van json, ontwikkeld door Amazon. Het heeft zowel een tekstuele als binaire voorstelling. Naast de gebruikelijke json-types, bevat het enkele uitbreidingen. (*Amazon Ion* 2020).

MessagePack Nog een binair formaat dat lichtjes op json gebaseerd is. Lijkt qua types sterk op json. Heeft implementaties in veel talen (Furuhashi 2018).

OGDL Afkorting voor *Ordered Graph Data Language*. Daar het om een serialisatieformaat voor grafen gaat, is het niet nuttig voor ons doel (*OGDL 2018.2* 2018).

OPC Unified Architecture Een protocol voor intermachinecommunicatie. Complex: de specificatie bevat 14 documenten, met ongeveer 1250 pagina's (*OPC unified architecture - Part 1: Overview and concepts* 2016).

OpenDLL Afkorting voor de *Open Data Description Language*. Een tekstueel formaat, bedoeld om arbitraire data voor te stellen. Wordt niet ondersteunt in veel programmeertalen, in vergelijking met bv. json (Lengyel 2017).

ProtocolBuffers Lijkt zoals vermeld sterk op FlatBuffers, maar heeft nog extra stappen nodig bij het encoderen en decoderen, wat het minder geschikt maakt (*ProtocolBuffers* 2019).

Smile Nog een binaire variant van json (Jackson JSON team 2010).

SOAP Afkorting voor *Simple Object Access Protocol*. Niet bedoeld als formaat voor dataserialisatie, maar voor communicatie tussen systemen over een netwerk (Mitra en Lafon 2007).

SDXF Binair formaat voor data-uitwisseling. Weinig talen ondersteunen dit formaat (Wildgrube 2001).

Thrift Lijkt sterk op ProtocolBuffers, maar geschreven door Facebook (Slee, Agarwal en Kwiatkowski 2007).

UBJSON Nog een binaire variant van json (*Universal Binary JSON* 2018).

```json
{
  "type": "list",
  "data": [
    {
      "type": "integer",
      "data": 5
    },
    {
      "type": "integer",
      "data": 15
    }
  ]
}
```

Codefragment 2.1.: Een lijst bestaande uit twee getallen, geëncodeerd in het serialisatieformaat.

Geen enkel overwogen formaat heeft grote voordelen tegenover een eigen structuur in json. Daarenboven hebben vele talen het nadeel dat ze geen json zijn, waardoor we een nieuwe taal moeten inbedden in het bestaande json-testplan. Dit nadeel, gekoppeld met het ontbreken van voordelen, heeft geleid tot de keuze voor json.

Dataschema

Het formaat is, zoals al vermeld, json. Het bijhorende dataschema is met opzet eenvoudig gehouden, om implementaties makkelijker te maken. Concreet wordt een waarde voorgesteld als een json-object dat bestaat uit de (geëncodeerde) waarde en het type van die waarde. Een voorbeeld is codefragment 2.1.

Datatypes

Naast de encodering van de data is er een tweede aspect van het serialisatieformaat: de datatypes. Het formaat ondersteunt de meeste basistypes die in bijna elke programmeertaal beschikbaar zijn. Hieronder volgt lijst met een korte omschrijving van de ondersteunde types. Hierbij is er een speciaal type, aangeduid met een ster (*), dat niet gebruikt wordt bij het encoderen van data. Het type `literal` is bedoeld voor waarden die eigenlijk geen data zijn, maar verwijzingen naar bv. een variabele (een *identifier* in de programmeertaal). Dit is nuttig bij functieargumenten (zo kunnen variabelen worden gebruikt bij een functieoproep).

`integer` Gehele getallen.

`rational` Rationale getallen.

`text` Een tekenreeks of string.

`literal*` Een tekstuele waarde die rechtstreeks als *identifier* wordt gebruikt. Deze waarde wordt enkel gebruikt bij het aangeven van de types van functie-argumenten.

`unknown` Dit type wordt gebruikt als er onbekende types zijn bij het encoderen van een waarde. Bij het omzetten van een waarde uit het serialisatieformaat naar een programmeertaal (deserialisatie) worden waarden van dit type genegeerd.

`boolean` Een Boolese waarde (of boolean).

`list` Een wiskundige rij, wat wil zeggen dat de volgorde belangrijk is en dat dubbele elementen toegelaten zijn. Merk op dat sommige talen meerdere implementaties hebben voor het concept van lijst. Het is de implementatie vrij om te kiezen welk concept gebruikt wordt. Zo wordt bijvoorbeeld in de Java-implementatie `List` in plaats van `array` gebruikt, om consistent te zijn met de implementatie van `set` en `object`.

`set` Een wiskundige verzameling, wat wil zeggen dat de volgorde niet belangrijk is en dat dubbele elementen niet toegelaten zijn.

`object` Een wiskundige afbeelding: elk element wordt afgebeeld op een ander element. In Java is dit bijvoorbeeld een `Map`, in Python een `dict` en in Javascript een `object`.

`nothing` Geeft aan dat er geen waarde is, ook wel `null`, `None` of `nil` genoemd.

2.2.3. Functieoproepen en assignments

Een ander onderdeel van het testplan verdient ook speciale aandacht: het toekennen van variabelen (*assignments*) en de functieoproepen.

In heel wat oefeningen, en zeker in objectgerichte programmeertalen, is het toekennen van een waarde aan een variabele, om deze later te gebruiken, onmisbaar. Bijvoorbeeld zou een opgave kunnen bestaan uit het implementeren van een klasse. Bij de evaluatie dient dan een instantie van die klasse aangemaakt te worden, waarna er methoden kunnen aangeroepen worden, zoals hieronder geïllustreerd in een fictief voorbeeld.

```
var variabele = new DoorDeStudentGemaakteKlasse();
assert variabele.testfunctie1() == 15;
assert variabele.testfunctie2() == "Vijftienduizend";
```

Concreet is ervoor gekozen om het testplan niet uit te breiden met generieke statements of expressions, maar de ondersteuning te beperken tot assignments en functieoproepen. Dit om de implementatie van de vertaling van het testplan naar de ondersteunde programmeertalen niet nodeloos ingewikkeld te maken. Een functieoproep ziet er als volgt uit:

```
{
  "type": "top|object|constructor|identity",
  "name": "Naam van de functie",
  "object": "Optioneel object bij de functie",
  "arguments": ["Lijst van argumenten"]
}
```

De lijst van argumenten is een lijst waarden, voorgesteld in het serialisatieformaat. Het type van de functie geeft aan welke soort functie het is. Mogelijke waarden zijn momenteel:

`top` Een *top-level* functie. Afhankelijk van de programmeertaal zal deze functie toch omgezet worden naar een functie op een object (bv. naar een statische functie in Java).

`object` Een functie van een object. De invulling hiervan is gedeeltelijk programmeertaalafhankelijk, en meer specifiek van welke invulling gegeven wordt aan een object. Bij dit soort functies moet het `object` gegeven worden, wat het object is waaraan de functie vasthangt. In de praktijk kan dit mechanisme ook gebruikt worden voor functies in bijvoorbeeld een namespace of een statische functie in Java.

`constructor` Deze soort heeft dezelfde semantiek als een top-level functie, met dien verstande dat het om een constructor gaat. In Java zal bijvoorbeeld het keyword `new` vanzelf toegevoegd worden. De functienaam doet dienst als naam van de klasse.

`identity` Dit is een speciaal geval: er mag geen functienaam gegeven worden en er moet exact één argument gegeven worden. Dat ene argument zal de returnwaarde van de functie zijn. In de implementaties wordt dit vaak ook niet vertaalt als een functie, maar wordt de waarde rechtstreeks gebruikt. De bestaansreden van dit soort functies is het toekennen van waarden aan variabelen, om redenen die we later zullen bespreken.

Een beperking is dat het niet mogelijk is om rechtstreeks een functieoproep te doen als argument voor een andere functie, of toch niet op een programmeertaalonafhankelijke manier. Een oproep als `oproep(hallo(), 5)` is niet mogelijk. Bij dergelijke dingen zal de functieoproep eerst aan variabele moeten toegekend worden, bv. `var param = hallo()`, waarna deze variabele als argument met type `literal` kan gegeven worden aan de oorspronkelijke functie: `oproep(param, 5)`. De aandachtige lezer zal opmerken dat met die functieargumenten van het type `literal` rond deze beperking kan gewerkt worden, aangezien de tekstuele waarde van een dergelijk argument letterlijk in de taal komt. We raden deze omweg echter ten sterkste af: dit maakt het testplan taalafhankelijk, want niet elke programmeertaal implementeert functieoproepen op eenzelfde wijze.

Dit brengt ons bij de variabeletoekenning of *assignment*. In ons testplan beperkt dit zich tot het toekennen van een naam aan het resultaat van een functieoproep. Dit is ook meteen de reden voor het bestaan van de functiesoort `identity`: via deze weg blijft het testplan eenvoudiger (de waarde van een assignment is altijd een functieoproep), maar toch kunnen gewoon waarden toegekend worden aan een variabelen. Concreet ziet een variabeletoekenning er als volgt uit:

```
{
  "name": "Naam van de variabele",
  "expression": "<Object voor functieoproep>",
  "type": "Optioneel type"
}
```

De `name` is de naam die aan de variabele gegeven zal worden. Het veldje `expression` moet een object zijn dat een functieoproep voorstelt, in het formaat zoals hiervoor besproken. In een beperkt aantal gevallen kan de judge het type van de variabele afleiden uit de functieoproep, maar in veel gevallen is het nodig om zelf het type mee te geven. Dit type moet een van de ondersteunde types zijn uit het serialisatieformaat, zij het dat er ondersteuning is voor eigen types (zoals een klasse die geïmplementeerd moest worden door de student).

Een gecombineerd voorbeeld staat hieronder. Hier wordt de string `'Dodona'` toegekend aan een variabele met naam `name`. De judge kan het type afleiden, dus we moeten niet opgeven dat `name` een `str` is.

```
{
  "name": "name",
  "expression": {
    "type": "identity",
    "arguments": [
      {
        "type": "text",
        "value": "Dodona"
      }
    ]
  }
}
```

2.2.4. Vereiste functies

We hebben in het overzicht al vermeld dat voor de evaluatie gecontroleerd wordt of een testplan de programmeertaal van gegeven oplossing ondersteunt. Concreet gebeurt dit door voor elk item in het testplan af te leiden welke functies nodig zijn in een programmeertaal om dat item te ondersteunen. Bevat een testplan bijvoorbeeld waarden met als type `set` (de verzameling), dan kunnen enkel programmeertalen die een verzameling ondersteunen gebruikt worden. Dat zijn bijvoorbeeld Python en Java, maar geen Bash. Het afleiden van de vereiste functies gebeurt volledig automatisch aan de hand van het testplan.

2.3. Uitvoeren van de oplossing

De eerste stap die wordt uitgevoerd bij de evaluatie van de oplossing is het genereren van de code, die de code van de student zal uitvoeren.

2.3.1. Genereren van code

Het genereren van de code gebeurt met een sjabloonsysteem genaamd Mako (Bayer e.a. 2020). Dit soort systemen wordt traditioneel gebruikt bij webapplicaties (zoals Ruby on Rails met ERB, Phoenix met EEX, Laravel met Blade, enz.) om html-pagina's te genereren. In ons geval zijn de sjablonen verantwoordelijk voor de vertaling van de programmeertaalonafhankelijke concepten in het testplan naar eigenlijke code in de programmeertaal van de oplossing. Hierbij denken we aan de functieoproepen, assignments, enz. Ook zijn de sjablonen verantwoordelijk voor het genereren van de code die de oplossing van de student zal oproepen en evalueren.

Sjablonen

Het aantal sjablonen en hoe ze geïmplementeerd worden is in principe vrij, zij het dat de judge wel enkele standaardsjablonen nodig heeft, waaraan vastgelegde parameters meegegeven worden. Deze verplichte sjablonen zijn:

`assignment` Vertaalt een assignment uit het testplan naar code.

`context` Het sjabloon dat de code genereert om een context te evalueren. Deze code moet uitvoerbaar zijn (d.w.z. een main-functie bevatten of een script zijn).

`selector` Het sjabloon de code genereert om een bepaalde context uit te voeren. Om performantieredenen (later hierover meer) wordt de code van alle contexten soms uit een keer gegenereerd en gecompileerd. Aan de hand van een parameter (de naam van de context), wordt bij het uitvoeren de code voor de juiste context gekozen.

`evaluator_executor` Genereert code om een aangepaste evaluator te starten.

`function` Vertaalt een functie-oproep naar code.

`value` Vertaalt een waarde uit het serialisatieformaat naar code.

Daarnaast moet het encoderen naar het serialisatieformaat ook geïmplementeerd worden in elke taal. Veel talen hebben dus nog enkele bijkomende bestanden met code. In alle bestaande implementaties is dit geïmplementeerd als een module of klasse met naam `Value`.

Modus

De judge ondersteunt twee uitvoeringsmodi:

Precompilatiemodus In deze modus wordt de code voor alle contexten in een keer gecompileerd. Dit wordt gedaan om performantieredenen. In talen die resulteren in een uitvoerbaar bestand (zoals Haskell, C/C++), resulteert deze modus in één uitvoerbaar bestand voor alle contexten. Bij het uitvoeren wordt dan aan de hand van een parameter de juiste context uitgevoerd (met het `selector`-sjabloon van hierboven).

Individuele modus Hierbij wordt elke context afzonderlijk gecompileerd.

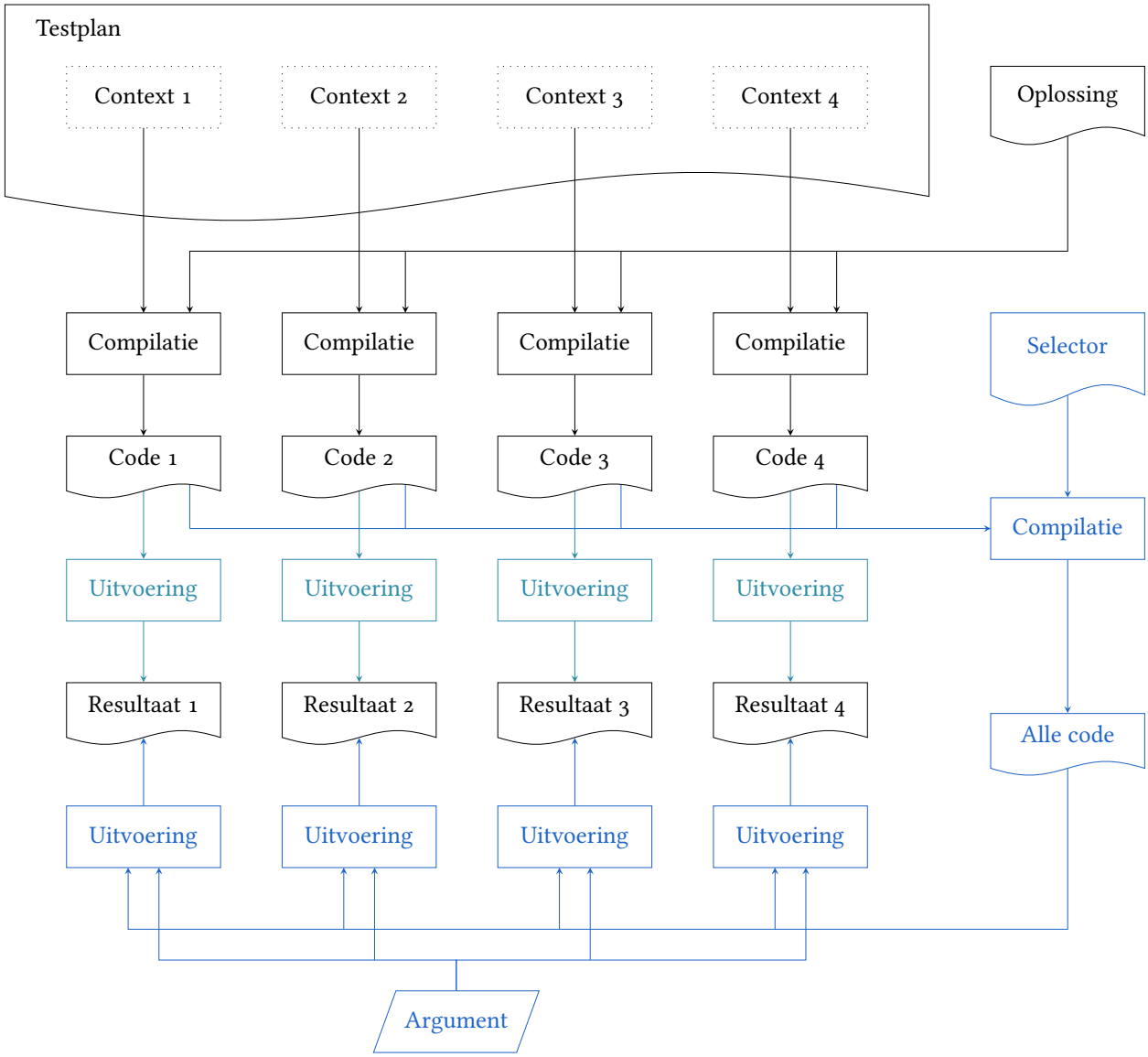
Een flowchart van het generen van de code is figuur 2.2, met een testplan met vier contexten. In dit diagram zijn de stappen voor de individuele modus in het aquablauw. De stappen van de precompilatiemodus zijn in het UGent-blauw. De gemeenschappelijke stappen zijn in het zwart.

Dit gedrag is configureerbaar in het testplan, maar standaard wordt de precompilatiemodus gebruikt, zij het met terugval op de individuele modus. Deze terugval is handig voor talen met sterke compilatie. Een voorbeeldscenario is als volgt: stel een oefening waarbij de student twee functies moet implementeren. De student implementeert de eerste functie en dient in om al feedback te krijgen. Bij talen als Java of Haskell zal dit niet lukken: daar alle contexten in een keer gecompileerd worden, zal de ontbrekende tweede functie ervoor zorgen dat de volledige compilatie faalt. In individuele modus is dit geen probleem: de contexten die de eerste functie testen zullen compileren en kunnen uitgevoerd worden. De individuele modus brengt wel een niet te onderschatten kost qua uitvoeringstijd met zich mee (zie ook hoofdstuk 5).

2.3.2. Uitvoeren van de code

Na het genereren wordt alle code gecompileerd (bij de talen waar dit mogelijk is). Dit gebeurt ofwel eenmaal voor alle contexten afzonderlijk, ofwel eenmaal voor alle contexten samen, afhankelijk van de modus. De werking hiervan wordt behandeld in paragraaf 2.3.1 en figuur 2.2.

Vervolgens wordt elke context uit het testplan uitgevoerd en wordt de uitvoer verzameld. Het uitvoeren zelf gebeurt op de normale manier dat een programmeertaal uitgevoerd wordt: via de



Figuur 2.2.: Schematische voorstelling van het genereren van de code. Gemeenschappelijke stappen zijn zwart, stappen voor de individuele modus aquablauw en stappen voor de precompilatiemodus UGent-blauw.

commandoregel. Deze aanpak heeft een voordeel: er is geen verschil tussen hoe de judge de code van de student uitvoert en hoe de student zijn code zelf uitvoert op zijn eigen computer. Dit voorkomt dat er subtiele verschillen in de resultaten sluipen.

Indien de configuratie het toelaat, worden de contexten in parallel uitgevoerd. Om te vermijden dat bestanden of uitvoer overschreven wordt, wordt de gecompileerde code gekopieerd naar een aparte map, waar de uitvoer gebeurt. Codefragment 2.2 illustreert dit met een voorbeeld voor een oplossing in Java. Deze mapstructuur stelt de toestand van de werkmap voor na het uitvoeren van de code. In de map `common` zit alle code en de gecompileerde bestanden. Voor elke context worden de gecompileerde bestanden gekopieerd naar een andere map, bv. `context-1`, wat de map is voor context `1` van het testplan.

2.3.3. Verzamelen van resultaten

De uitvoering van een oplossing genereert resultaten die door de judge geïnterpreteerd moeten worden. Er zijn verschillende soorten uitvoerresultaten (zoals vermeld heeft elke soort uitvoer een aparte test in het testplan). We noemen de verschillende soorten uitvoer de *uitvoerkanalen*. Twee ervan, het standaarduitvoer- en standaardfoutkanaal komen overeen met de standaarduitvoer- en standaardfoutstroom van het proces dat de code uitvoert. Uitvoer naar een bestand (het bestandskanaal) resulteert in een bestand en vormt ook geen probleem. De overige uitvoerkanalen, het kanaal voor exceptions (uitzonderingenkanaal) en het returnkanaal (voor returnwaarden) worden geschreven naar een bestand. Het is namelijk niet in elke taal mogelijk om nieuwe kanalen te openen. De sjablonen krijgen de verwachte namen van die bestanden mee van de judge, maar zijn wel verantwoordelijk voor het openen, schrijven en sluiten van deze bestanden. Deze naam bevat willekeurige tekens, zodat de kans dat deze bestanden overschreven worden door de oplossing minimaal is. De bestanden waar de exceptions and returnwaarden naartoe geschreven worden, worden na de uitvoering gelezen door de judge. Hierna worden alle kanalen op dezelfde manier behandeld door de judge.

In de bestaande implementaties ligt de verantwoordelijkheid om naar deze bestanden te schrijven bij de `Value`-module van hierboven.

2.4. Evalueren van een oplossing

Na de uitvoering van elke context heeft de judge alle relevant uitvoer verzamelt, zoals de standaardkanalen. Deze uitvoer moet vervolgens beoordeeld worden om na te gaan in hoeverre deze uitvoer voldoet aan de verwachte uitvoer. Dit kan op drie manieren:

1. Ingebouwde evaluator: de oplossing wordt geëvalueerd in de judge zelf.
2. Aangepaste evaluator: de oplossing wordt geëvalueerd door eigen code, maar dezelfde wordt gebruikt voor alle programmeertalen, in het evaluatieproces.
3. Taalspecifieke evaluator: de oplossing wordt onmiddellijk na de uitvoering geëvalueerd in het uitvoeringsproces.

```
workdir                          //Werkmap van de judge
├── common                       //Gemeenschappelijke code
│   ├── context_0_0.py           //Broncode voor context 0-0
│   ├── context_0_0.pyc          //Gecompileerde code voor context 0-0
│   ├── context_0_1.py
│   ├── context_0_1.pyc
│   ...
│   ├── context_0_49.py
│   ├── context_0_49.pyc
│   ...
│   ├── context_1_49.py
│   ├── context_1_49.pyc
│   ├── submission.py            //Code van de student
│   ├── submission.pyc
│   ├── values.py                //Values-module
│   └── values.pyc
├── context_0_0                  //Map voor context 0-0
│   ├── FaLd6WGRN_exceptions.txt //Uitzonderingskanaal
│   ├── FaLd6WGRN_values.txt     //Kanaal voor returnwaarden
│   ├── context_0_0.pyc          //Code voor context 0-0
│   ├── submission.pyc
│   └── values.pyc
...
└── evaluators                   //Aangepaste evaluatoren
    ├── buzzchecker_Aao9H18Ve    //Map voor elke context
    │   ├── buzzchecker.py       //Code aangepaste evaluator
    │   ├── evaluation_utils.py
    │   ├── evaluator_executor.py
    │   └── values.py
    ├── buzzchecker_B5WViK0zQ
    │   ├── buzzchecker.py
    │   ├── evaluation_utils.py
    │   ├── evaluator_executor.py
    │   └── values.py
    ...
```

Codefragment 2.2.: Mapstructuur na het uitvoeren van de evaluatie van een oplossing in Python. Context 0-0 staat voor de eerste context van het eerste tabblad.

2.4.1. Ingebouwde evaluator

Voor eenvoudige evaluaties volstaat de ingebouwde evaluator van de judge. Momenteel zijn er drie soorten ingebouwde evaluatoren, die hieronder besproken worden.

Tekstevaluator

Deze evaluator vergelijkt de verkregen uitvoer van een uitvoerkanaal (standaarduitvoer, returnwaarde, ...) met de verwachte uitvoer uit het testplan. Alle data worden als string behandeld. Deze evaluator biedt enkele opties om het gedrag aan te passen:

`ignoreWhitespace` Witruimte voor en na het resultaat wordt genegeerd.

`caseInsensitive` Er wordt geen rekening gehouden met het verschil tussen hoofdletters en kleine letters.

`tryFloatingPoint` De waarde moet geïnterpreteerd worden als een zwevendekommagetal (*floating point*), waarbij rekening gehouden wordt met de foutmarge.

`applyRounding` Of zwevendekommagetallen afgrond moeten worden. Indien wel wordt het aantal cijfers genomen van de optie `roundTo`.

`roundTo` Het aantal cijfers na de komma. Enkel nuttig als `applyRounding` waar is.

Bestandsevaluator

Hiermee kan een geproduceerd bestand vergeleken worden met een gegeven bestand uit het testplan. Het gaat om tekstuele bestanden. Deze evaluator kan werken in drie modi:

`exact` Beide bestanden moet exact hetzelfde zijn, inclusief regeleindes.

`lines` Elke regel wordt vergeleken met overeenkomstige regel in het andere bestand. De evaluatie van de lijnen is exact, maar zonder de regeleindes.

`values` Elke regel wordt geïnterpreteerd als een tekstuele waarde en vergeleken met de tekstevaluator. In deze modus worden kunnen ook alle opties van de tekstevaluator gebruikt worden.

Waarde-evaluator

Deze evaluator vergelijkt twee waarden, zoals gedefinieerd door het serialisatieformaat. De twee waarden moeten exact overeenkomen, met uitzondering van zwevendekommagetallen.

```python
def evaluate(expected, actual, arguments):
    """
    :param expected: The expected value from the testplan.
    :param actual: The actual value produced by the student's code.
    :param arguments: Arguments from the testplan.
    """
    pass
```

Codefragment 2.3.: De definitie van de aangepaste evaluator.

2.4.2. Aangepaste evaluator

Voor de aangepaste evaluator moet een bestand geschreven worden in een programmeertaal naar keuze. Het resultaat van de uitvoering wordt vervolgens geserialiseerd en gedeserialiseerd naar het evaluatieproces. Hoe een evaluator moet geïmplementeerd worden, hangt af van de programmeertaal.

In Python bestaat de aangepaste evaluator uit een module met een functie die voldoet aan de definitie, zoals gegeven in codefragment 2.3. De judge stelt ook een module `evaluation_utils` ter beschikking. De functie van hierboven moet dan één oproep doen naar de functie `evaluated()`. Deze module is redelijk eenvoudig, zoals te zien in codefragment 2.4.

In de Java-implementatie is de situatie gelijkaardig: het gaat om het implementeren van een abstracte klasse, die ook dienst doet als de module van Python. Deze klassen en haar ouder staan in codefragmenten 2.5 en 2.6.

Taalspecifieke evaluator

De taalspecifieke evaluator lijkt sterk op de aangepaste evaluator. is de eenvoudigste: deze neemt een codefragment met daarin één functie `evaluate`, die één argument aanvaardt, de geproduceerde waarde. Waar de geproduceerde waarde bij de aangepaste evaluator in het serialisatieformaat moet kunnen, is dit hier niet het geval: de functie wordt rechtstreeks opgeroepen tijdens de uitvoering. In Python wordt dit codefragment 2.7, in Java codefragment 2.8. Om het resultaat van de evaluatie aan de judge te geven, wordt dezelfde `evaluated`-functie als bij de aangepaste evaluator gebruikt (respectievelijk codefragmenten 2.4 en 2.5)

```python
import values
import sys

from typing import List, Optional


def evaluated(result: bool,
              readable_expected: Optional[str] = None,
              readable_actual: Optional[str] = None,
              messages: Optional[List[str]] = None):
    """
    Report the result of an evaluation to the judge. This method should only
    be called once, otherwise things will break.

    :param messages: Optional list of messages to be shown to the student.
    :param readable_actual: A string version of the actual value. Optional; if
                            not given, the judge will produce one on a best-
                            efforts basis.
    :param readable_expected: A string version of the expected value. Optional;
                              if not given, the judge will produce one on a
                              best-efforts basis.
    :param result: The result of the evaluation.
    """
    if messages is None:
        messages = []

    values.send_evaluated(sys.stdout,
                          result, readable_expected, readable_actual, messages)
```

Codefragment 2.4.: De implementatie van de module `evaluation_utils`

```java
import java.io.Closeable;
import java.io.IOException;
import java.io.OutputStreamWriter;
import java.util.Collection;
import java.util.List;

abstract class AbstractEvaluator implements Closeable {

    protected final OutputStreamWriter writer;

    public AbstractEvaluator() {
        this.writer = new OutputStreamWriter(System.out);
    }

    @Override
    public void close() throws IOException {
        this.writer.close();
    }

    /**
     * Report the result of an evaluation to the judge. This method should only
     * be called once, otherwise things will break.
     *
     * @param result           The result of the evaluation.
     * @param readableExpected Optional string version of the expected value.
     * @param readableActual   Optional string version of the actual value.
     * @param messages         Optional list of messages to pass to the student.
     */
    protected void evaluated(boolean result, String readableExpected,
                             String readableActual,
                             Collection<String> messages) throws IOException {
        Values.evaluated(writer,
                result, readableExpected, readableActual, messages);
    }

    protected void evaluated(boolean result, String readableExpected,
                             String readableActual) throws IOException {
        Values.evaluated(writer,
                result, readableExpected, readableActual, List.of());
    }
}
```

Codefragment 2.5.: De implementatie van de klasse `AbstractEvaluator`.

```java
import java.io.IOException;
import java.util.List;

abstract class AbstractCustomEvaluator extends AbstractEvaluator {

    abstract void evaluate(Object expected,
                           Object actual,
                           List<Object> arguments) throws IOException;
}
```

Codefragment 2.6.: De implementatie van de klasse `AbstractCustomEvaluator`.

```python
def evaluate(actual):
    """
    :param actual: The actual value produced by the student's code.
    """
    pass
```

Codefragment 2.7.: De definitie van de taalspecifieke evaluator.

```java
import java.io.IOException;
import java.util.List;

abstract class AbstractSpecificEvaluator extends AbstractEvaluator {

    abstract void evaluate(Object actual) throws IOException;
}
```

Codefragment 2.8.: De implementatie van de klasse `AbstractSpecificEvaluator`.

3. Case-study: configureren van een programmeertaal in TESTed

Allerlei uitleg

4. Case-study: nieuwe oefening

In dit hoofdstuk behandelen we het toevoegen van drie oefeningen in handleidingsstijl. Elke oefening gebruikt een andere evaluatiemethode uit paragraaf 2.4. We gaan er van uit dat de map voor de repo voor deze oefeningencollectie al bestaat. Deze moet voldoen aan de mappenstructuur voor oefeningen, opgelegd door Dodona[1].

4.1. ISBN

Bij de eerste oefening gebruiken we enkel ingebouwde evaluators.

4.1.1. Voorbereiding

We beginnen met het maken van een nieuwe map `isbn` in onze map voor de oefeningencollectie. Vervolgens maken we in deze map een configuratiebestand `config.json` voor onze oefening, met deze inhoud:

```
{
  "description": {
    "difficulty": 2.0,
    "names": {
      "en": "ISBN",
      "nl": "ISBN"
    }
  },
  "access": "public",
  "programming_language": "python",
  "evaluation": {
    "plan_name": "plan.json"
  }
}
```

In dit bestand doen we drie belangrijke dingen:

1. We geven onze oefening een naam.
2. We duiden aan dat het om een Python-oefening gaat (zie hoofdstuk 5 voor waarom we dit doen).
3. We zeggen dat ons testplan als naam `plan.json` heeft.

[1]Hier beschikbaar: https://dodona-edu.github.io/en/references/exercise-directory-structure/

4.1.2. Opgave

In deze stap stellen we de opgave op. Om dit deel kort te houden, gaan we er van uit dat de opgave al bestaat. Maak een map `description` in onze oefeningenmap, en kopieer de opgavebestanden naar deze map. Uiteindelijk moet deze map er zo uitzien:

```
isbn
└── description
    ├── description.en.html
    ├── description.nl.html
    └── media
        └── ISBN.gif
```

Om toch een idee te krijgen van waarover de oefening gaat, is hieronder een samenvatting van de opgave:

- Schrijf een functie `is_isbn` waaraan een string c (`str`) moet doorgegeven worden. De functie moet een Booleaanse waarde (`bool`) teruggeven, die aangeeft of c een geldige ISBN-code is. De functie heeft ook nog een optionele tweede parameter `isbn13` waaraan een Booleaanse waarde (`bool`) kan doorgegeven worden die aangeeft of het om een ISBN-10 code (`False`) of om een ISBN-13 code (`True`, standaardwaarde) moet gaan.
- Schrijf een functie `are_isbn` waaraan een lijst (`list`) met $n \in \mathbb{N}$ codes moet doorgegeven worden. De functie moet voor alle codes uit de gegeven lijst aangegeven of ze geldige ISBN-codes voorstellen. De functie heeft ook nog een tweede optionele parameter `isbn13` waaraan een Booleaanse waarde (`bool`) kan doorgegeven worden die aangeeft of het om ISBN-10 codes (`False`) of om ISBN-13 codes (`True`) moet gaan.
  Als er niet expliciet een waarde wordt doorgegeven aan de parameter `isbn13`, dan moet het type van elke code uit de lijst bepaald worden op basis van de lengte van die code. Als een code geen string (`str`) is, dan wordt die *a priori* als ongeldig bestempeld. Voor codes van lengte 13 moet getest worden of het geldige ISBN-13 codes zijn, en voor codes van lengte 10 of het geldige ISBN-10 codes zijn. Codes met afwijkende lengtes (geen 10 en geen 13) worden ook *a priori* als ongeldige ISBN-codes bestempeld.
  De functie moet een nieuwe lijst (`list`) met n Booleaanse waarden (`bool`) teruggeven, die aangeven of de code op de corresponderende positie in de gegeven lijst een geldige ISBN-code is.

4.2. Testplan

Nu komen we aan het belangrijkste deel van het opstellen van een oefening: het testplan. Om te beginnen kunnen we even nadenken over hoe de structuur er zal uitzien. We hebben twee functies die getest moeten worden. Laten we de in- en uitvoer voor deze functies X_i en X_o noemen. Aangezien elke functieoproep onafhankelijk is, kiezen we ervoor om ze in aparte contexten van het testplan te steken. Verder kunnen we al zeggen dat elke context juist één testgeval zal hebben: de functieoproep. Elk testgeval zal ook maar één test hebben, waar we de returnwaarde controleren.

Alle andere uitvoerkanalen mogen de standaardwaarde behouden (wat betekent dat er bv. geen exceptions mogen zijn, een lege standaardfoutstroom, de standaarduitvoerstroom wordt genegeerd, enz.). We verdelen de contexten bovendien in twee tabbladen, een voor elke functies. De structuur van het testplan ziet er dus als volgt uit:

- Tabblad voor `is_isbn`
  - Context voor invoer X_{i1} en uitvoer X_{o1}
  - Context voor invoer X_{i2} en uitvoer X_{o2}
  - Context voor invoer X_{i3} en uitvoer X_{o3}
  - Context voor invoer X_{i4} en uitvoer X_{o4}
  - ...
- Tabblad voor `are_isbn`
  - Context voor invoer X_{i1} en uitvoer X_{o1}
  - Context voor invoer X_{i2} en uitvoer X_{o2}
  - Context voor invoer X_{i3} en uitvoer X_{o3}
  - Context voor invoer X_{i4} en uitvoer X_{o4}
  - ...

In plaats van dit testplan manueel te schrijven, kiezen we ervoor om een Python-scriptje te schrijven, dat dit testplan voor ons genereert. Om te beginnen:

- Kopieer `solution.py` naar de map `solution`. Dit is de voorbeeldoplossing.
- Kopieer `values.py` naar de map `preparation`. Deze module encodeert waarden in het serialisatieformaat. In de toekomst is dit misschien niet meer nodig, moest dit als Python-package gepubliceerd worden.

Maak nu een bestand in de map `preparation` genaamd `generator.py`. Hieronder volgt het script dat het testplan genereert. Het is rijkelijk voorzien van commentaar, zodat alles duidelijk is.

```
import random
import string
import json

# De implementatie van het serialisatieformaat.
from isbn.preparation import values
# De voorbeeldoplossing.
from isbn.solution import solution

# Met een vaste seed krijgen we deterministische resultaten.
random.seed(123456789)

# We halen de naam van het testplan op uit de configuratie van de oefening.
with open("../config.json", "r") as config_file:
    config = json.load(config_file)
testplan_name = config["evaluation"].get("plan_name", "plan.json")
```

```python

# Of alle testgevallen in dezelfde context moeten plaatsvinden of niet.
# Aangezien ze onafhankelijk zijn, doen we dit niet.
# Meerdere contexten hebben echter wel een performantiekost, waardoor we de
# optie toch voorzien.
ONE_CONTEXT = False


def check_digit10(code):
    """Bereken het controlecijfer voor een ISBN van lengte 10."""
    check = sum((i + 1) * int(code[i]) for i in range(9)) % 11
    # Zet het controlecijfer om naar een string.
    return 'X' if check == 10 else str(check)


def check_digit13(code):
    """Bereken het controlecijfer voor een ISBN van lengte 13."""
    check = sum((3 if i % 2 else 1) * int(code[i]) for i in range(12))
    # Zet het controlecijfer om naar een string.
    return str((10 - check) % 10)


def random_characters(length, alphabet):
    """Genereer een aantal willekeurige tekens uit een alfabet."""
    return ''.join(random.choice(alphabet) for _ in range(length))


def generate_code():
    """Genereer een ISBN-code, met zowel lengte 10 als 13."""
    length = random.choice([10, 13])
    code = random_characters(length - 1, string.digits)
    if length == 10:
        if random.random() < 0.5:
            code += check_digit10(code)
        else:
            code += random.choice(string.digits + 'X')
    else:
        if random.random() < 0.5:
            code += check_digit13(code)
        else:
            code += random.choice(string.digits)
    return code


def generate_is_isbn():
    """
    Genereer de contexten voor de "is_isbn"-functie.
    :return: De gegenereerde contexten.
    """
```

```
    contexts = []

    # Genereer eerst de argumenten voor de functie "is_isbn".
    # We beginnen met wat vaste combinaties.
    args = [
        ('9789027439642', False),
        ('9789027439642', True),
        ('9789027439642', None),
        ('080442957X', None),
        ('080442957X', False),
        (9789027439642, None),
    ]
    # Voor de rest vullen we aan met willekeurige argumenten.
    while len(args) < 50:
        code = generate_code()
        args.append((code, random.choice([None, True, False])))

    # Genereer de eigenlijke contexten.
    for code, isbn13 in args:
        # Eerst doen we de functieoproep. We geven zeker de ISBN mee.
        function_arguments = [
            values.encode(code)
        ]
        # Indien nodig geven we ook het tweede argument mee.
        if isbn13 is not None:
            function_arguments.append(values.encode(isbn13))

        # Bereken het resultaat met de gegeven argumenten.
        result = solution.is_isbn(code, isbn13 if isbn13 is not None else True)

        # Ons testgeval bevat de functieoproep als invoer, en de berekende waarde
        # als verwachte uitvoer.
        testcase = {
            "input":  {
                "function": {
                    "type":      "top",
                    "name":      "is_isbn",
                    "arguments": function_arguments
                }
            },
            "output": {
                "result": {
                    "value": (values.encode(result))
                }
            }
        }

        # Steek het testgeval in een context.
        context = {
```

```
            "normal": [testcase]
        }
        contexts.append(context)

    return contexts


def generate_are_isbn():
    """
    Genereer de contexten voor de "are_isbn"-functie.
    :return: De gegenereerde contexten.
    """

    # Vaste invoerargumenten die we zeker in het testplan willen.
    codes = [
        '0012345678', '0012345679', '9971502100', '080442957X', 5, True,
        'The Practice of Computing Using Python', '9789027439642', '5486948320146'
    ]
    codes2 = ['012345678' + str(digit) for digit in range(10)]
    args = [
        (codes, None),
        (codes, True),
        (codes, False),
        (codes2, None),
        (codes2, True),
        (codes2, False),
    ]
    # Vul opnieuw de rest aan tot we aan 50 zitten.
    while len(args) < 50:
        codes = [generate_code() for _ in range(random.randint(4, 10))]
        args.append((codes, random.choice([None, True, False])))

    # Genereer de eigenlijke contexten.
    contexts = []
    for index, (codes, isbn13) in enumerate(args):
        index += 1
        # Deze keer pakken we het iets anders aan: de lijst van ISBN's kennen we
        # eerst toe aan een variabele, en geven het niet rechtstreeks mee als
        # argument. Maak het testgeval voor de assignment.
        assignment_testcase = {
            "input": {
                "assignment": {
                    "name": f"codes{index:02d}",
                    "expression": {
                        "type": "identity",
                        "arguments": [
                            values.encode(codes)
                        ]
                    }
```

```
                }
            }
        }

        # Stel de functieargumenten op. We geven opnieuw sowieso de variabele die
        # we eerst hebben aangemaakt mee als argument.
        function_arguments = [
            {
                "type": "literal",
                "data": f"codes{index:02d}"
            }
        ]

        # Voeg het tweede argument toe indien nodig.
        if isbn13 is not None:
            function_arguments.append(values.encode(isbn13))

        # Bereken het resultaat.
        result = solution.are_isbn(codes, isbn13 if isbn13 is not None else None)

        # Maak het normale testgeval. We hebben opnieuw als invoer de functie en als
        # uitvoer de verwachte waarde.
        testcase = {
            "input":  {
                "function": {
                    "type":      "top",
                    "name":      "are_isbn",
                    "arguments": function_arguments
                }
            },
            "output": {
                "result": {
                    "value": (values.encode(result))
                }
            }
        }

        # Voeg beide testcases toe aan een context.
        # Elke testcase heeft hoogstens één functieoproep of assignment, maar niet
        # beide. Daarom hebben we nu twee testgevallen. Het eerste testgeval voor
        # de assignment heeft enkel de standaardtests, wat wil zeggen dat er bv.
        # geen uitvoer op stderr mag zijn.
        context = {
            "normal": [assignment_testcase, testcase]
        }

        contexts.append(context)
    return contexts

```

```python

def flatten_contexts(contexts):
    """Voeg alle testgevallen samen tot 1 context."""
    testcases = [context["normal"] for context in contexts]
    flat = [item for sublist in testcases for item in sublist]
    new_context = {
        "normal": flat
    }
    return new_context


# Creëer de contexten.
tab_1_contexts = generate_is_isbn()
tab_2_contexts = generate_are_isbn()

# Creëer het testplan met de twee tabbladen.
plan = {
    "tabs": [
        {
            "name": "is_isbn",
            "contexts": tab_1_contexts
        },
        {
            "name": "are_isbn",
            "contexts": tab_2_contexts
        }
    ]
}

# Indien we alles in één context willen, doen we dat.
if ONE_CONTEXT:
    new_tab1_context = flatten_contexts(plan["tabs"][0]["contexts"])
    plan["tabs"][0]["contexts"] = [new_tab1_context]
    new_tab2_context = flatten_contexts(plan["tabs"][1]["contexts"])
    plan["tabs"][1]["contexts"] = [new_tab2_context]

# Terugvallen op individuele modus is niet nuttig in Python, dus laten we dat niet
# toe. Indien het terugvallen niet nuttig is, is het sneller om het uit te zetten.
plan["configuration"] = {
    "allow_fallback": False
}

# Schrijf het testplan.
with open(f"../evaluation/{testplan_name}", 'w') as fp:
    json.dump(plan, fp, indent=2)
```

```
isbn
├── config.json
├── description
│   ├── description.en.html
│   ├── description.nl.html
│   └── media
│       └── ISBN.gif
├── evaluation
│   └── plan.json      # Gegenereerd door generator.py
├── preparation
│   ├── generator.py
│   └── values.py
└── solution
    └── solution.py
```

Codefragment 4.1.: Finale mapstructuur van de oefening ISBN.

4.2.1. Afsluiting

Op dit moment is de oefening klaar. Ter controle: de complete mapstructuur van de uiteindelijke oefening wordt nog eens getoond in codefragment 4.1. Nu rest nog enkel de oefening importeren in Dodona. Is de oefening toegevoegd aan een bestaande repo, dan volstaat het om op de knop „Alle oefeningen opnieuw verwerken" te drukken. Gaat het om een nieuwe oefeningenrepo, dan moet deze toegevoegd worden[2].

[2]De handleiding hiervoor staat op https://dodona-edu.github.io/en/guides/new-exercise-repo/

5. Beperkingen en toekomstig werk

Wat kunnen we al en vooral wat niet? Waar kan nog aan gewerkt worden?

Korte samenvatting

5.1. Performance

-> Uitleg over eerste implementatie met jupyter kernels

-> Uitleg over verschillende stadia van codegeneratie (alles apart -> zoveel mogelijk samen)

5.2. Functies

-> Dynamisch testplan –> Dingen meerdere keren uitvoeren –> Dingen wel of niet uitvoeren op basis van vorige uitkomst

-> Functies/assignments –> Functie als functie-argumenten zonder tussenstap met assignments

-> Meertaligheid

-> Beschrijving van de oefening (programmeertalen)

A. Specificatie van het serialisatieformaat

TODO: hier misschien json-schema?

Bibliografie

Amazon Ion (15 januari 2020). Amazon. URL: https://amzn.github.io/ion-docs/ (bezocht op 27-01-2020).

Apache Avro™ 1.9.1 Documentation (9 februari 2019). The Apache Foundation. URL: http://avro.apache.org/docs/1.9.1/ (bezocht op 27-01-2020).

Apache Parquet (13 januari 2020). The Apache Foundation. URL: https://parquet.apache.org/documentation/latest/ (bezocht op 27-01-2020).

Bayer, Michael e.a. (20 januari 2020). *Mako Templates for Python*. URL: https://www.makotemplates.org/.

Bormann, Carstenn en Paul Hoffman (oktober 2013). *Concise Binary Object Representation (CBOR)*. RFC 7049. Internet Engineering Task Force. URL: https://tools.ietf.org/html/rfc7049.

BSON 1.1 (19 juli 2019). MongoDB. URL: http://bsonspec.org/ (bezocht op 17-01-2020).

Cohen, Bram (4 februari 2017). *The BitTorrent Protocol Specification*. URL: http://bittorrent.org/beps/bep_0003.html.

Furuhashi, Sadayuki (17 september 2018). *MessagePack*. URL: https://msgpack.org/.

Information technology – Abstract Syntax Notation One (ASN.1): Specification of basic notation (augustus 2015). Recommendation X.608. International Telecommunications Union. URL: https://www.itu.int/rec/T-REC-X.680-201508-I/en.

Information technology – Generic applications of ASN.1: Fast infoset (14 mei 2005). Recommendation X.881. International Telecommunications Union. URL: https://www.itu.int/rec/T-REC-X.891-200505-I/en.

Jackson JSON team (2010). *Smile Data Format*. URL: https://github.com/FasterXML/smile-format-specification.

Lengyel, Eric (17 januari 2017). *Open Data Description Language (OpenDDL)*. URL: http://openddl.org/.

Mitra, Nilo en Yves Lafon (april 2007). *SOAP Version 1.2 Part 0: Primer (Second Edition)*. TR. http://www.w3.org/TR/2007/REC-soap12-part0-20070427/. W3C.

OGDL 2018.2 (25 februari 2018). URL: https://msgpack.org/ (bezocht op 28-01-2020).

Oortmerssen, Wouter van (24 april 2019). *FlatBuffers*. Google. URL: https://google.github.io/flatbuffers/.

OPC unified architecture - Part 1: Overview and concepts (10 mei 2016). IEC TR 62541-1:2016. International Electrotechnical Commission. URL: https://webstore.iec.ch/publication/25997.

ProtocolBuffers (13 december 2019). Google. URL: https://developers.google.com/protocol-buffers/.

Ramos, Bernardo (25 september 2019). *Binn*. URL: https://github.com/liteserver/binn.

Slee, Mark, Aditya Agarwal en Marc Kwiatkowski (2007). „Thrift: Scalable cross-language services implementation". In: URL: http://thrift.apache.org/static/files/thrift-20070401.pdf.

Universal Binary JSON (25 februari 2018). URL: http://ubjson.org/.

Wikipedia-bijdragers (25 januari 2020). *Comparison of data-serialization formats. Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=Comparison_of_data-serialization_formats&oldid=937433197.

Wildgrube, Max (maart 2001). *Structured Data Exchange Format (SDXF)*. RFC 3072. Internet Engineering Task Force. URL: https://tools.ietf.org/html/rfc3072.