Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

KATEDRA AUTOMATYKI

PRACA DYPLOMOWA INŻYNIERSKA

KONRAD MALAWSKI

PROTODOC
IMPLEMENTACJA ODPOWIEDNIKA NARZĘDZIA JAVADOC DLA JĘZKA DEFINICJI INTERFEJSÓW GOOGLE PROTOCOL BUFFERS

PROMOTOR:
dr inż. Jacek Piwowarczyk

Kraków 2012

OŚWIADCZENIE AUTORA PRACY

OŚWIADCZAM, ŚWIADOMY ODPOWIEDZIALNOŚCI KARNEJ ZA POŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZĄ PRACĘ DYPLOMOWĄ WYKONAŁEM OSOBIŚCIE I SAMODZIELNIE, I NIE KORZYSTAŁEM ZE ŹRÓDEŁ INNYCH NIŻ WYMIENIONE W PRACY.

. . . . . .

PODPIS

AGH

University of Science and Technology in Krakow

Faculty of Electrical Engineering, Automatics, Computer Science and Electronics

DEPARTMENT OF AUTOMATICS

BACHELOR OF SCIENCE THESIS

KONRAD MALAWSKI

PROTODOC
DEVELOPMENT OF A JAVADOC TOOL EQUIVALENT FOR THE GOOGLE PROTOCOL BUFFERS INTERFACE DESCRIPTION LANGUAGE

SUPERVISOR:
Jacek Piwowarczyk Ph.D

Krakow 2012

Spis treści

1. Wprowadzenie

Poniższy rozdział ma za cel przedstawienie czytelnikowi dziedziny oraz problemu jaki rozwiązuje zaimplementowany w ramach tej pracy inżynierskiej program - *ProtoDoc*. Następnie zostaną przedstawione rozdziały pracy oraz jakie tematy dany rozdział porusza.

1.1. Przedstawienie dziedziny zagadnienia

Google Protocol Buffers są zespołem narzędzi i binarnego protokołu służącymi do transportu danych w możliwie najefektywniejszy sposób. Można o nich myśleć w kategorii następcy takich formatów danych jak XML lub ostatnio popularny JSON. Główną przyczyną dlaczego Protocol Buffers, dalej zwanymi ProtoBuf, zaczynają powoli wypierać zastosowanie wcześniej wspomnianych formatów reprezentacji danych jest fakt wynikający bezpośrednio z części „binarny" w opisie ProtoBuf – wiadomości kodowane przy pomocy tego narzędzia, mogą być znacznie efektywniej kompresowane oraz szybciej rekonstruowane. W przypadku systemów „High Velocity" gdzie opóźnienia rzędu milisekund już są znaczące okazuje się, że nie ma już od dawna miejsca dla uznawanych dotychczas przez branżę standardów komunikacji, takich jak na przykład SOAP, parsowanie którego wiadomości z racji formatu danych, zwyczajnie nie jest w stanie sprostać wymaganiom wydajnościowym tych systemów. Jednym z przykładów największych instalacji ProtoBuf jest oczywiście samo Google, stosujące je w ramach komunikacji między większością swoich aplikacji. Protocol Buffers nie są same w ruchu wielkich firm ku binarnym protokołom - jako przykład może posłużyć Apache Thrift [Thr12], będący odpowiednikiem ProtoBuf, uwolnionego jako open source spod skrzydła Facebook.

Zasada działania tych protokołów jest prosta, oraz w pewien sposób zbliżona do tego co świat enterprise zna od lat, pod formą WSDL, to znaczy „kontraktów" między dostawcą serwisu oraz jego klientami. W Protocol Buffers taki kontrakt spisuje się przy pomocy specjalnego IDL (*Interface Description Language* - ang. Język opisu interfejsu) o tej samej nazwie. Język ten jest przedmiotem niniejszej pracy inżynierskiej. Definiowane przy jego pomocy typy wiadomości są następnie wykorzystywane przez kompilator *protoc*, celem generowania „potrafiących się przeczytać ze strumienia bajtów" klas, w dowolnym języku programowania. Pozwala to na automatyczne generowanie web serwisów, analogicznie jak miało to miejsce w przypadku generowania kodu z WSDL za czasów SOAP.

Miejscem w którym *ProtoDoc*, aplikacja powstała w ramach tej pracy inżynierskiej, znajduje swoją niszę, jest dokumentacja wiadomości Protocol Buffers. Niestety okazuje się, iż dokumentacja wiadomości Protocol Buffers staje się dużym problemem w korporacjach gdzie mamy do czynienia setkami

wiadomości, które zostały pierwotnie stworzone wiele lat temu. Niestety nie istnieją obecnie narzędzia generujące dokumentację na podstawie kodu ProtoBuf. Pomogłoby to rozproszonym dużym zespołom uniknąć nieporozumień dotyczących poprawnego wykorzystywania wiadomości. Narzędziem które może ten stan rzeczy zmienić, jest *ProtoDoc*.

1.2. Cel pracy

Celem projektu jest implementacja narzędzia generującego dokumentację na podstawie plików *.proto zawierających zapisane przy pomocy „języka definicji interfejsów" (tzw. *Interface Description Language*, w skrócie *IDL*) - Google Protocol Buffers.

Potrzebę implementacji takiego narzędzia motywuję doświadczeniem w pracy z Protocol Buffers, gdy mamy do czynienia z dużą ilością plików *.proto (setki). Brak automatycznie generowanej dokumentacji tak dużego zbioru wiadomości znacznie utrudniał zapoznanie się z systemem oraz przystąpienie do sprawnej pracy z nim. Gdyby taka, zawsze aktualna, dokumentacja była dostępna w firmowym intranecie na przykład, komunikacja między zespołami o wiadomościach byłaby znacznie prostsza - możliwe byłoby wówczas przesłanie sobie między programistami linku do właściwej wiadomości „to tej wiadomości szukasz", włącznie z upewnieniem się, że na pewno wskazana wiadomość nie jest przestarzała - zawierałaby wówczas odnośnik do wiadomości którą obecnie powinno się stosować.

Proces generowania dokumentacji jest analogiczny do znanego z świata Javy narzędzia JavaDoc [Jav12] - stąd zainspirowana JavaDociem nazwa tego projektu. Sam proces generowania dokumentacji polega na dostarczeniu narzędziu plików *.proto, które następnie są parsowane oraz na podstawie tego procesu, generowana jest strona www zawierające wszystkie zebrane informacje, włącznie z komentarzami oraz dodatkowymi informacjami typu „*deprecated*" (ang. przestarzałe). Jako dodatkowy krok wygenerowana strona mogłaby automatycznie zostać opublikowana w firmowym intranecie.

Cały proces możliwe jest w pełni zintegrować z narzędziami stosowanymi do budowania projektów np. Javowych. W przypadku projektów Javowych, obecnym *de facto* standardem w wielu firmach stał się Apache Maven [Mav12]. ProtoDoc może zostać użyty razem z Maven aby automatycznie, podczas budowania projektu generować dokumentację. Możliwe jest uruchomienie tego zadania samodzielnie, lub jako jeden z etapów budowy projektu - dzięki czemu nie konieczne jest pamiętanie oraz ręczne aktualizowanie dokumentacji - byłaby automatycznie generowana podczas buildu, na przykład na serwerze ciągłej integracji.

1.3. Zawartość pracy

W rozdziale 2 (*Analiza obecnie dostępnych rozwiązań*) zostaną przedstawione obecnie istniejące implementacje parserów Protocol Buffers. Poddane rozważaniom zostanie opłacalność skorzystania ze wspomnianych rozwiązań open source jako bazy tego projektu. Ostatecznie jednak przedstawione zostanie ostatecznie wybrane rozwiązanie implementacji parsera zastosowane w tym projekcie - kombinatory parserów języka Scala.

W rozdziale 3 (*Projekt systemu*) omówiona zostanie ogólna architektura aplikacji oraz jej podział na funkcjonalnie niezależne moduły. Następnie zostaną opisane odpowiedzialności każdego z modułów oraz czytelnik zostanie przeprowadzony przez proces tworzenia hierarchii typów służącej w całym projekcie to wiernego oraz praktycznego oddania typów oraz pól mogących pojawić się w opisie wiadomości Protocol Buffers.

W kolejnym Rozdziale 4 (*Szczegóły implementacyjne*) zostaną omówione najważniejsze szczegóły implementacyjne każdego z wprowadzonych w poprzednim rozdziale komponentów. Szczególny nacisk zostanie położony na właściwości języka Scala oraz wybory dokonane w związku z zakresem oraz ewentualną kontunuacj projektu.

Następnie w rozdziale 5 (*Przykład zastosowania ProtoDoc w realnym projekcie*) przedstawione zostanie jak w praktyce powinno się korzystać z zaimplementowanego w ramach tego projektu narzędzia. Przedstawiony zostanie również plugin do narzędzia Maven [Mav12] dzięki któremu pełna automatyzacja oraz włączenie ProtoDoc w proces budowania projektu jest bardzo proste.

W przedostatnim rozdziale, tj. 6 (*Rola testów w procesie tworzenia aplikacji*) przedstawiona zostanie metodyka Test Driven Development (ang. „Prowadzenie Projektu Testami" - tłumaczenie własne), dzięki której każda funkcjonalność przedstawiania w poniższej aplikacji pokryta została również automatycznym testem, mogącym potwierdzić poprawne funkcjonowanie aplikacji.

Ostatni rozdział 7 (*Podsumowanie*) podsumowuje pracę oraz przedstawiane są w nim możliwości dalszego rozwoju aplikacji.

Jako załączniki do pracy zostały dołączone dwa dodatki:

– w Dodatku A omawiane są Protocol Buffers – skąd i dlaczego powstały oraz przedstawia podstawowe zastosowanie ich w praktyce,

– w Dodatku B natomiast przedstawiany jest język Scala oraz Parser Combinators. Przedstawiony w tym dodatku zakres informacji o Scali powinien być wystarczający aby swobodnie czytać przykłady umieszczone w części implementacyjnej tej pracy.

2. Analiza obecnie dostępnych rozwiązań

Niestety na chwilę obecną nie są dostępne narzędzia pozwalające na generowanie dokumentacji z plików Protocol Buffers. *Analiza obecnych rozwiązań zatem organiczy się do rozważenia opłacalności wykorzystania jakiegoś projektu open source jako bazy dla ProtoDoc.*

Jak się okaże, najopłóacalniejsza z perspektywy programisty jak i użytkownika końcowego gotowej aplikacji ProtoDoc, będzie implementacja parsera, przy wykorzystaniu języka Scala, a nie wykorzystanie istniejących rozwiązań - które na przykład posiadają bardzo duże zewnętrzne zależności, lub ich dopasowanie do potrzeb tego projektu byłby zbyt dużym przedsięwzięciem.

Celem ułatwienia zrozumienia poniższego, oraz kolejnych rozdziałów w przypadku gdy czytelnik nie miał jeszcze styczności z Google Protocol Buffers zalecane jest wpierw zapoznanie się z *Dodatkiem A*, gdzie wyjaśniane jest dokładnie jak oraz dlaczego działają Protocol Buffers.

2.1. Google Protoc

Protoc jest „oryginalnym" kompilatorem plików *.proto. Zawiera ręcznie zaimplenentowany przez inżynierów google skaner oraz parser, potrafiący obsłużyć 100% specyfikacji ProtoBuf. Jego źródła są dostępne na stronie Google Code: http://code.google.com/p/protobuf/source/browse/ Projekt objęty jest licencją *New BSD License* [BSD12].

Warto również uwypuklić pewien problem z udostępnianym przez Google kompilatorem Protocol Buffers IDL - *protoc*. Otóż nawet jeżeli źródłowy plik *.proto posiada komentarze, kompilator *protoc* nie przeniesie je do wynikowych plików, np. *.java. Parser ten niestety ignoruje całkowicie komentarze.

Po wstępnej analizie kodu parsera dostarczanego przez Google doszedłem do wniosku, że niestety wykorzstanie go jako bazy ProtoDoc nie byłoby opłacalne, ze względu na bardzo dużą ilość zmian które trzeba by wprowadzić w *core* parsera - zaimplementowanego „ręcznie", bez zastosowania znanych generatorów parserów, w C++.

2.2. Idea plugin protobuf

Innym projektem open source zzawierającym zaimplementowany parser ProtoBuf jest plugin do „IntelliJ IDEA", popularnego w świecie programistów JVM IDE programistycznego. Źródła znajdują się na Google Code pod adresem: http://code.google.com/p/idea-plugin-protobuf/source/browse Projekt udostępniany jest na warunkach *Apache 2.0 License* [Apa12].

Z perspektywy ProtoDoc, interesującymi fragmentami tego projektu jest skaner oraz parser. Skaner jest generowany przy pomocy *JFlex* [1], , odpowiednika narzędzia GNU Flex, dla języka Java. Skaner teoretycznie nadawałby się do ponownego wykorzystania - obsługiwane są tutaj również komentarze.

Niestety druga z interesujących nas części aplikacji, parser, jest *ściśle związany ze środowiskiem IntelliJ IDEA*, dla którego to powstał ten projekt. IntelliJ dostarcza własny mechanizm parsowania do którego pluginy jedynie mogą się podpinać, oraz pomagać w przeprowadzeniu parsingu pliku, nie można w tym przypadku powiedzieć że projekt zawiera całą implementację parsera. Część źródeł IntelliJ IDEA co prawda jest otwarta, jednak skorzystanie z podejścia dołączenia całego IDE, aby być w stanie parsować pliki, wydaje się bardzo nie optymalna - rozmiar dystrybucji ProtoDoc stałby się bardzo duży (rzędu setek MB, z racji dołączonych zależności w postaci IntelliJ).

Jak widać, również i ten projekt nie dostarcza w pełni funkcjonalnej oraz łatwej to rozbudowania o potrzebne w projekcie ProtoDoc funkcjonalności implementacji parsera Protocol Buffers. W związku z powyższym, postanowiłem wybrać sposób własnoręcznej implementacji parsera, aby proces był jednak możliwie przyjemny, oraz możliwy do utrzymania w przyszłości - na przykład przez społeczność Open Source. Ostatecznie wybrana przezemnie technika implementacij parsera zostanie przedstawiona w kolejnej sekcji.

2.3. Decyzja: własnoręczna implementacji Parsera przy pomocy *Scala Parser Combinators*

Podsumowując, istnieją implementacje parserów Protocol Buffers na wolnych (jak wolność) licencjach, jednak rozbudowa ich o pożądane funkcjonalności albo byłaby zbyt czasochłonna by nazwać ją opłacalną (zmiany manualnie implementowanego parsera *protoc*), albo wymagałyby przepisania parsera w całości, z powodu korzystania przez nie z zewnętrznych zależności których nie da się w prosty sposób dostarczyć.

Tabela 2.1 przedstawia małe podsumowanie zastosowanych technik implementacji parserów w omawianych projektach.

Po przeanalizowaniu powyższych projektów i porzuceniu pomysłu rozwinięci istniejącej już implementacji o potrzebne elementy, rozpocząłem wybór generatora parserów / skanerów który chciałbym zastosować podczas tego projektu.

[1] JFlex (Fast Lexical Analyzer for Java)

Projekt	Metoda impl. skanera	Metoda impl. parsera
Google Protoc	"manualnie", C++	"manualnie", C++
Idea-Plugin-Proto	JFlex, Java	dostarczany z IntelliJ, Java

Tablica 2.1: Zestawienie sposobów implementacji parserów w rozważanych projektach open source

Pierwotnym kandydatem do zastosowania jako generator parsera był powszechnie znany *GNU Bison* [Bis11], który w połączeniu z Flexem pozwolił na wygenerowanie parsera w „znajomy" sposób. Oba te narzędzia są dobrze znane oraz sprawdzone od wielu lat oraz posiadają dobrą dokumentację. W ramach poszukiwań innych rozwiązań natknąłem się jednak na tak zwane „kombinatory parserów", a następnie na fakt iż istnieje ich implementacja wewnątrz biblioteki standardowej języka Scala.

Scala jest statycznie typowanym językiem programowania na platformę Java który wspiera zarówno *obiektowy* jak i *funkcyjny* paradygmat programowania. Tak zwane „*Parser Combinators*" o których tutaj mowa nie są ideą nową. Pojawiły się wraz z językami funkcyjnymi, a pierwsze publikacje naukowe na ich temat można było już napotkać w 1996 roku [GH96] w publikacji Hutton oraz Meijer. Pojęcie „parsowania przy pomocy kombinatorów parserów" najłatwiej jest wytłumaczyć jako:

> „Budowanie parserów rekursywnie zstępujących poprzez modelowanie parserów jako funkcji i definiowanie funkcji wyższego rzędu (zwanych kombinatorami) które implementują konstrukcje takie jak sekwencjonowanie, wybór oraz powtórzenie. [...]" [GH96]

Będziemy mieli zatem w efekcie do czynienia a parserem „rekursywnie zstępującym" (*LL(k)*). Przedstawicielem generatorów tworzących tego typu parsery jest na przykład ANTLR [Ant11], opublikowany po raz pierwszy w roku 1992 jako następca *Purdue Compiler Construction Tool Set* który powstał jeszcze w roku 1989 (sic). Ten typ parserów dodaje do znanej klasy LL(k) funkcjonalność „wycofania się", z dowolnej głębokości look-ahead (parser może pracować z dowolnie dużym *k*, i zawsze będzie w stanie wykonać nawrot oraz wypróbować inną ścieżkę). Korzystanie z nawrotów przez parser oczywiście niesie z sobą zmniejszenie jego wydajności, jednakw wielu przypadkach (jak choćby Protocol Buffers, które mają stosunkowo prostą gramatykę), obawa przed spadniem wydajności nie odzwierciedla się zbytnio w rzeczywistości. Dobrą wiadomością jest natomiast, że w *Scala Parser Combinators* możemy korzystać z wersji metod z dodanym wykrzyknikiem oznaczającym, że pragniemy aby dany fragment był faktycznie klasy *LL(1)* - jeżeli gramatyka nie jest na tyle jednoznacza aby dało się uzyskać *LL(1)* w danym parserze zostaniemy powiadomieni o tym pod postacią błędu.

Jednym z wyróżniających *Scala Parser Combinators* czynników jest fakt iż zamiast pisać pliki w których deklarujemy naszą gramatykę a sam kod źródłowy parsera jest dopiero generowany na jego podstawie w przypadku Scali i wspomnianej biblioteki zdefiniować gramatykę parsera, w połączeniu z blokami kodu które miałyby dokonać odpowiednich transformacji sparsowanych tokenów dokładnie w tym samym pliku który jest „plikiem źródłowym parsera". Dzięki temu oszczędzamy na „kroku" generowania kodu źródłowego parsera, który dopiero później zostałby skompilowany oraz wykonany. Daje

to ogromną przewagę podczas poszukiwania błędów w parserze - ponieważ ewentualne problemy bezpośrednio odwołują się do tego co my napisaliśmy, a nie do odrębnego pliku który powstał na podstawie naszego pliku.

Pomimo tak wielu zalet sama struktura definicji parsera pozostaje podobna do znanej z Bisona oraz nadal jest złudnie podobna do notacji *BNF*[2] - która jest bardzo przejrzysta oraz zazwyczaj znana, lub łatwa to utworzenia podczas pisania parsera znanego języka.

Kolejną zaletą jest umieszczenie definicji tokenów w tym samym pliku co definicja skanera - ponownie unikamy kroku generowania skanera (na przykład przy użyciu *JFlex*a). Zmniejszenie ilości miejsc gdzie konieczne jest wprowadzenie modyfikacji, jest zatem kolejną z zalet tego podejścia.

Metoda impl. skanera	Metoda impl. parsera
Parser Combinators, Scala	

Tablica 2.2: Przedstawienie jednolitości rozwiązania z zastosowaniem *Scala Parser Combinators*

W Tabeli 2.2) zostało przedstawione jak można umieścić *Parser Combinators* w poprzednio przedstawianych kategoriach. Oczywistą zaletą jest tutaj korzystanie w obu cześciach implementacji - zarówno skanerze jak i parserze - z tego samego języka oraz biblioteki, dzięki czemu możemy się skupić na implementacji, a nie nauce kolejnej składni.

Reasumując poniższe zalety są przyczyną wyboru tego podejścia do generowania parsera ponad klasyczne narzędzia typu Flex/Bison:

– Brak konieczności dodatkowego kroku generowania kodu źródłowego:

  dla skanera (brak osobnego pliku ze spisem tokenów),

  dla parsera (brak osobnego języka, dedykowanego definiowaniu.

– Wykonywany kod jest bezpośrednio związany z pisaną przez nas definicją parsera, co pozwala na łatwe poszukiwanie błędów,

– Minimalizacja miejsc w których konieczne jest wprowadzanie zmian, cała implementacja znajduje się w 1 miejscu.

Opis działania *Parser Combinators* znajduje się w kolejnej sekcji, oraz w *Dodatku B*, gdzie wytłumaczone zostały wszystkie zasady konstruowania parserów przy pomocy tej biblioteki.

Jeżeli czytelnik jeszcze nie miał styczności z Scalą oraz dostarczaną przez wraz z nią biblioteką *Parser Combinators* zalecane jest zapoznanie się z *Dodatkiem B*, gdzie szczegółowo omówiono zasady działania samego języka jak i Parser Combinators.

[2]Notacja BNF - „Backus Naur Form”, metoda zapisu reguł gramatyki kontekstowej

3. Projekt systemu

Rozdział ten ma za cel przedstawienie w sposób holistyczny architektury aplikacji *ProtoDoc*. Dopiero w kolejnym rozdziale (Rodział 4) zostaną opisane szczegóły implementacyjne, oraz podjęte decyzje na poziomie kodu, na razie zostanie opisana jestnie generalna architektura oraz podjęte w tej warstwie abstrakcji decyzje.

Główne odpowiedzialności aplikacji zostały podzielone pomiędzy poniższe klasy (konkretniej, klasy typu **object** - omówionych dokładniej w Dodatku B, dotyczącym języka Scala):

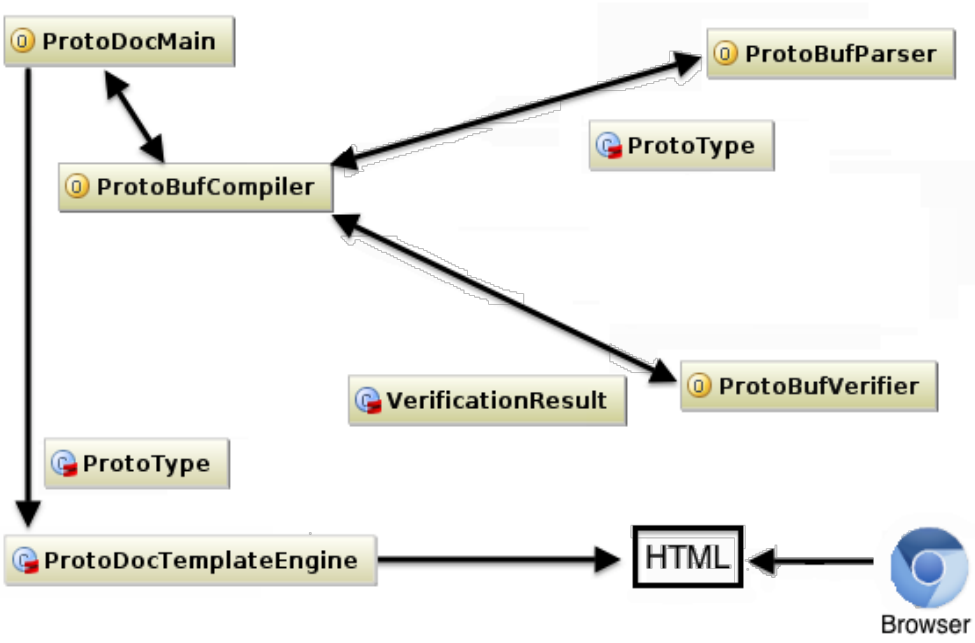
- `ProtoDocCompiler` - jest fasadą nad `ProtoBufParser` oraz `ProtoBufVerifier`. Został wprowadzony celem ułatwienia pracy na „zweryfikowanych" już obiektach typu ProtoType, będącymi wynikami parsowania, jednak dopiero po wykonaniu weryfikacji, możemy być pewni że utworzone obiekty z pewnością są poprawne. Compiler enkapsuluje parsowanie oraz weryfikację w jedno publiczne API, znacznie upraszczając testowanie oraz korzystanie z ProtoDoc na poziomie kodu (a nie zewnętrzej aplikacji).

- `ProtoBufParser` - jest implementacją parsera przy pomocy *Scala Parser Combinators*. Odpowiedzialny jest również za analizę syntaktyczną - błędy odnaleziome podczas parsowania (na przykład „za duży" **tag** dla pola wiadomości) zostałby wykryty już podczas parsowania. Miłym akcentem jest tutaj iż parser jest w stanie podać kontekst w którym wystąpił błąd, zaznaczając go znakiem `^`, co znacznie uprzyjemnia korzystanie z niego użytkownikom końvowym. Omówiony zostanie szczegółowo w sekcji 4.1.

- `ProtoBufVerifier` - „weryfikator" zajmujący się sprawdzaniem poprawności semantycznej sparsowanych plików *.proto. W przypadku napotkania błędów krytycznych, działanie protodoc może zostać w tym miejscu przerwane, oraz zwrócony zostałby komunikat informujący o przyczynie błędu.

- `ProtoDocTemplateEngine` - „generator kodu", w naszym przypadku generowanym „kodem" jest HTML. TemplateEngine wykorzystuje wewnątrz silnik renderowania szablonów *Mustache*, który zostanie omówiony w sekcji 4.3.1. Mechanizm działania generatora jest stosunkowo prosty oraz sprowadza się do podstawiania odpowiednich wartości w odpowiednie „zmienne" celem udostępnienia ich do wypisania przez system szablonów *Mustache*.

W kolejnych sekcjach tego rozdziału zostaną przedstawione w formalny sposób interakcje oraz relacje między komponentami systemu. Poruszone zostaną również typy którymi zamodelowano w języku Scala odpowiednie typy pól mogące pojawić się w wiadomości ProtoBuf. Wpierw zostanie jednak pokazany w sposób bardziej wizualny niż formalny, jak komponenty się ze sobą komunikują - służy temu Rysunek 3.1.



Rysunek 3.1: Nieformalna wizualizacja interakcji pomiędzy komponentami

Dzięki nie formalnej wizualizacji interakcji między komponentami na Rysunku 3.1 można łatwo zobrazować sobie jak komponenty między sobą wymieniają informacje. Pliki *.proto są parsowane przez parser, po czym reszta komunikacji odbywa się na poziomie klas typu `ProtoType`, oraz jej specjalizacji.

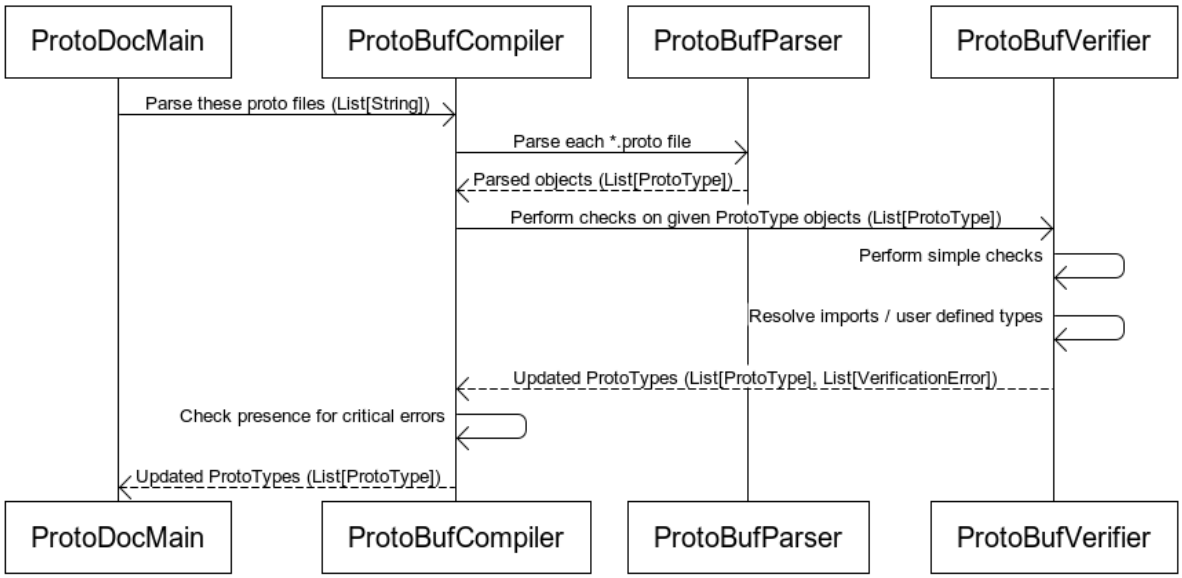
Jak widać mamy wydzielone klasyczne dla kompilatorów odpowiedzialności: parsowanie/skanowanie, weryfikację semantyczną oraz generowanie kodu. W kolejnych sekcjach zostaną przedstawione diagramy oraz wizualizacje interakcji między tymi komponentami.

3.1. Architektura aplikacji

W tej sekcji przedstawione zostaną diagramy sekwencji opisujące interakcje między wcześniej już wstępnie opisanymi komponentami.

Na diagramie sekwencji 3.2 zostało przedstawione szczegółowo do jakich interakcji między wcześniej opisywanymi (nie formalnie) komponentami dochodzi podczas procesu parsowania.

Punktem wejściowym aplikacji jest klasa `ProtoDocMain`, po sparsowaniu argumentów wejściowych specjalnym Domain Specific Language (innym niż Parser Combinators, zostanie on omówiony i przedstawiony w sekcji poświęconej tej klasie), przekazuje przygotowane do parsowania zawartości



Rysunek 3.2: Diagram sekwencji parsowania oraz weryfikowania wiadomości

plików do obiektu `ProtoBufCompiler`. Compiler jedynie deleguje parsowanie każdego z plików do `ProtoBufParser`, który wykonuje parsowanie przy pomocy *Scala Parser Combinators*.

Ponieważ jeden plik *.proto może zawierać więcej niż jedną wiadomość (lub enumerację), dla każdego sparsowanego pliku proto zwracana jest lista typów konkretnych *(omówionych szczegółowo w sekcji 3.1.1)*, dziedziczących po abstrakcyjnej klasie `ProtoType`.

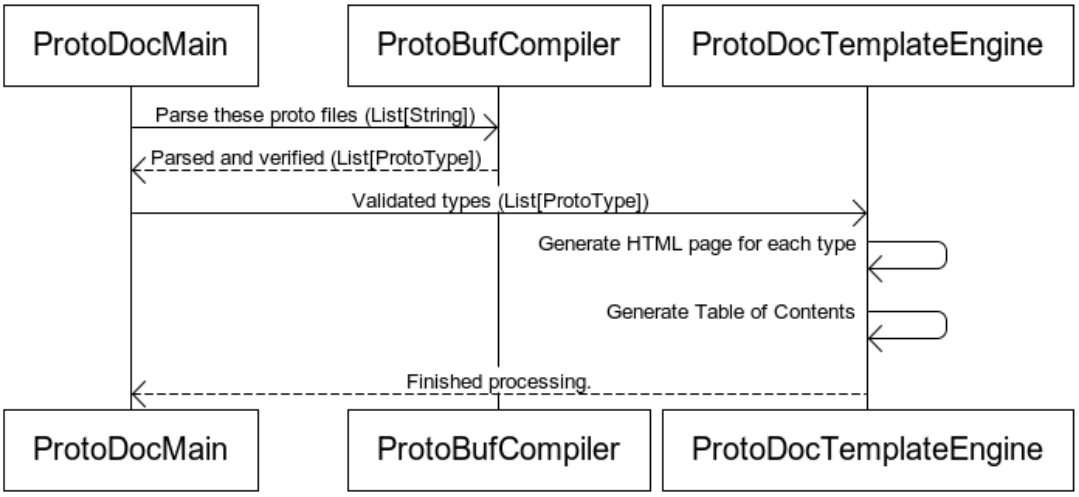
Kolejnym krokiem jest przekazanie otrzymanych właśnie `ProtoType` do Instancji `ProtoBufVerifier`, który zajmuje się weryfikacją poprawności typów. Może on, ponieważ otrzymuje *wszystkie* sparsowane typy, również rozstrzygnąć czy *import* z innego pakietu jest poprawny czy też nie. Innymi słowy, jedną z jego odpowiedzialności jest rozwiązanie problemów widoczności typów. Nie mógł, oraz nie powinien, przeprowadzać tego Parser, ponieważ nie posiadał jeszcze pełnego zestawu danych (typów) koniecznych do rezolucji typów. Sposób w jaki informacja o „nie rozstrzygniętym typie" jest przekazywana do weryfikatora, jest dość prosty: Parser podczas napotkania typu którego jeszcze nie zna, tworzy taki sam obiekt referencji do pola, jaki stworzyłby gdyby znał ten typ, jednak dodatkowo zaznacza iż typ nie był widoczny podczas parsowania. Gdy weryfikator dostaje wszystkie typy oraz ich pola, przeszukuje pola, zważywszy na wspomnianą flagę - oraz próbuje dokonać rezolucji typu, mając już pełne informacje o dostępnych typach w danym kontekście. Przyjęta przez niego strategia zostanie przybliżona dokładniej w sekcji 4.2.

`ProtoBufVerifier` odpowiada listą komunikatów mogących być albo błędami, albo ostrzeżeniami, a `ProtoBufCompiler` może na nie zareagować przerwaniem wykonania aplikacji oraz wypisaniem wszystkich problemów, lub może zwrócić wszystkie zweryfikowane typy (obiekty typu `ProtoType`) do głównej klasy programu, która pierwotnie rozpoczęła proces parsowania plików proto — do `ProtoDocMain`.

Kolejnym krokiem w kierunku wytworzenia wyniku działania aplikacji — gotowej strony www dokumentacji ProtoDoc — jest przekazanie zweryfikowanych obiektów `ProtoType` z `ProtoDocMain` do ostatniego z istotnych komponentów aplikacji — do generatora kodu, tj. do instancji klasy `ProtoTemplateEngine`. Warty zauważenia jest powrót to nazewnictwa *Proto**Doc**...* tej klasy, w przeciwieństwie do *Proto**Buf**...* występującego jako prefix poprzednich komponentów. Przyczyną takiego rozróżnienia w nazewnictwie jest, iż zarówno parser jak i weryfikator, nie są ściśle związane z ProtoDoc - mogą parsować oraz weryfikować dowolne pliki proto, oraz zwracają ich obiektową reprezentację. Równie dobrze komponenty te można by wykorzystać w innych celach - nie tylko celem generowania dokumentacji.

Proces generowania kodu jest bardzo prosty. Jako, że `ProtoBufCompiler` dostarczył już „gotowe" obiekty, które na pewno są poprawne, jedyne co pozostaje generatorowi do zrobienia to wygenerowanie strony HTML na podstawie każdego z obiektów proto-typów, które zostały mu przekazane przez `ProtoDocMain`.

Diagramie sekwencji 3.3 obrazuje proces generowania kodu. Pierwsze dwie wymiany wiadomości zostały szczegółowo przedstawione na poprzednim diagramie (3.2).



Rysunek 3.3: Diagram sekwencji generowania stron HTML dokumentacji

3.1.1. Zastosowany model typów wiadomości Protocol Buffers w Scala

Celem wykorzystania w pełni z potencjału statycznego typowania języka Scala — zamiast pracowania wprost na listach i mapach wartości które w domyślnym przypadku zostałyby zwrócone przez parser, konieczne było stworzenie modelu typów, odzwierciedlającego w Scali typy mogące pojawić się w ProtoBuf. Konieczne było zamodelowanie zarówno istniejącej wewnętrznie w Protocol Buffers jak i łatwej do rozszerzania o typy definiowanie przez użytkownika struktury typów.

Zdefiniowanie nowego typu przez użytkownika jest główną czynnością podczas opisywania interfejsu przy pomocy Google Protocol Buffers, także typy te powinny również mieć istotne odzwiercie-

dlenie w ProtoDoc. W ramach przypomnienia, Listing 3.1 przedstawia składnię zdefiniowania prostej wiadomości (`message`) lub enumeracje (`enum`) - jedynych interesujących nas w zakresie ProtoDoc typów.

Listing 3.1: Przykład zdefiniowania nowych typów w Protocol Buffers IDL

```
package pl.project13;

/** some comments */
message MyNewType { required string pole = 1; }

/** some comments */
enum MyEnumeration { VALUE = 1; }
```

Okazuje się, że enumeracja i wiadomość dzielą wiele wspólnych cech, oba typy:

– mogą znajdywać się w *package*,

– mają nazwę,

– mają „w pełni kwalifikowaną nazwę", składającą się z połączenia nazwy typu,

– zawierają pola. Mimo, że w przypadku enumeracji pole wygląda nieco inaczej, można przyjąć pewne uogólnienie i traktować je tak samo jak pozostałe pola.

W związku zauważeniem powyższych wspólnych cech został wprowadzony wprowadzony typ `ProtoType`, będący super-typem dla obu tych klass.

Dodatkowo warto zauważyć, że komentarze mogą być umieszczone *„na"* każdym z wspomnianych typów, oraz to samo można powiedzieć o polach umieszczonych wewnątrz tych typów - jak obrazuje to Listing 3.2.

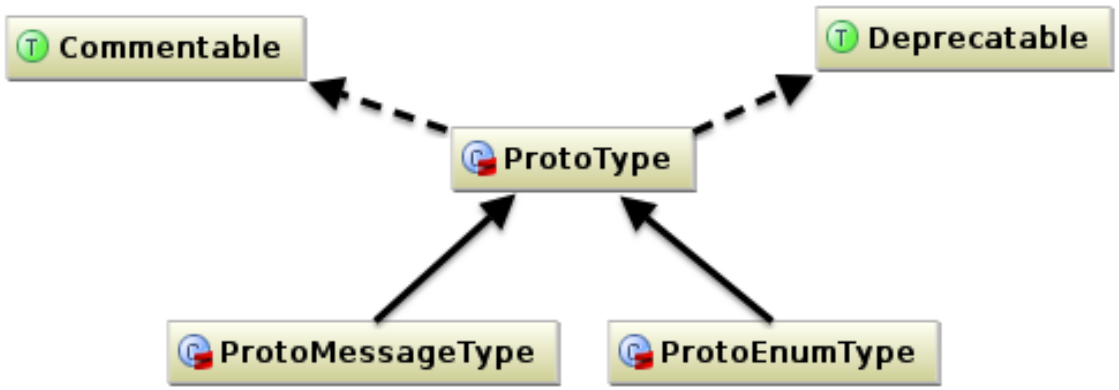
Listing 3.2: Przykład umieszczenia komentarza ProtoDoc na polach wiadomości oraz enumeracji

```
message MyNewType { /** docs */ required string field = 1; }

enum MyNewType { /** docs */ SOME_VALUE = 1; }
```

Umieszczenie pola *comment* wewnątrz klasy `ProtoType` nie byłoby zatem dobrym pomysłem, ze względu na nie-możliwość ponownego użycia interfejsu typu „*X* ma komentarz" na polach. Rozwiązaniem jest wprowadzenie interfejsu „`Commentable`", które eksponuje metody związane z posiadaniem komentarza, na przykład „czy komentarz jest obecny?" etc. Pozwoliło to na implementację metod pracujących na tym interfejsie, nie ważne czy ów komentowalny element jest polem czy nowym typem, zatem zmniejszyło ścisłość wiązań w API tworzonego systemu.

W przypadku języka Scala, mowa tutaj nie o interfejsach a *Trait*ach (rozpoznawalnych na załączonych diagramach po zielonej ikonie **T**), jednak jako że celem tego rozdziału pracy nie przedstawienie implementacji a generalnego design projektu można przyjąć że Trait może być rozumiany jako „interfejs". Szczegółowy opis czym są *Trait*y oraz jak wpływają na design projektu, można przeczytać w Sekcji B.3, w Dodatku B.



Rysunek 3.4: Zastosowany model typów definiowanlnych przez użytkownika typów

Rysunek 3.4 przedstawia zastosowaną w projekcie strukturę typów. Warto zauważyć dodatkowy Trait o nazwie `Deprecatable`, oznaczający „coś co może zostać adnotowane jako przestarzałe". Adnotowanie pól jako `@deprecated` jest dobrą praktyką programistyczną oraz bezcennym źródłem dodatkowej informacji o typie dla takiego narzędzia jak ProtoDoc, które może takie pola wyszarzyć lub zasugerować jakiego typu powinno się użyć zamiast oznaczonego taką adnotacją. Przyczyną wprowadzenia tego typu jako Trait ponownie jest fakt, że wiele elementów protocol buffers może zostać oznaczona jako przestarzała, nie jedynie same deklaracje typów.

3.1.2. Zastosowany model typów pól Protocol Buffers w Scala

Analogiczny jak przedstawiony powyżej problem dotyczy pól, które mogą być reprezentować jeden z predefiniowanych typów protocol buffers, lub mogą być enumeracją lub wiadomością definiowaną przez użytkownika.

Celem skorzystania również i tutaj z statycznego typowania również podczas pracy na polach, również tutaj konieczne było wprowadzenie modelu typów. Tutaj wymaganiem ponownie było aby możliwie prosto dało się wspierać już wbudowane w ProtoBuf jak i nowo definiowane przez użytkownika typy.

Listing 3.3 przedstawia różne możliwości jakie mamy do dyspozycji podczas definiowania pola w wiadomości lub enumeracji. W przypadku gdy czytelnik chciałby się zagłębić w szczegóły składni oraz znaczenia poszczególnych deklaracji, zachęcam do przeczytania Dodatku A, w którym wszystkie te

elementy są szczegółowo omawiane. Przykład tutaj przytaczany jest celem unaocznienia ponownie cech wspólnych dla wspomnianych elementów.

Listing 3.3: Deklaracja pola w wiadomości oraz enumeracji

```
// message fields:
optional int32 age = 1;
required string age = 2 [default = "Konrad"];
repeated sint64 numbers = 3;

optional MyType it = 4;
optional MyEnum other = 5 [default = VALUE];

message MyType { }
enum MyEnum {
  // enum values:
  VALUE = 3;
}
```

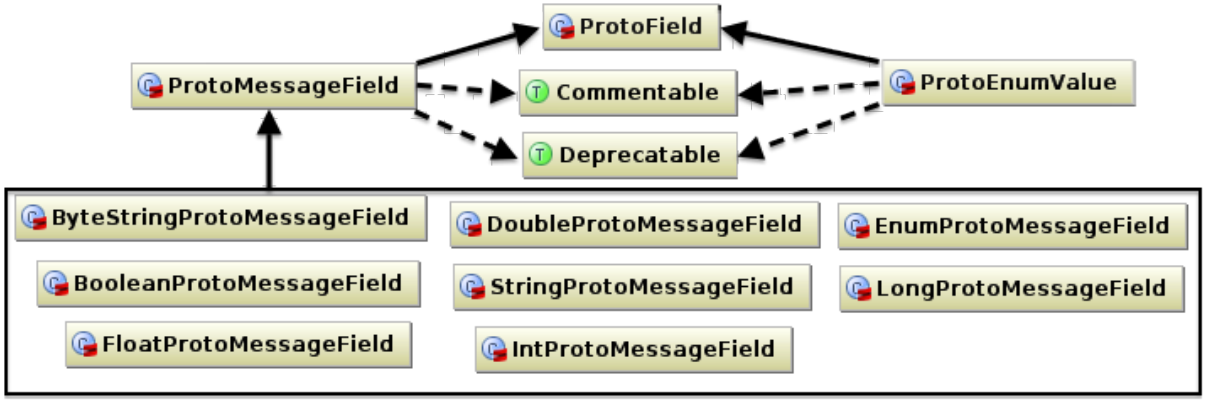
Deklarowanie pól w wiadomościach jest oczywiście bardziej interesujące niż pola w enumeracjach, co powyższy przykład powinien dość ładnie obrazować. Mimo wszystko, ponownie jesteśmy w stanie znaleźć pewne wspólne elementy pomiędzy wszelkimi „polami", włączając w to również wartości enumeracji. Przy okazji, pamiętajmy iż każde z tych pól może być *Deprecated* oraz może zostać okomentowane komentarzem ProtoDoc.

- pole posiada nazwę,
- pole posiada przypisany mu *tag* (liczba umieszczona po znaku równości, po szczegółowe wyjaśnienie czym tag jest, odsyłam do Dodatku A),
- pole może posiadać komentarz,
- pole może być ozaczone jako przestarzałe (*Deprecated*).

Powyższa lista elementów wspólnych w sposób oczywisty doprowadziła do powstania klasy `ProtoField` będącej podstawą ku wszystkim pozostałym typom. Dzięki wydzieleniu Traitów `Commentable` oraz `Deprecatable`, również teraz można je zastosować aby osiądnąć te same funkcjonalności w nowo przedstawionych klasach.

Pole będące wartością enumeracji będziemy traktować analogicznie jak zwyczajny `ProtoField`, nie jest ono wyjątkowe z perspektywy struktury samej z siebie. Co je odróżnia od `ProtoMessageField` jest natomiast fakt iż po `ProtoMessageField` dziedziczą wszystkie predefiniowane typy. Dokładne oddanie relacji pomiędzy utworzonymi tutaj typami a typami umieszczanymi w plikach proto nie jest w naszym przypadku konieczne, a być może nawet nie wskazane. W przypadku korzystania z protocol buffers na platformie JVM, nie istnieją odpowiedniki typów unsigned / signed - także korzystamy w tym przypadku po prostu z typu `scala.Int`, do przechowania np.

wartości domyślnej takiego pola. Nazwa typu którego dana klasa jest mapowaniem zostaje jednak utrzymana w ciele klasy, abyśmy mogli podczas generowania dokumentacji przywołać jaki dane pole miało typ w pliku *.proto. Pełen spis przyjętych mapowań został przedstawiony w formie Tabeli 3.1.



Rysunek 3.5: Zastosowany model typów pól mogących wystąpić w ProtoBuf IDL

W ramach ciekawostki, chciałbym w tym miejscu przedstawić implementację dwóch przykładowych z omawianych klas.

Listing 3.4: Pełna implementacja klasy ProtoEnumValue

```
case class ProtoEnumValue(valueName: String, tag: ProtoTag)
  extends ProtoField
  with Commentable
  with Deprecatable
```

Listing 3.5: Pełna implementacja klasy StringProtoMessageField

```
case class StringProtoMessageField(override val fieldName: String,
                           override val tag: ProtoTag,
                           override val modifier: ProtoModifier,
                           override val defaultValue: String = None)
  extends ProtoMessageField(fieldName,
                     "string", // type name in *.proto
                     "java.lang.String", // JVM representation
                     tag,
                     modifier,
                     defaultValue)
```

Warto zwrócić uwagę, że mimo iż klasy przedstawiona na Listingach 3.4 oraz 3.5 nie posiadają stricte zdefiniowanego ciała, dzięki pewnym konwencjom oraz mechanizmom udostępniamym przez Scala, takie definicja są wszystkim co jest potrzebne do zdefiniowania w pełni funkcjonalnych klas, włącznie z polami, akcesorami dla atrybutów etc.

Typ pola w Protocol Buffers	Odpowiednik w ProtoDoc
double	DoubleProtoMessageField
float	FloatProtoMessageField
int32	IntProtoMessageField
int64	LongProtoMessageField
uint32	IntProtoMessageField
uint64	LongProtoMessageField
sint32	IntProtoMessageField
sint64	LongProtoMessageField
fixed32	IntProtoMessageField
fixed64	LongProtoMessageField
sfixed32	IntProtoMessageField
sfixed64	LongProtoMessageField
bool	BooleanProtoMessageField
string	StringProtoMessageField
bytes	ByteStringProtoMessageField
enumeration type	EnumProtoMessageField(type)
user defined message type	ProtoMessageField(type)
enum value	ProtoEnumValue

Tablica 3.1: Typy pól oraz odpowiadające im typy Scala (bazując na reprezentacji typów przez protoc na JVM)

Czytelnika zainteresowanego mechanizmami kryjącymi się za zastosowanymi tutaj „`case class`ami” zachęcam do przeczytania sekcji B.4, dotyczącej tego działu języka Scala, z Dodatku B.

4. Szczegóły implementacyjne

Poniższy rozdział przedstawi szczegóły implementacyjne poszczególnych komponentów składających się na *ProtoDoc*. Ogólna architektura oraz interakcie między komponentami zostały już opisane w poprzednim rozdziale, stąd te kwestie nie bądą tutaj ponownie poruszane, uwaga natomiast zostanie skoncentrowana na szczegółach implementacyjnych, oraz ewentualnych wyjaśnieniach nowo wprowadzanych pojęć lub bibliotek.

Omówione zostaną zarówno ogólne założenia przyjęte podczas projektowania systemu, jak i struktura klas przyjęta celem modelowania struktury typów Protocol Buffers. Następnie przedstawione zostaną poszczególne komponenty aplikacji, z naciskiem na uzasadnienie wybranych rozwiązań oraz rozważenia ich zalet, wad oraz potencjalnych możliwości usunięcia zauważonych wad.

Na zakończenie rozdziału przedstawione zostaną elementy kodów źródłowych, skrócone do postaci wystarczającej na cel omówienia danego tematu - w przypadku chęci zapoznania się ze całością implementacji np. komponentu parsera, zachęcam do zapoznania się z załączonymi do pracy plikami źródłowymi projektu.

Jako że poniższy rozdział wymaga od czytelnika minimalnej choćby znajomości Scala oraz Protocol Buffers IDL, zalecane jest zapoznanie się z dodatkami A (Protocol Buffers) oraz B (Podstawy języka Scala).

4.1. ProtoBufParser

Parser został zaimplementowany przy pomoy wspominanych wielokrotnie już kombinatorów parserów, dostarczanych wraz z biblioteką standardową Scali. Jako wiadomość referencyjną, służącą jako przykład parsowanego pliku *.proto, będziemy posługiwać się w tym rozdziale bardzo prostą wiadomością, obejmującą jednak podstawowe funckojalności definiowania typów w Protocol Buffers, przedstawioną na Listingu 4.1.

Listing 4.1: Przykład wiadomości, służący łatwiejszej wizualizacji działania parsera

```
message MyMessage {
  required FullName name = 1;
  optional int32 age = 2 [default = 1];
  optional Gener gender = 3;

  message FullName {
    required string firstname = 1;
    required string lastname = 2;
  }

  enum Gender {
    MALE = 1;
    FEMALE = 2;
  }
}
```

Parser jest zdefiniowany jako obiekt dziedziczący po klasie `RegexParsers` (patrz Listing 4.2), będącą domyślną implementacją dostarczającą metody `parse` oraz `parseAll`, konieczne do implementacji parsera. Ponad wspomniane funkcje, udostępnia również możliwość ignorowania białych znaków oraz kilka metod pomocniczych. Wszystkie najważniejsze metody zawarte są w Traicie `Parsers`, którego RegexParsers dostaje wmieszanego (jak wytłumaczono w Dodatku B).

Listing 4.2: Definicja obiektu parsera

```
object ProtoBufParser
  extends RegexParsers // includes the parser DSL
  with ImplicitConversions // conversions for list flattening
  with ParserConversions // my implicic conversions
  with Logger { // include a logger
```

Dodatkowo wmieszane zostały dwa Traity udostępniające implicit konwersje (patrz Dodatek B, Sekcja B.5 – Implicit Conversions), ułatwiające pracę z listami oraz konwertowanie między typami takimi jak proste typy liczbowe do typu reprezentującego tag protobufowy (klasa `ProtoTag`).

Idąc dalej przyjrzymy się najprostrzemu parserowi w tej klasie. Będzie to proste wyrażenie regularne matchujące identyfikatory które można stosować w plikach *.proto. Intentyfikatory te mają podobne warunki istnienia jak indentyfikatory w Java, także wyrażenie (de facto, cały „parser indentyfikatora") tak jak to przedstawiono na Listingu 4.3

Listing 4.3: Bardzo prosty parser, zdefiniowany za pomocą implicita, który rozszerza API klasy String o metodę *r* tworzącą wyrażenie regularne (instancję klasy *scala.util.matching.Regex*)

```
val ID = """[a-zA-Z_]([a-zA-Z0-9_]*|_[a-zA-Z0-9]*)*""".r
```

Następnie zdefiniujemy mały parser wykorzystując powyższy (Listing 4.4):

Listing 4.4: Definicja parsera identyfikatora wiadomości

```
def messageTypeName /*: Parser[String]*/ = "message" ~> ID
```

Znak ~> oznacza *kombinację parserów, przy odrzuceniu lewej strony*. W efekcie zdefiniowaliśmy właśnie część gramatyki oznaczającą iż jeżeli pojawi się słowo message, powinien po nim nastąpić pewien identyfikator, ale dla celów obróbki tej informacji, będzie nas interesować tylko parser po prawej stronie znaku ~>. Lewa strona zostanie jedynie użyta podczas parsowania, oraz słowo „message" nie będzie widoczne podczas dalszego kombinowania tego parsera z pozostałymi.

Warto zauważyć iż znak ~ jak i jego odpowiedniki „porzucające" lewą (~>) lub prawą (<~) stronę sekwencji parserów, po pierwsze: nie są operatorami, a metodami, oraz po drugie: są dostarczane przez implicit konwersje na parserach lub typach prostych. (Szczegółowy opis znajduje się w Dodatku B).

Kolejnym potrzebnym do zdefiniowania parserem jest właściwe ciało wiadomości. Definicja jest bardzo prosta i czytelna, ponieważ korzystamy tutaj z zdefiniowanych gdzie indziej parserów poszczególnych pól. Listing 4.5 przedstawia parser ciała wiadomości, który niebawem zostanie użyty wspólnie z parserem nazwy wiadomości celem budowy pełnego obiektu wiadomości przy pomocy kombinacji tych parserów.

W ramach przypomnienia zanim spojrzymy na kod parsera ciała wiadomości przypomnijmy sobie jakie elementy może ono zawierać:

– definicję pola wiadomości,

– definicję zagnieżdżonej wiadomości,

– definicję zagnieżdżonej enumeracji.

Powyższe trzy typy mogących wystąpić w ciele wiadomości deklaracji bardzo łatwo jest przetłumaczyć na parser, przedstawiony na Listingu 4.5.

Listing 4.5: Parser ciała wiadomości

```
def messageBody = "{" ~>
                  rep(instanceField | enumTypeDef | messageTypeDef)
                <~ "}"
```

Listing 4.5 poza znanym nam już kombinatorem ~> wykorzystuje jeszcze kombinator `rep()`, oznaczający po prostu iż dany element może powtarzać się wielokrotnie. Jego działanie jest analogiczne do znaku + w wyrażeniach regularnych.

Poza parserem ciała wiadomości umieszczonym na Listingu 4.5, konieczne oczywiście jest zdefiniowanie parserów `enumTypeDef`, `messageTypeDef` oraz `instanceField`. Implementacje tych parserów, są stosunkowo długie także nie zostaną umieszczone w całości w tym dokumencie. Jako przykład zostanie podane jednak pierwsze kilka linii metody definicji, pozwalającej nam zapznać się z generalną zasadą działania wszystkich zaimplementowanych w projekcie parserów, ich kombinacji oraz

jak dochodzi to przekształcenia zwyczajnego drzewa syntaktycznego do zaprojektowanych przez nas w porzednim rozdziale klas.

Listing 4.6: Skrócona implementacja parsera pola wiadomości

```
def instanceField = opt(comment) ~ modifier ~! (protoType |
    userDefinedType) ~ ID ~! "=" ~! integerValue ~ opt(defaultValue) <~ ";"
    ^^ {
   case doc ~ mod ~ pType ~ id ~ eq ~ tag ~ defaultVal =>
     val comment = doc.getOrElse("") // : Option[String]
     // ...

     // return the prepared instance
     if(isResolvedType) {
       if(isKnownProtoEnumField(pType))
         // ...
         ProtoMessageField.toEnumField(id, itsEnumType, tag, mod, defaultVal)
       else
         // ...
         ProtoMessageField.toProtoField(pType, id, tag, mod, defaultVal)
     } else {
       // ...
       ProtoMessageField.toUnresolvedField(pType, id, tag, mod, defaultVal)
     }
}
```

Z nowych elementów pojawiły się tutaj metody | (działającej analogicznie do logicznej alternatywy) oraz ^^ będącej operacją transformacji przeparsowanego ciągu. Operacja ^^ jest jedną z najistotniejszych ponieważ pozwala zmianę produkcji parsera z zwyczajnych list, do faktycznych obiektów, które zamodelowaliśmy w poprzednim rozdziale. Szczegóły składni jej wykonania są dość złożone, jednak w skrócie, mamy tutaj do czynienia z funkcją wyższego rzędu, której przekazujemy funkcję która będzie w stanie transformować sparsowany ciąg, do typu ProtoType. W naszym przykładzie zwracany jest `ProtoMessageField` z pobranych podczas parsowania elementów. Słowo kluczore `return` znane z innych języków programowania nie jest tutaj konieczne.

Pozostałe parsery są bardzo podobne w implementacji do przedstawionych powyżej, oraz generalnie sprowadzają się głównie do sprawdzania poprawności, na przykład wartości domyślnej w przypadku znanego przez parser już typu oraz przypisań odpowiednich elementów do odpowiednich typów, które szczegółowo omówiono w poprzednim rozdziale.

4.2. ProtoBufVerifier

Klasa `ProtoBufVerifier` pojawiła się z konieczności przeprowadzania rozwiązania typów (ang. *rype resolution*), w przypadku gdy mamy do czynienia z odwołaniem się do typu zdefiniowanego albo „poniżej" typu zawierającego to odwołanie, lub na przykład w innym pliku. Parser niestety nie jest wówczas w stanie rozstrzygnąć samodzielnie czy dany typ, na przykład zastosowany na polu wewnątrz wiadomości, jest nie zdefiniowany czy po prostu został zdefiniowany w innym miejscu, do którego parser jeszcze nie doszedł. Verifier unika tych problemów, pracę po zakończeniu działania parsera, kiedy to wszystkie informacje o typach są już dostępne.

Przechodząc do klasy `ProtoBufVerifier`, mamy już zakończony parsing oraz przekazane wszystkie sparsowane typy do instancji tej klasy. `ProtoBufVerifier`, podobnie jak Parser, również jest zdefiniowany jako **object**, co odrobinę upraszcza korzystanie z niego. Patrząc na Verifier z zewnątrz (z poziomu `ProtoDocCompilera`) użycie go sprowadza się do wykonania weryfikacji na komplecie przygotowanych przez parser obiektów ProtoType, oraz następnie sprawdzenie czy zawierają błędy „krytyczne". Fragment kodu odpowiedzialny za te czynności, umieszczony w `ProtoDocCompiler` został przedstawiony na listingu 4.7

Listing 4.7: Przykład wykorzystania weryfikatora

```
val verification= ProtoBufVerifier.verify(parsedProtos)

if(verification.invalid) {
  verification.errors.foreach(error(_)) // print all errors
  throw new ProtoDocVerificationException(verification)
}
```

Implementacja weryfikatora bazuje głównie na zebranych informacjach oraz pojęciu „kontekstu" (na przykład „wewnątrz tej wiadomości") dzięki któremu jest w stanie rozstrzygnąć czy widoczność typów jest poprawna etc. Zaimplementowane weryfikacje różnią się dla typów wiadomości oraz enumeracji. Poszczególne weryfikacje zostaną opisane w kolejnej sekcji. Listing 4.8 przedstawia na jakiej zasadzie Weryfikator deleguje wykonanie sprawdzeń poprawności do konkretnych wyspecjalizowanych metod.

Listing 4.8: Delegacja sprawdzania poprawności typów

```
def verify(protoTypes: List[ProtoType]): VerificationResult =
  VerificationResult((for (protoType <- protoTypes)
                  yield check(protoType, protoTypes)).flatten)
```

```
def check[T <: ProtoType](protoType: T,
                    protoTypes: List[T]): List[VerificationError] =
  protoType match {
    case msgType: ProtoMessageType =>
      checkMessageType(msgType, protoTypes)
```

```
    case enumType: ProtoEnumType =>
      checkEnumType(enumType, protoTypes)
}
```

Tutaj zastosowanie znajduje **pattern matching**, znany z języków funkcyjnych, takich jak *Haskell* lub *Erlang*. Konstrukcja **match** pozwala na dekonstrukcję matchowanego typu oraz wiele bardzo potężnych operacji które w przeciwnym przypadku byłyby bardzo długą serią instrukcji warunkowych oraz rzutowań typów. Metoda `check` jak widać, deleguje wiadomość jeszcze głębiej, jednak tym razem już do specjalizowanych metod odpowiadających za sprawdzanie poprawności typu. Przykładem implementacji checków jest `checkMessageType` umieszczony na listingu 4.9.

Listing 4.9: Metoda checkMessageType, sprawdzająca poprawność wiadomości

```
def checkMessageType(msgType: ProtoMessageType, protoTypes:
    List[ProtoType]) = {
  info("Running verifications on "+b(msgType)+" message")

  // errors
  val tagErrors = TagVerifier.validateTags(msgType,
      msgType.fields.map(_.tag))

  val enumErrors = for (enum <- msgType.enums) yield checkEnumType(enum,
      protoTypes)
  val fieldErrors = for (field <- msgType.fields) yield checkField(msgType,
      field, protoTypes)
  val innerMsgErrors = for (innerMsg <- msgType.innerMessages) yield
      checkInnerMsg(msgType, innerMsg, protoTypes)

  // warnings
  val deprecatedItemsCount = checkDeepDeprecation(msgType)
  warn("Found "+deprecatedItemsCount+" deprecated fields/types in
      ["+msgType.fullName+"]")

  tagErrors ::: fieldErrors.flatten ::: enumErrors.flatten :::
      innerMsgErrors.flatten ::: Nil // return a list of all errors
}
```

4.2.1. Obsługiwane weryfikacje

ProtoDoc w momencie pisania tej pracy obsługuje następujące weryfikacje poprawności (dla wszystkich typów pól oraz deklaracji typów jakie parser obecnie jest w stanie zrozumieć):

– poprawność **tag**ów:

  czy tag zawiera się w specyfikowanym przez Protocol Buffers zakresie dozwolonych liczb,

czy tag jest niepowtarzalny w zasięgu bloku danej wiadomości lub enumeracji.

– czy typ definiowany przez użytkownika posiada niepowtarzalną w pełni kwalifikowaną nazwę,

– czy wartość domyślna pola w wiadomości posiada typ zgodny z typem tego pola (np. liczbę całkowitą dla pól int32),

– czy w przypadku korzystania z definiowanego przez użytkownika typu, wykorzystanego podczas deklaracji pola, typ ten jest „widoczny" z wewnątrz kontekstu tego pola,

– czy wiadomość lub enumeracja nie posiada zduplikowanych nazw pól.

Najciekawszą z wspomnianych weryfikacji jest definitywnie sprawdzanie widoczności typu, z racji wielu możliwości odwołania się do niego, na przykład podczas deklaracji pola w wiadomości. Całość implementacji jest b. długa, jednak celem zobrazowania jak można budować czytelne weryfikacje przy pomocy pisania własnych implicit conversions, zostaną przytoczone w Listingu 4.10 najciekawsze fragmenty metody odpowiedzialnej za sprawdzenie czy typ danego pola jest „w zasięgu".

Listing 4.10: Ważniejsze fragmenty implementacji sprawdzania widoczności typu

```
// by using the find method on List
val fullyQualifiedMatch = allParsed.find(_.fullName == typeName)
if(fullyQualifiedMatch.isDefined) {
  field resolveTypeTo(fullyQualifiedMatch.get)
  return NoErrorsEncountered
}

//by using an infix notation, via implicit conversions
if(typeName isDefinedWithin fromContext) {
  val resolvedType = typeName getResolvedTypeWithin fromContext
  field resolveTypeTo(resolvedType)
  return NoErrorsEncountered
}
// ...
// unable to resolve, report error
error("the field: ["+field.fieldName+"] was unresolvable at this
    point...")
return UndefinedTypeVerificationError(field.fieldName, "Unable to resolve
    type ["+typeName+"] from ["+fromContext+"] context.") :: Nil
```

Dodatkowo zostały zaimplementowane ostrzeżenia, które nie są błędne w kontekście Protocol Buffers, jednak są „podejrzane" stąd warto wskazać je programiście podczas poszukiwania problemów w kodzie:

– czy enumeracja nie jest pusta (nie posiada żadnych wartości),

– czy wiadomość nie jest pusta (nie posiada żadnych pól).

Weryfikacje te były bardzo proste w implementacji dzięki zaprojektowanemu modelowi danych. Jako przykład może posłużyć sprawdzenie czy typ jest pusty, przedstawione w całości w Listingu 4.11.

Listing 4.11: Implementacja ostrzeżenia czy enumeracja jest pusta

```
def checkIfEmpty(enumType: ProtoEnumType) {
  if(enumType.values.isEmpty)
    warn("""|The enum "+enumType.fullName+" has no values.
            |This could be a possible typo with missplacing the "}"..."""
            .stripMargin)
}
```

Weryfikacje są bardzo ciekawym elementem ProtoDoc, oraz jednym z najbardziej obiecujących miejsc do dalszego rozwoju. Nie trudno jest sobie wyobrazić wiele heurystycznych metod mogących tutaj znaleźć zastosowanie, oraz pomóc programiście pracować nad poprawianiem jakości tworzonych przez siebie wiadomości. Tymczasem, podstawowe checki, które zostały zaimplenentowane w tym projekcie powinny być wystarczające - zważywszy na fakt iż są dodatkiem, a nie właściwym celem aplikacji.

4.3. ProtoDocTemplateEngine - generator kodu

Ostatim komponentem składającym się na ProtoDoc jest generator kodu, w naszym przypadku jego rolę pełni klasa `ProtoDocTemplateEngine`.

Jego rola jest bardzo prosta, jecyne co musi zrobić to dla każdego otrzymanego typu, wygenerować stronę HTML, z przygotowanego wcześniej szablonu. Dodatkowym krokiem jest wygenerowanie strony z indeksem wszystkich typów, aby użytkownik mógł wygodnie wyszukiwać.

Strony generowane są przy wykorzystaniu biblioteki *Scalate* [Sca12a] oraz silnika **Mustache** dostarczanego przez nią. Mustache jest prostym językiem definiowania szablonowym (ang. *template*), który umieszcza się np. w HTMLu. Silnik mustache następnie jest w stanie iterować np. po dostarczonej mu liście pól danej wiadomomości oraz wyświetlać fragment szablonu dla każgego z nich.

Na tym kończy się odpowiedzialność `ProtoDocTemplateEngine` — proste zmapowanie pól obiektów, na odpowiadające im nazwy w szablonach. Przykład przekazania informacji o typie dla strony o typie enumeracji można zobaczyć na Listingu 4.12. Metoda `engine.layout` zwraca wygenerowaną stronę jako String, pozostaje jedynie ją zapisać na dysk.

Listing 4.12: Przykład zastosowania Scalate (z silnikiem renderowania szablonów Mustache)

```
def renderEnumPage(enum: ProtoEnumType) = {
  import enum._

  val data = Map("enumName" -> typeName,
    "packageName" -> packageName,
    "comment" -> comment,
    "values" -> values)

  engine.layout("enum".mustache, data)
}
```

4.3.1. Język szablonów - Mustache

Celem krótkiego przedstawienia języka Mustache [Mus12], oraz uzasadnienia wybrania takiego silnika renderującego szablony chciałbym przedstawić kilka elementów Mustache. Posłużymy się jedynie kilkuliniowymi przykładami, aby nie zaciemniać obrazu HTMLem który aż tak interesujący z naszej perspektywy nie jest.

Nazwa silnika *Mustache* pochodzi od angielskiego słowa oznaczającego wąsy. Może się to wydawać dziwne, lecz po pierwszym spojrzeniu na znaczniki mustache etymologia nazwy staje się oczywista: Wszystkie odwołania do zmiennych, jak i operacje warunkowe obejmowane są następującym znakiem: `{{  }}`, przypominającym wąsy. Konkretny przykład zastosowania umieszczania zmiennych w HTML można zobaczyć na Listingu 4.3.1.

```
<title>ProtoDoc for: {{packageName}}.{{fieldName}}</title>
```

Warunki natomiast zapisuje się przy pomocy znaku rozpoczynającego warunek: `{{#boolean_var}}` (lub jego odpowiednikowi negującemy wartość w sprawdzanej zmiennej: `{{^boolean_var}}`) oraz kończącego blok warunkowy `{{/boolean_var}}`. Identyczną składnią (`{{#fields}}`) można iterować po wartościach zawartych w liście, co w ProtoDoc ma miejsce podczas np. renderowania tabeli z informacjami wszystkich pól danego typu.

Mustache udostępnia odrobinę więcej funkcjonalności niż omówione powyżej, jednak w przypadku generowania prostych szablonów, z jakimi mamy do czynienia w ProtoDoc, nie były one konieczne do zastosowania.

5. Przykład zastosowania ProtoDoc w realnym projekcie

W poniższym rozdziale zostanie przedstawione zastosowanie narzędzia ProtoDoc, w parze z Apache Maven [Mav12], obecnym de facto standardem zarządzania dużymi projektami w świecie aplikacji klasy enterprise.

W tym celu został zaimplementowany dodatkowy „*plugin*" (ang. wtyczka) do Maven, pozwalający na automatyczne wykonanie ProtoDoc podczas procesu budowy projektu. Zostanie również przedstawiony sposób w jaki ten plugin można skonfigurować, z poziomu pliku konfiguracyjnego Mavena.

5.1. Przykład połączenia ProtoDoc z Maven, celem automatycznego generowania dokumentacji

Ponieważ takie narzędzie jak *ProtoDoc* powinno być stosowane w sposób w pełni automatyczny, logicznym krokiem w rozpowszechnieniu jego użycia była integracja z najpopularniejszym (de facto „standardowym") narzędziem zarządzania cyklem życia projektu, jakim w świecie Javy jest **Maven**.

W związku z powyższym w ramach projektu został zrealizowany plugin do narzędzia Maven. Wspierane są zarówno wersje 2.x jak i 3.x tego narzędzia. Nie będziemy tutaj wnikać w szczegóły i zasady działania narzędzia Maven, ponieważ jest to bardzo obszerny temat, jednak przedstawię jak można wykorzystać *ProtoDoc* wraz z Mavenem do automatycznego generowania dokumentacji wszystkich plików proto umieszczonych w projekcie.

Wszystkie zależności jak i pluginy deklaruje się w mavenie w pliku **pom.xml**. W naszym przypadku konieczne będzie dodanie elementu `<plugin>`, wewnątrz `project -> build -> plugins`.

Listing 5.1: Deklaracja korzystania z pluginu ProtoDoc

```
<plugin>
  <groupId>pl.project13.maven</groupId>
  <artifactId>protodoc-maven-plugin</artifactId>
  <version>1.0</version>
  <executions>
    <execution>
      <goals>
        <goal>package</goal>
      </goals>
    </execution>
```

Listing 5.2: Kontunuacja listingu 5.1

```
      </executions>
  <!-- Optional overrides of default properties:
      <configuration>
        <protoDir>${project.basedir}/src/main/proto</protoDir>
        <outDir>${project.basedir}/target/protodoc</outDir>
        <verbose>false</verbose>
      </configuration> -->
    </plugin>
```

Jest to wystarczające aby ProtoDoc mógł **automatycznie**, podczas budowania projektu generować dokumentację dla wszystkich plików proto umieszczonych wewnątrz domyślnego folderu - **src/main/proto**. Ścieżka ta jest zgodna z konwencjami przyjętymi w Maven, jednak w razie potrzeby istnieje możliwość nadpisania tego ustawienia poprzed dodanie elementu configuration, tak jak to przedstawiono na Listingu 5.1.

5.2. Szczegóły implementacyjne

Uważny czytelnik zauważył że przeniknęliśmy właśnie ze świata **Scala** do świata **Javy**. Wartym podkreślenia jest jak sprawnie da się korzystać z obu tych języków „jednocześnie", dzięki dobrodziejstwom platformy JVM [1] — wspólnego runtime na którym różne języki bezproblemowo mogą się między sobą komunikować.

Dzięki poprzednim doświadczeniom w tworzeniu pluginów mavenowych, oraz przygotowaniu ProtoDoc w taki sposób aby był bardzo łatwo używalny nie tylko z poziomu command line, ale również poziomu kodu źródłowego, integracja z mavenem przebiegła bardzo prosto. Pliku pom.xml zawierającego zależności pluginu nie będę tutaj umieszczał z racji jego rozmiaru (150 linii), jednak okazuje się że cała implementacja pluginu, bezproblemowo nadaje się to umieszczenia w tym miejscu.

Listing 5.3: Pełna implementacja pluginu mavenowego korzystającego z ProtoDoc

```
public class ProtoDocMojo extends AbstractMojo {
  /** @parameter default-value="${project.basedir}/src/main/proto" */
  private String protoDir;
  /** @parameter default-value="${project.basedir}/target/protodoc" */
  private String outDir;
  /** @parameter default-value="false" */
  private boolean verbose;

  public void execute() throws MojoExecutionException {
    ProtoDocMain.generateProtoDoc(protoDir, outDir, verbose);
  }
}
```

[1] JVM - Java Virtual Machine

Ciekawostką jest że pierwotne wersje Maven2, były tak stare, iż nie były jeszcze dostępne adnotacje w Javie. Stąd uciekano się do *meta-programowania* w komentarzach, poprzez tak zwane docklety — adnotacje (słowa poprzedzane znakiem `@`) umieszczone nad zmiennymi prywatnymi powyższego Mojo faktycznie **są znaczące** oraz deklarują iż w te zmienne powinny zostać wstrzyknięte przez mavena odpowiednie ustawienia z sekcji `<configuration/>` pluginu.

5.3. Efekt działania ProtoDoc wraz z Maven

Efektem zastosowania powyższego pluginu w projekcie mavenowym, jest automatyczne wygenerowanie się dokumentacji. Aby unaocznić bardziej strukturę plików oraz proces korzystania z Apache Maven [Mav12], poniżej przedstawiam krótką sekwencję komend wydaną w zgognej z POSIX linii poleceń, obrazującej działanie mavena oraz samego pluginu.

```
$ tree
+-- pom.xml
+-- src
    +-- main
        +-- java
        |  +-- pl
        |      +-- project13
        |          +-- ProtoDocMojo.java
        +-- proto
            +-- amazing_message.proto
            +-- common_message.proto
            +-- simple.proto
$ mvn clean install
# ............. maven output .............
$ tree
+-- pom.xml
+-- src/ ...
+-- target
    +-- clases/ ...
    +-- protodoc
    |   +-- images/ ...
    |   +-- js/ ...
    |   +-- index.html
    |   +-- pl.project13.AmazingMessage.EnumType.html
    |   +-- pl.project13.CommonMessage.html
    |   +-- pl.project13.MessageWithInner.html
    |   +-- pl.project13.MessageWithInner.InnerMessage.html
```

Powyższy listing ładnie obrazuje jaką strukturę katalogów generuje ProtoDoc w efekcie działania. Folder `target` jest w projektach mavenowych dedykowany wszystkim produktom danego procesu budowania projektu. ProtoDoc umieszcza swoje pliki analogicznie jak uczyniłby to JavaDoc plugin (generujący dokumentację dla kodu zapisanego w języku Java), wewnątrz folderu `target/` z własnym sub katalogiem o nazwie `protodoc`.

Aby zobaczyć jak wygenerowana dokumentacja wyląda na żywo, można uruchomić na przykład poniższą komendę, uruchamiającą przeglądarkę Google Chromium:

```
$ chromium-browser target/index.html
```

Przykładowe zrzuty ekranu z działającej aplikacji przedstawione zostały w na następnej stronie, w seksji 5.4.

5.4. Zrzuty ekranu wygenerowanej dokumentacji

W tej sekcji zostały umieszczone zrzuty ekranu z przykładowych wygenerowanych stron dla różnych typów wiadomości.

Rysunek 5.1: Widok wygenerowanej strony dla typu **message**

Na Rysunku 5.1 przedstawiona została strona wygenerowana na podstawie prostej wiadomości. Widok wiadomości składa się na jego nazwę, wyróżnioną jako nagłówek strony, dokumentację umieszoną na jego typie, informacji o pakiecie w którym została ona zdefiniowana oraz liście pól oraz wewnętrznych enumeracji oraz wiadomości zdefiniowanych w niej. Każde z pól dokumentowane jest swoją nazwą, odpowiednim typem na platformie JVM (co w przypadku typów definiowanych przez użytkownika może być praktyczne), oraz oczywiście właściwą dokumentacją umieszczoną na danym polu.

Widoczna jest również po lewej stronie lista wszystkich wiadomości oraz enumeracji przeparsowanych podczas generowania dokumentacji dla tego projektu. Dodatkowo możliwe jest wyszukiwanie po wspomnianych elementach poprzez umieszone nad nimi pole temu służące. Wyszukiwarka ta jest praktyczna ponieważ filtruje ona listę wiadomości po dowolnym podciągu znaków, dzięki czemu łatwo jest znaleźć wszystkie wiadomości mające na przykład w nazwie „Person”.

Rysunek 5.2: Widok wygenerowanej strony dla typu **enum**

Na Rysunku 5.2 przedstawiona jest strona enumeracji. Celem pokazania całej tej strony, a nie „całości” widoku jak miało to miejsce na Rysunku 5.1 została tutaj pominięta lewa ramka (spis typów). Z racji, iż mamy do czynienia z enumeracją, kolor strony jest zmieniony oraz ikonka **M** została zmieniona na **E**. Dalsze części strony są zbudowane analogicznie - wpierd dokumentaja umieszczona na typie enumeracji a następnie lista jej pól, które w przypadku enumeracji należy rozumieć jako wartości. Każda wartość może mieć przypisaną dokumentację, umieszczoną w prostokącie poniżej jej nazwy, oraz wartość taga.

6. Rola testów w procesie tworzenia aplikacji

Rozdział ten ma za cel przybliżyć czytelnikowi jak istotną rolę w rozwoju *ProtoDoc* miały testy jednostkowe (oraz w pewnym sensie również integracyjne). Dzięki zastosowaniu poniżej opisanej metodyki, możliwe było otrzymanie kodu którego większość funckojalności jest pokryta testami a API jest na tyle stabilne, iż wprowadzanie zmian w jednym miejscu nie ma znaczącego wpływu na pozostałe komponenty.

W pierwszej sekcji przedstawiona zostanie metodyka *Test Driven Development*, w skrócie TDD, oraz jak należy ją stosować.

W drugiej sekcji zostanie przedstawione wyjście jakie ujrzy się w momencie uruchomienia testów w ProtoDoc. Zaskakujące może się wydać jak czytelne są wiadomości logowane przez framework testowy ScalaTest [Sca12b].

6.1. Metodyka TDD i jej pozytywny wpływ na projekt

Projekt prowadzony był zgodnie z zasadami Test Drivem Development (zwanego dalej *TDD*), co znacznie ułatwiło ustabilizowanie API oraz głównych konceptów jeszcze we wczesnych etapach tworzenia aplikacji. Ponad to, metodyka ta umożliwiła pracę z dotychczas nieznanym mi API bez obaw o zniszczenie zaimplementowanych wcześniej funkcjonalności.

Metodykę *TDD* możnaby opisać jako cylk składający się z trzech faz:

– napisanie najpierw (sic!) testu, sprawdzającego automatycznie czy stawiane przed nami oczekiwanie zostało spełnione

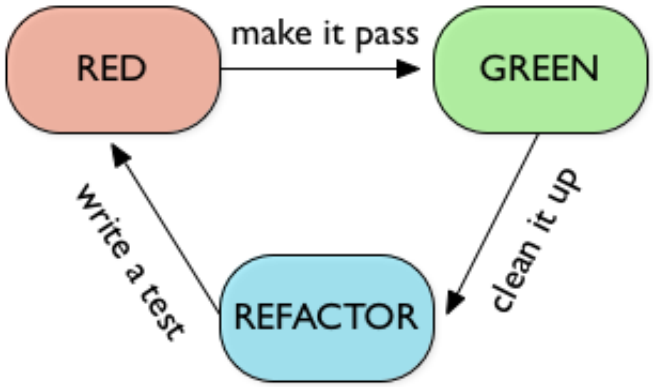
  pewną sub-fazą jest upewnienie się że test faktycznie na stan obecny aplikacji nie przechodzi. Najlepiej aby wiadomość niepowodzenia jasno wskazywała na to co jest przyczyną problemu. Jest to istotne nie tyle teraz, podczas implementacji, jednak podczas dalszego rozwoju aplikacji, kiedy to być może sprawimy, że warunek sprawdzany ten test przestanie być prawdziwy - wówczas, *„kilka tygodni później”*, pomocny komunikat o przyczynie problemu znacznie przyśpieszy zlokalizowanie oraz naprawienie problemu.

– implementacji funkcjonalności, tak aby warunki w teście zostały spełnione.

  należy pamiętać aby była to implementacja minimalna - nie wolno wychodzić „do przodu” z implementacją, nawet jeżeli uważa się, że pewna funkcjonalność *prawdopodobnie* będzie niebawem implementowana.

– oraz refaktoringu właśnie zaimplementowanych komponentów aplikacji, lub zauważonych podczas implementacji ewentualnych powtórzeń kodu itp.

Fazy te w literaturze znane są jako „Red - Green - Refactor”, i obrazuje się ją przy pomocy przedstawionego na Rysunku 6.1 grafu.



Rysunek 6.1: Schemat obrazujący fazy pracy w metodyce *TDD* (źródło: własne)

Przedstawiony powyżej cykl zazwyczaj trwa pomiędzy kilkoma a trzydziestoma minutami. Technika ta jest ściśle związana z samo-dyscypliną programisty i stosunkowo trudna do zastosowania w przypadku nie stosowania jej na codzień - jednak rezultaty, pod postacią wzrostu jakości kodu oraz zmniejszeniu czasu traconego na poszukiwania błędów są znaczne.

Oprócz pisania testu zanim powstanie jakakolwiek implementacja, bardzo ważnym elementem fazy implementacji jest jest aby jej celem było napisanie *minimalnej ilości kodu doprowadzając test to „przejścia”* (spełniania wymagań w nim stawianych). Przykładowo, nie dozwolone jest implementowanie dodatkowych funkcjonalności („na zapas”), nawet jeżeli uważa się iż będą niebawem konieczne podczas fazy implementacji związanej z właśnie napisanym testem. Faza implementacji nie może zostać zakończona w przypadku uszkodzenia (sprawienia że inny niż obecnie rozwijany test „nie przejdzie”).

Dzięki zastosowaniu tej metodyki, nie dość że kod tworzyło się łatwiej – dzięki skupianiu się na konkretnym celu, dostarczającym konkretnych wartości dla projektu – ale również podczas wprowadzania dużych zmian, mogłem być pewien, że nie uszkadzam przypadkiem istniejących oraz działających sprawnie elementów programu. Szczerze zalecam stosowanie tej metodyki, a nawet jeżeli nie czystego TDD, które może byż z początku przytłaczającym podejściem do wytwarzania oprogramowania, to z pewnością do samego rozpoczyniania pracy od napisania testu, a dopiero następnie przestępowania do programowania.

Ogółem w projekcie zostało zaimplementowane ponad 46 testów, w których zastosowano ponad 200 asercji.

6.2. Zaimplementowane specyfikacje

Ponieważ wyjście z Runnera („tego który uruchamia") testów stosowanego przezemnie w tym projekcie, są bardzo czytelne, oraz ładnie dokumentują jakie funkcjonalności dokładnie zostały zaimplementowane oraz automatycznie przetestowane. Poniżej umieszczone zostały jedynie niektóre z wiadomości, ponieważ umieszczenie wszystkich zajęłoby niestety kilka stron – mam jednak nadzieję, że ich czytelność mile zaskoczy czytelnika oraz zachęci do stosowania tego typu podejścia do testowania aplikacji.

```
ProtoBufVerifierTest:
The Verifier should validate field types
- should detect an unresolvable field
  + Given a message with an invalid fieldtype
  + When the message is parsed and verified
  + Then the Verifier report it as invalid
  + And it should point out that the UnknownType is unresolvable
- should have no problems with resolvable field Type
  + Given a message with valid, resolvable fieldtype, defined before the messa
  + When the message is parsed and verified
  + Then the result should contain one HasResolvableField message
  + And the field should be resolved to the propper type

RealSimpleParsingTest:
Parsing of an real message, with outer enum
- should be parsed properly
  + Given A real proto file
  + When it is parsed
  + And it is verified
  + Then parsed size should be 2
  + And the inner message should be detected
  + And the inner message should be named properly
  + And the enum field should have the proper type resolved
  + And it's tag should be equal 3
  + And it's resolved type should be the outer enumeration
  + And the outer enum should be parsed and named properly
...
```

Jak widać, informacje drukowane podczas uruchamiania testów mogą wręcz stanowić swego rodzaju dokumentację jego zachowania. Są czytelne oraz zachowują wspólną formę dzięki czemu łatwo jest znaleźć przy jakich warunkach początkowych, można oczekiwać jakiego rezultatu działania aplikacji.

7. Podsumowanie

Celem powstania aplikacji ProtoDoc było udostępnienie programistom, pracującym z Google Protocol Buffers na codzień, narzędzia automatycznie dokumentującego tworzone przez nich wiadomości. Cel został osiągnięty poprzez implementację parsera oraz generatora kodu, działającego na zasadach podobnych do JavaDoc, znanego i popularnego narzędzia generującego dokumentację dla klas zapisanych w języku Java. Przy okazji implementacji tych dwóch komponentów okazało się, że dodanie trzeciego - weryfikatora, nie dość że byłoby bardzo proste, to również zwiększyłoby znacznie przydatność ProtoDoc.

W finalnej wersji aplikacji, weryfikator został zaimplementowany oraz posiada kilka podstawowych weryfikacji, takich jak sprawdzanie poprawności przypisanych wartości domyślnych, wartości tagów czy też raportowania napotkanych pól Deprecated (ang. „przestarzałych"). Weryfikacje stanowią największe potencjalne pole do rozwoju narzędzia, ponieważ można by zaimplementować liczne dodatkowe heurystyki sprawdzające poprawność tworzonej wiadomości pod kątem utrzymywalności w przyszłości.

Wszystkie funkcjonalności aplikacji, włączając w to parsowanie „nietypowych sytuacji" oraz wszystkie weryfikacje, sprawdzane na różnych przypadkach zostały pokryte testami automatycznymi. W momencie pisania projekt zawierał 49 testów, zawierających łącznie ponad 200 asercji, z których wszystkie podczas uruchamiania testów są spełniane. Dzięki rygotystycznie stosowanej podczas implementacji tego projektu metodyce Test Driven Development, udało się utrzymać kod na wysokim poziomie czystości oraz komponeny nie posiadające dużych zależności, co dobrze świadczy o przyjętym projekcie systemu. Kod oczywiście dzięki temu był i jest również prosty do testowania o czym świadczyć może wspomniana już duża liczba testów automatycznych, dających pewność, że zaimplementowane funkcjonalności faktycznie działają.

W ramach umożliwienia faktycznego stosowania ProtoDoc w rzeczywistości został również zaimplementowany plugin do systemu zarządzania cyklem życia projektu - Apache Maven [Mav12] - dzięki któremu dołączenie ProtoDoc do istniejących projektów jest bardzo proste – ogranicza się do konfiguracji pluginu w pliku konfiguracyjnym docelowego projektu. Wspomniany plugin został również dołączony do niniejszej pracy, jako jej naturalne rozszerzenie.

Podsumowując projekt zakończył się sukcesem – wszystkie wymagania dotyczące generowania dokumentacji oraz automatyzacji tego procesu zostały spełnione. Zakres projektu został nawet nieco poszerzony z racji dołączenia komponentu weryfikującego kod. Aplikacja posiada duży potencjał na dalszy rozwój w kierunku stania się używanym „w rzeczywistości" narzędziem, a punkty w których można by kontunuować ów rozwój zostały wymienione w kolejnej, ostatniej już, sekcji też pracy.

7.1. Plany rozwoju aplikacji

Aplikacja w przyszłości może zostać rozszerzona aby wspierać jeszcze większą część standardu Protocol Buffers, w szczególności, można wyróżnić niektóre interesujące zadania do wykonania:

- zaimplementować rozpoznawanie słowa kluczowego **option**, służącego przekazywaniu pewnych meta danych do kompilatora *protoc*,
- zaimplementować bardziej dokładne rozwiązywanie widoczności, biorące pod uwagę różne notacje z kropką w nazwie wiadomości w przypadku stosowania słowa kluczowego *import*,
- dobudować dodatkowy moduł, który dokumentowałby elementy **service**, które nie zostały pokryte w tej pracy,
- zaimplementować więcej, oraz bardziej wnikliwych weryfikacji, służących nie tylko jako sprawdzenie poprawności, ale jako rady oraz utrzymanie dobrego stylu implementowanych wiadomości,
- zaimplementować *ProtoDoc* jako plugin do innych systemów zarządzania cyklem życia projektu. Zadanie to jest stosunkowo proste, jak już przedstawiono na przykładzie Maven, a mogłoby znacznie pomóc społeczności użytkowników języków takich jak Groovy bądź Scala, gdzie Maven jest rzadziej wykorzystywany niż w projektach „czysto Javowych”.

Projekt planuję rozwijać w czasie wolnym, ponieważ uważam że potrzeba takiego narzędzia faktycznie istnieje oraz przy okazji jego implementacji nauczyłem się wiele o Scali jak i samych Protocol Buffers.

A. Google Protocol Buffers

W tym dodatku zostanie omówiona idea oraz szczegóły implementacyjne stojące za Google Protocol Buffers, dalej zwanym *ProtoBuf*. Omówione zostaną również przykładowe zastosowania, oraz dlaczego warto porzucić w niektórych sytuacjach komunikację przy pomocy XML bądź JSON, na rzecz Protocol Buffers lub mu podobnym binarnym protokołom komunikacji.

A.1. Zasada działania

Protocol Buffers powstało pierwotnie jako wewnętrzny projekt Google, mające na celu zmniejszenie ilości ruchu na łączach oraz zwiększenia szybkości z jaką wiadomości wysyłane między serwerami są serializowane oraz deserializowane. W momencie pisania niniejszej pracy wewnątrz Google znajduje się 48,162 typów wiadomości, umieszczonych w 12,183 plikach *.proto – są to imponujące liczby, obrazujące jak ważnym elementem infrastruktury Google stał się ProtoBuf. [Goo11]

Protocol Buffers, dalej zwane ProtoBuf, opierają się w dużej mierze o język definicji interfejsu (*IDL*) o tej samej nazwie. Przy jego pomocy definiuje się tak zwane wiadomości, które mogą zawierać pola lub zagnieżdżone wiadomości i/lub enumeracje. Tak przygotowany opis wiadomości, zapisywane jest tak zwanym proto-pliku („*proto-file*"), z rozszerzeniem `*.proto`. Następnie plik taki poddawany jest „kompilacji", przy pomocy narzędzia *protoc* (*Proto*col Buffers *C*ompiler) czego wynikiem są pliki źródłowe klas (o ile takie pojęcie istnieje w docelowym języku) które udostępniają bezpieczne (co do typów) API do manipulacji tymi wiadomościami. Najważniejszą funkcją wygenerowanych klas jest jednak to iż korzystając z nich, możemy otrzymawszy strumień bajtów, o którym wiemy że zawiera wiadomości tego typu, łatwo zdeserializować ów bajty to postaci obiektu tej klasy. Udostępniana jest równocześnie automatyczna metoda konwersji w drugą stronę, z obiektu do strumienia bajtów.

Jako przykład możemy spojrzeć na poniższą deklarację wiadomości (Listing A.1), zaczerpniętą z [Goo11]. Zawiera ona podstatowe typy pól, oraz wiadomość wewnątrzną.

```
message Person {
  required string name = 1;
  optional string email = 3;
  repeated PhoneNumber phone = 4;

  message PhoneNumber {
    required string number = 1;
    optional PhoneType type = 2 [default = HOME];
  }
}
```

Po wygenerowaniu z powyższego pliku *.proto kodu w języku C++ uzyskamy klasę którą można wykorzystać a następujący sposób:

Listing A.1: Przykład klasy ProtoBuf w C++, serializowanej do pliku

```
Person person;
person.set_name("John Doe");
person.set_id(1234);
person.set_email("jdoe@example.com");

fstream output("myfile.data", ios::out | ios::binary);
person.SerializeToOstream(&output);
```

A następnie możliwe byłoby dokonanie opisywanej wcześniej deserializacji z utworzonej właśnie tablicy bajtów. Ten przykład przedstawimy w języku Java, w ramach przypomnienia iż ProtoBuf jest niezależny od języka ani platformy:

Listing A.2: Przykład klasy ProtoBuf w Javie, odczytującej „się” z zapisanego wcześniej binarnego pliku

```
FileInputStream is = new FileInputStream("myfile.data");

Person person = Person.parseFrom(is);
assert person.getName().equals("John Doe");
assert person.getId().equals(1234);
assert person.getEmail().equals("jdoe@example.com");

is.close();
```

Analogicznie wynegerowane klasy można otrzymać w praktycznie dowolnym języku programowania, przy czym wspieranymi przez Google językami są C++, Java oraz Python. Do generowania klas dla pozostałych języków służą implementowane przez społeczność open source pluginy do kompilatora protoc.

Łatwo jest sobie wyobrazić powyższy przykład, piszący zamiast do pliku, na strumień będący odbieranym przez klienta po drugiej stronie. Tak właśnie w dużym przybliżeniu implementowane są web serwisy przy wykorzystaniu Protocol Buffers. Istnieją dodatkowe pomocnicze fabryki oraz metody korzystania z ProtoBuf jako zasobów sieciowych, na przykład implementacje Protocol Buffers w oparciu o standard JAX-RS [Pul08].

A.2. Wyjaśnienie znaczenia pól Protocol Buffers

Listing A.3: Przypomnienie wyglądu deklaracji pola wiadomości

```
required string name = 45 [ default = "sample" ];
```

W przedstawionej w 1 sekcji tego rozdziału wiadomości (Listing A.1) można zauważyć iż każde pole składa się z następujących elementów:

– **Kwalifikatora**, który może przyjąć wartości:

  required - oznaczającym że pole *musi* zostać ustawione. Generalnie odradza się jego stosowania, ponieważ pole raz oznaczone tym kwalifikatorem, musi zawsze pozostać required. Nawet jeżeli przestanie być kiedyś w przyszłości używane.

  optional - oznaczającym że pole może ale nie musi zostać ustawione. Jest to zalecany kwalifikator w większości przypadków.

  repeated - oznaczającym listę. Pole oznaczone tym kwalifikatorem może powtarzać się dowolną ilość razy w wiadomości.

– **Typu pola** - mogącym być jednym z predefiniowanych w Protocol Buffers typów (wylistowanych w Tabeli A.1, lub typem wiadomości lub enumeracji zdefiniowanym przez użytkownika.

– **Nazwy pola** - nazwa pola, poprawne indentyfikatory są zgodne z standardem stosowanym przez Javę

– **TAG**u - tag jest specjalną liczbą, nie mogącą powtarzać się w obrębie jednej wiadomości. Liczba ta jest wykorzystywana podczas kodowania wiadomości do jej możliwie najoptymalniejszej postaci. Warto pamiętać że umieszczenie często używanych pól jako pierwszych *ma wpływ* na ostateczny rozmiar wiadomości. Ponieważ w przypadku pól o wysokich tagach, które rzadko, lub nigdy nie są ustawiane, nie zostaną one w ogóle umieszczone w zserializowanej postaci wiadomości - oszczędzając cenne bajty.

– **default** - wartość domyślna *nie jest konieczna* jednak możliwa do podania przy każdym polu – sens ma oczywiście jednak tylko w przypadku podania jej dla pola będącego typu **optional**, ponieważ w przypadku **required** wartość ta i tak zostałaby nadpisana. W przypadku pola typu enumeracyjnego, możliwe jest podanie stałej z wewnątrz tej enumeracji jako wartości domyślnej.

sint32	bytes	double	uint64
sint64	string	float	uint32
fixed32	bool	int32	int64
sfixed32	sfixed64	fixed64	

Tablica A.1: Predefiniowane typy w ProtoBuf

Poza wspomnianymi predefiniowanymi typami, pola mogą oczywiście mieć typ innej wiadomości. Tym sposobem można modelować nawet skomplikowane modele domenowe w obrębie wiadomości ProtoBuf. Innym typem który mogą przyjąć pola, są type enumeracyjne. Deklaruje się je w podobny sposób jak wiadomości, jednak ich pola mają mniej elementów koniecznych do podania. Przykład deklaracji enumeracji można zobaczć na Listingu A.4.

Listing A.4: Deklaracja oraz zastosowanie typu enumeracyjnego

```
enum TheEnum {
  SMALL = 1;
  LARGE = 2;
}

message It {
  optional TheEnum itsName = 1 [ default = LARGE ]
}
```

Jak widać, możliwe jest również podanie nazwy `LARGE` jako wartości domyślnej pola.

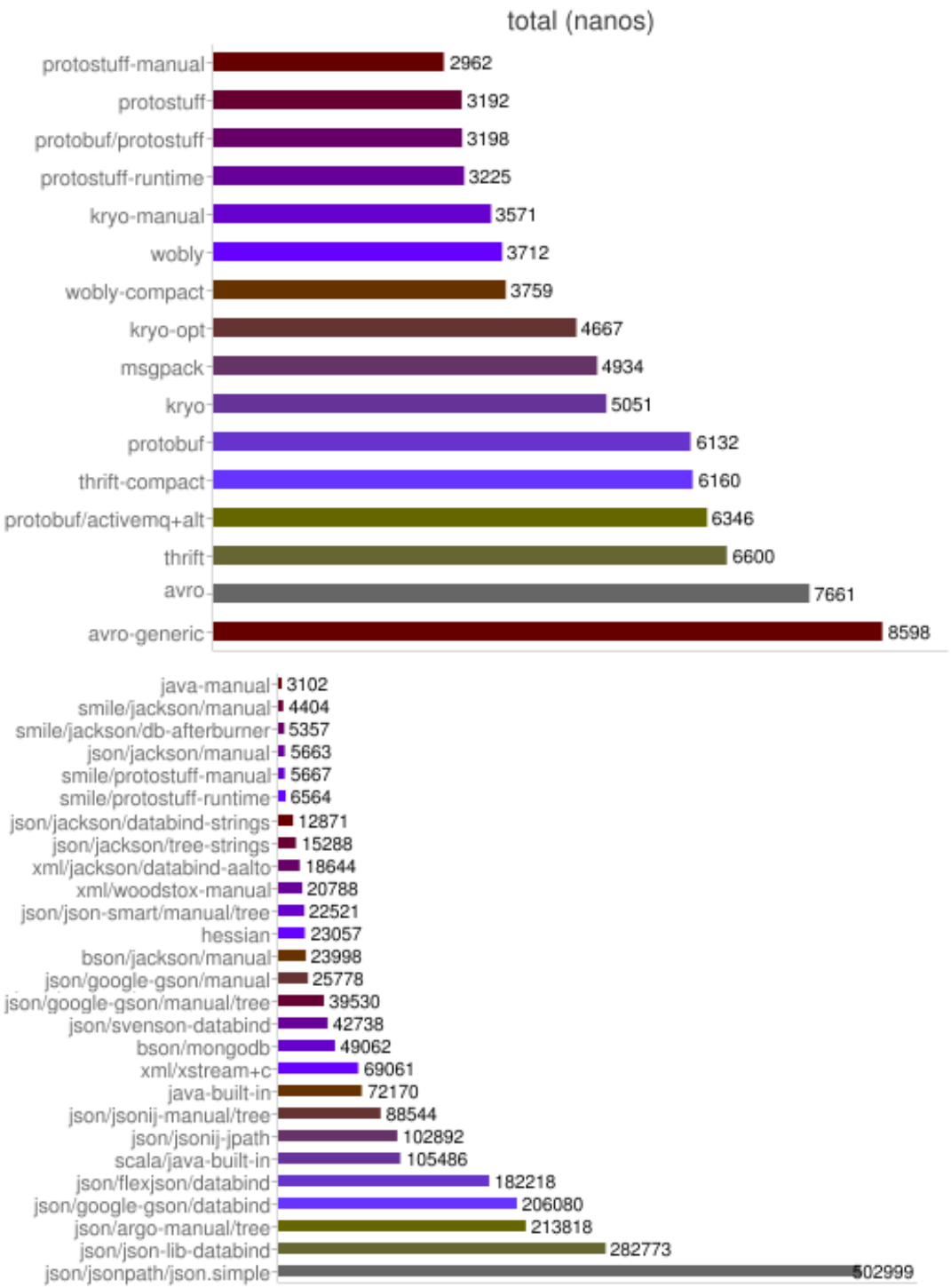
A.3. Zestawienie wydajności mechanizmów serializacji na JVM

Poniewać w efekcie, ProtoBuf służy efektywnej serializacji i deserializacji obiektów, bardzo ważna jest jego wydajność pracy w rzeczywistych warunkach, oraz porównanie go z innymi dostępnymi rozwiązaniami. Benchmark taki został przeprowadzony przez Pana w ramach projektu **jvm-serializers** [Smi11] Testy takie przeprowadzono w ramach projektu umieszczonego na githubie pod adresem. Benchmark ten został wykonany w roku 2011, oraz uwzględnia najważniejsze metody serializacji danych na JVMie, takie jak natywna serializacja Java, serializacja danych do JSON przy pomocy różnych bibliotek oraz oczywiście ProtoBuf oraz jego konkurencji - Thrift.

Serializacja + Deserializacja:

Czas spędzony na serializacji oraz deserializacji tego samego obiektu został przedstawiony na Rysunku A.1.

ProtoStuff, alternatywna implementacja protokołu ProtoBuf wiedzie prym w tym benchmarku, jednak protobuf nadal jest stanowczno wiadącą implementacją. Najważniejszą obserwacją tutaj jest jednak porównanie binarnych protokołów, które zmieściły się na powyższym wykresie, do metod serializacji JSONowej. Dla przykłądu Google Gson, jedna z popularniejszych bibliotek serializacji Java -> JSON,

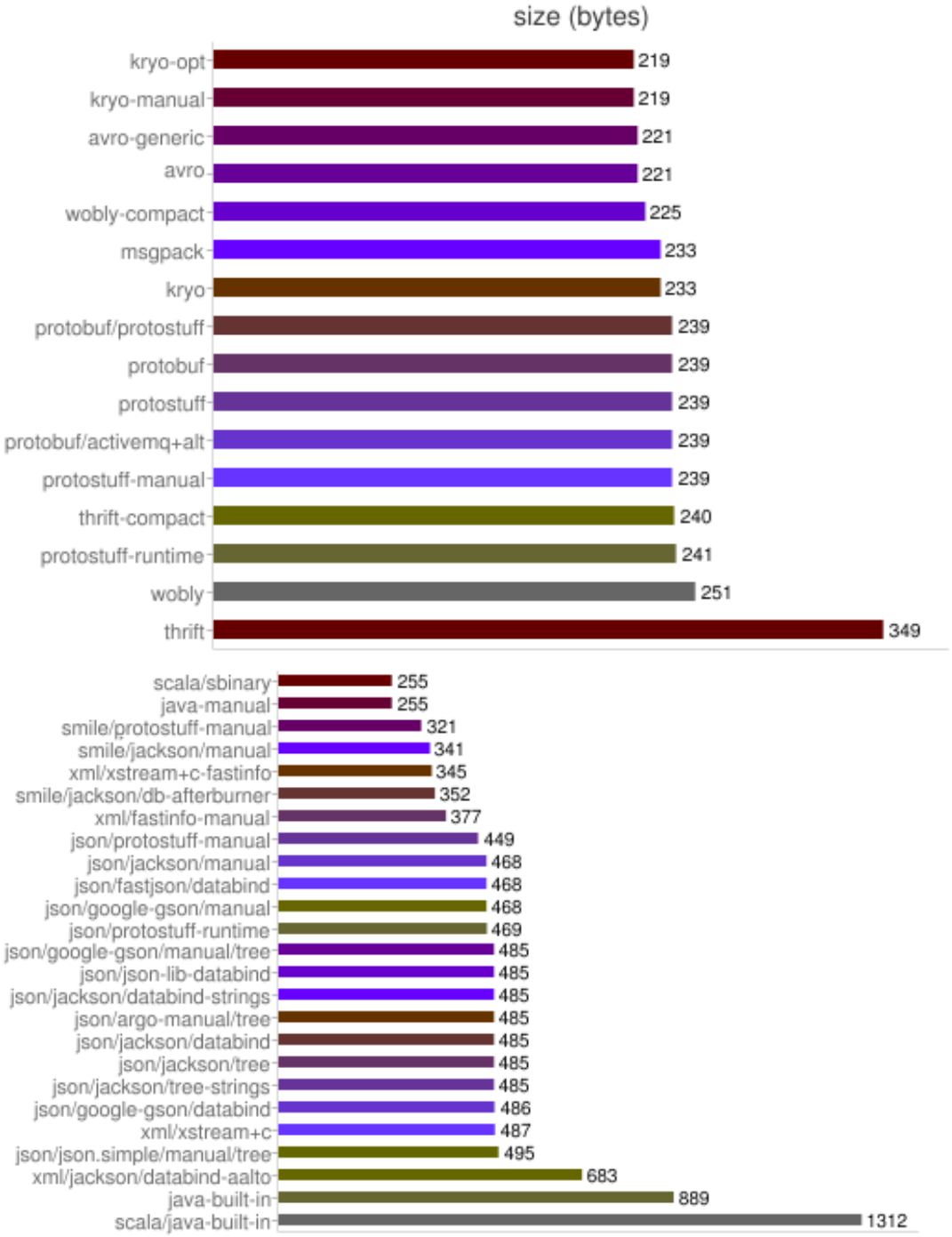


Rysunek A.1: Wykresy czasu serializacji + deserializacji obiektu przy zastosowaniu różnych bibliotek (źródło: [Smi11])

tą samą serializację przeprowadziła w **25778ns**, co w porównaniu z czasem osiągniętym przez ProtoBuf, równym **6132ns** jasno wyraża przewagę z jaką ProtoBuf wyprzedza serializację „klasyczną”.

Rozmiar zserializowanej wiadomości:

Rysunek A.2 przedstawia natomiast rozmiar wiadomości po serializacji przy wykorzystaniu wybranych bibliotek (liczony w bajtach).



Rysunek A.2: Wykresy rozmiaru rozmiaru zserializowanego obiektu przy zastosowaniu różnych bibliotek (źródło: [Smi11])

Przy porównywaniu rozmiaru zserializowanego obiektu ponownie protokoły binarne, z ProtoBuf na czele wiodą prym... Konkurencja typu serializacji Javowej lub JSON niestety nie ma tutaj wiele do powiedzenia, powodem jest iż Java serializując obiektu musi zatrzymać wszystkie nazwy klas aby potrafić

się do nich odwołać podczas deserializacji; W przypadku JSONa, mamy do czynienia z dużą ilością nadmiarowych danych w postaci reprezentowania liczb jako tekst, oraz dużej ilości nadmiarowych znaków w postaci { oraz }. To samo można powiedzieć o wszelkich tekstowych reprezentacjach danych, które tłumaczą każde pole danych jako element XML - niosący ze sobą jeszcze większą ilość nadmiarowych danych (elementy otwierające i zamykające dane pole). Ciekawostą jest natomiast to iż nawet tak ociążała postać danych jaką jest XML, ma szansę wyprodukować mniejszy rozmiach zserializowanego obiektu niż mechanizmy domyślnie stosowane przez języki takie jak Scala lub Java. Przyczyną tutaj jest konieczność serializacji przez mechanizmy bazujące bezpośrednio na JVM, również wszystkich nazw typów każgego z pól które jest serializowane – tym samym, dla każdego pola będącego ciągiem znaków, dodatkowo w zserializowanej postaci obiektu pojawi się ciąg znaków „java.lang.String”.

A.4. Przykładowe definicje wiadomości

Poniżej zostanie przedstawione kilka przykładowych definicji wiadomości, celem zapoznania czytelnika z formatem i składnią protobuf w sposób bardziej praktyczny (wizualny). Przykłady te modelują pytanie oraz odpowiedź, jakie mogłyby być używane podczas używania wyszukiwarki Google.

Listing A.5: Definicja wiadomości zawierająca enum oraz wartości domyślne

```
message SearchRequest {
  required string query = 1;
  optional int32 page_number = 2;
  optional int32 result_per_page = 3 [default = 10];

  enum Corpus {
    UNIVERSAL = 0;
    WEB = 1;
    IMAGES = 2;
    LOCAL = 3;
    NEWS = 4;
    PRODUCTS = 5;
    VIDEO = 6;
  }

  optional Corpus corpus = 4 [default = UNIVERSAL];
}
```

Powyższy listing obrazuje jak mogłaby wyglądać wiadomość Protocol Buffers stosowany przez wyszukiwarkę Google. SearchRequest przedstawiałby w takim przypadku żądanie nadane w efekcie przez użytkownika, wgłąb systemu.

Listing A.6: Definicja wiadoomści wykorzystującej inną wiadomość

```
message SearchResponse {
  repeated Result result = 1;
}

message Result {
  required string url = 1;
  optional string title = 2;
  repeated string snippets = 3;
}
```

Listing A.7: Definicja wiadomości korzystającej z wewnętrznej wiadomości

```
message SearchResponse {
  message Result {
    required string url = 1;
    optional string title = 2;
    repeated string snippets = 3;
  }
  repeated Result result = 1;
}
```

Więcej przykładów można zobaczyć na stronie domowej Protocol Buffers - http://code.google.com/intl/pl-PL/apis/protocolbuffers/docs/proto.html.

B. Podstawy języka Scala oraz Scala Parser Combinators

Celem tego dodatku jest przybliżenie czytelnikowi języka „Scala” aby w wystarczająco płynny sposób mógł czytać przykłady kodu używane w tym dokumencie.

B.1. Krótka historia języka

Język Scala („Scalable Language”) najłatwiej jest przedstawić jako hybrydę dwóch znanych nurtów programowania: programowania obiektowego oraz funkcyjnego, wraz z powiązanymi z nimi językami programowania. Twórca języka Scala, Martin Oderski [1] był ściśle związany z językiem Java - był głównym projektantem generyków w Javie (*Java Generics*) oraz głównym autorem utrzymywanej po dziś dzień serii kompilatorów **javac** [MO].

Jako konkretnych „rodziców” można by wskazać:

– **Java** - jako reprezentant nurtu obiektowego

– oraz języki: **Haskell**, **SML** oraz pewne elementy języka **Erlang** (głównie *Actor model*).

O języku Scala można myśleć jako połączeniu tych nurtów. Dostępne są klasyczne elementy języków funcyjnych, takie jak pattern matching czy nacisk na immutability wszystkich tworzonych obiektów. Jest to zwłaszczw widoczne w domyślnych implementacjach kolekcji, których nie można modyfikować - a tworzy się za każdym razem nową kolekcją, współdzieląc między różnymi kolekcjami elementy które są niezmienne.

B.2. Podstawy

Ta sekcja służy przybliżeniu czyletnikowi języka *Scala* na poziomie wystatczającym aby swobodnie czytać przykłady kodu umieszczone w tej pracy. W niektórych przykładach pomijane są przypadki skrajne lub nietypowe, celem szybkiego oraz jasnego przedstawienia minimum wiedzy na temat języka aby móc swobodnie go „czytać”.

Scala jest językiem statycznie typowanym posiadającym lokalne „Type Inferrence”. Pozwala to kompilatorowi *scalac* na „odnajdywanie” typów wszystkich zmiennych oraz typów zwracanych przez metody podczas kompilacji, bez potrzeby definiowania ich wprost (poza konkretnymi wyjątkami np. funkcji wprost rekurencyjnych).

[1] Martin Odersky - Strona domowa: http://lamp.epfl.ch/ odersky/

Użycie nawiasów `()`, średnika `;` oraz kropki `.` jest analogiczne jak w przypadku Javy, jednak w wielu przypadkach opcjonalne gdyż kompilator jest w stanie wydedukować gdzie powinny się znaleźć.

```
val value: Int = Option(42);
val other: Int = value.orElse(0);

// moze zostac zastpione
val value = Option(42)
val other = value orElse 0 // "infix notation"
```

Jednym z ciekawych przykładów stosowania notacji bez nawiasów i kropek jest *ScalaTest*[2] (przy którego pomocy pisano testy w tym projekcie). Przykładowa *asercja* napisana w *DSL*u definiowanym przez tą bibliotekę wygląda następująco:

```
messages should (contain key ("Has") and not contain value ("NoSuchMsg"))
```

Dostępne jest wiele sposobów definiowania metod / pól w klasie, w efekcie (na poziomie bytecode), wszystkie przekładane są na wywołania metod. Dostępne są słowa kluczowe:

– **def**, *def*iniujący zwyczajną metodę instancyjną. Warto nadmienić że Javowa koncepcja pojęcia *static* nie jest dostępna z poziomu Scala.

– **val**, deklarujący „stałą” - to jest metodę która raz zawołana, zwróci wartość oraz pole to będzie konsekwentnie zwracać tą samą wartość. Dodatkowym efektem jest traktowanie zmiennych tego typu analogicznie do Javowych zmiennych z modyfikatorem **final**.

– **var**, deklaruje zwyczajną „zmienną”, do jakiej przyzwyczajeni jesteśmy z Java.

– modyfikator **lazy**, wpływający na moment inicjalizacji zmiennej - metody zadeklarowane z modyfikatorem **lazy** zostaną dopiero zainicjalizowane podczas pierwszego odwołania się do tego pola z innego miejsca w kodzie. W przypadku pary **lazy val**, metoda ta zostanie zawołana jedynie jednokrotnie, a zwrócona po raz pierwszy wartość zostanie zapisana w cache oraz będzie konsekwentnie zwracana podczas ponownych wywołań tej metody.

Modyfikator lazy pozwala na budowę eleganckich konstrukcji z jakimi mamy do czynienia w przypadku na przykład Parser Combinators, omówionych szczegółowo w kolejnych sekcjach.

B.3. Traits - wmieszanie zachowania do klasy

Słowo kluczowe **trait** rozpoczyna definicję typu zwanego traitem. Implementacja nie różni się (na potrzeby tego szybkiego omówienia) od implementowania klasy, jednak różnica jest podczas „dziedziczenia” przy wykorzystaniu traitów. Nie mówimy bowiem o „dziedziczeniu” w przypadku *trait*ów, a o „wmieszaniu” (ang. *mixin* - wmieszanie) zachowania do klasy konkretnej.

[2] ScalaTest - framework do testowania - http://www.scalatest.org

Poniżej został przedstawiony najprostrzy trait zawierający jakieś zachowanie, oraz jeden ze sposobów jego wmieszania do klasy konkretnej. Warto zauważyć że w przypadku wmieszania *trait*a `A` do klasy `Test`, wprowadzamy między nimi relację „Test **IS-A** A", analogicznie jak w przypadku dziedziczenia.

```
trait A {
  def test = "A" // definicja metody zwracajcej "A"
}

class Test extends A { } // wmieszanie A

(new Test).test // skompiluje i wykona sie poprawnie
```

Co ciekawe, nie zauważamy różnicy w przypadku składni odnoszącej się do dziedziczenia dwóch klas konkretnych, oraz wmieszania traita. Składnia ulega zmianie w przypadku korzystania z więcej niż jeden trait lub domieszania traita do klasy która już dziedziczy po innej klasie, wówczas zamiast słowa kluczowego **extends** należy stosować **with** (nie dozwolone jest wielokrotne zapisanie **extends**, jednak wielokrotne **with** są często spotykane). Przykład wmieszania większej ilości traitów zostanie przedstawiony poniżej.

Jest to namiastka dziedziczenia wielobazowego jednak Scala dzięki swojemu bardzo rygorystycznemu kompilatorowi jest w stanie uniknąć sytuacji gdzie dziedziczenie wielobazowe byłoby niebezpieczne (klasyczne przykłady problematycznych sytuacji w przypadku dziedziczenia wielobazowago można przeczytać w „Symfonii C++", autorstwa pana Gręboszа [Gre08]).

Kompilator *scalac* przy napotkaniu konflintów nazw mogących doprowadzić do niejasności „którą metodę należy zawołać", nie skompiluje takiego kodu oraz poprosi o rozwiązanie konfliktu w sposób explicite. Jako przykład rozważmy dwa *trait*y udostępniające metodę `def test: String`:

```
trait A { def test = "A" }
trait B { def test = "B" }

class Example extends A with B {
  // blad kompilacji!
}
```

Przy napotkaniu problemu tego typu kompilator zgłosi:

```
error: overriding method test in trait A of type => java.lang.String;
                    method test in trait B of type => java.lang.String needs `ov
class Example extends A with B {
```

Dzieje się tak ponieważ **scalac** próbuje odnaleźć która metoda powinna mieć większą wagę, a tymsamym powinna zostać wywołana. Ponieważ nie jesteśmy w stanie dodać modyfikatora **override** do żadnego z *trait*ów (ponieważ nie nadpisują one tej metody, a jedynie deklarują), jedynym możliwym miejscem na rozwiązanie tego konfliktu jest uzupełnienie `Example` o następujący fragment kodu, rutujący poprawnie nasze wywołanie metody:

```
class Example extends A with B {
 // selektywne odwolanie sie do metody konkretnego supertypu
 override def test = super[B].test
}

new Example().test // poprawne
```

B.4. Case Class oraz Pattern Matching

Klasy deklarowane przy pomocy `case class` niosą ze sobą pewne ułatwienia, które generuje za nas kompilator. Case klasy są wykorzystywane w bardzo wielu miejscach w ProtoDoc, ze względu na ich znaczną zwięzłość zapisu oraz wygodną współpracę z konstrukcją `match`, która zostanie za moment wyjaśniona.

Case klasę (pozwolę sobie przyjąć taki, mieszający dwa języki, sposób nazywania jej, z racji problematycznego sesownego przetłumaczenia słowa case (ang. przypadek) w tym zwrocie) definiujemy przy pomocy konstrukcji przedstawionej na Listingu B.1.

Listing B.1: Deklaracja case klasy

```
case class SampleProtoField(name: String, value: Long)
```

oraz odpowiada w przybliżeniu implementacji przedstawionej poniżej, na Listingu B.2:

Listing B.2: Ręczna implementacja case classy

```
class SampleProtoField {
 private def this(__name: String) = {
  this()
  _name = __name
 }

 private val _name = null

 def name = _name

 def equals = /**/
 def hashCode = /**/
 def toString = /**/
}

object SampleProtoField {
 def apply(name: String) = new SampleProtoField(name)
 def unapply(field: SampleProtoField) = field.name
}
```

Dzieje się tutaj bardzo dużo ciekawych rzeczy sięgających głęboko po możliwości Scala, jednak w efekcie umożliwia nam:

- domyślne utworzenie niezmiennych pól dla każdego argumentu w konstruktorze case klasy (val)
- automatyczne wygenerowanie getterów dla tych pól
- przyjemną dla oka implementację `toString()`
- implementacje `equals()` and `hashCode()` (programiści Java znają ból generowania tych metod ręcznie)
- wygenerowanie „companion object” pozwalającego na:

  tworzenie konstrukcji typu `SampleProtoField("")` zamiast `new SampleProtoField("")` (poprzez implementację apply)

  korzystanie z tej klasy w konstrukcji match (poprzez implementację unapply)

Najciekawsze dla nas są automatyczne implementacja apply / unapply, które w efekcie pozwalają na następujące linie:

```
val it: SampleProtoField = SampleProtoField("name") // apply

val SampleProtoField(name) = "some name"
```

A jeżeli sięgnąć po konstrukcję match, pozwala ona na konstrukcje „wyjmujące” wartości pól z skomplikowanych case class:

```
myCaseClass match {
  case SampleProtoField(name) => println(name)
  case ComplicatedProtoField(name, type, _, _) => println(name +" & "+ type)
}
```

Konstrukcja `match`, wywodzi się z programowania funkcyjnego. Mowa tutaj o „pattern matching” znanym z chociażby **Erlanga**. Dzięki tej konstrukcji da się ominąć wiele linii kodu w stylu `if(x) {x = x.getX(); method(x);}`. Technika ta została zastosowana w bardzo wielu miejscach aplikacji, włącznie z parserem oraz weryfikatorem.

B.5. Implicit Conversions - konwersje „domniemane”

Scala pomimo że jest językiem silnie statycznie typowanym pozwala na pewne zabiegi aby ułatwić pracę w tak rygorystycznym systemie typów. Jednym z tych rozwiązań są tak zwane „Implicit Conversions”, będące typem metod, które kompilator może próbować zastosować podczas gdy potrzebna jest automatyczna konwersja z typu A na B. Najłatwiej będzie omówić to na prostym przykładzie (Listing B.3, także spójrzmy na poniższe przypisanie liczby typu `scala.Int` do zmiennej typu `java.lang.String`:

Listing B.3: Przykład wystąpienia implicit conversion

```
val num: Int = 42
val string: String = num // compile time error!
```

Przykład przedstawiony na Listingu B.3 *nie skompiluje się*, a kompilator odpowiedziałby następującym komunikatem:

```
<console>:8: error: type mismatch;
found   : Int
required: String
      val string: String = num
```

Dopisanie implicit konwersji w zasięgu widoczności tego przypisania, pozwoli natomiast kompilatorowi zauważyć iż dostępna jest metoda potrafiąca przeprowadzić konwersję z typu `Int` na `String`, oraz ją zastosować. Implementację takiej konwersji przedstawiono na Listingu B.4.

Listing B.4: Implementacja oraz zastosowanie konwersji domniemanej — *Implicit Conversion*

```
implicit def num2str(num: Int) = num.toString
// == implicit def num2str(num: Int): String = num.toString

val num: Int = 42
val string: String = num // ok!
```

To co się rzeczywiście dzieje podczas kompilacji, to zwyczajne wstawienie metody num2str, w linii z przypisaniem liczby do zmiennej typu String, w następujący sposób: `num2str(num)`. Istnieje więcej szczegółowych zasad dotyczących konwersji domniemanych, na przykład która konwersja powinna zostać zastosowana w przypadku większej ilości metod pozwalających na poprawne wykonanie przypisania, jednak nie będziemy w nie wnikać, ponieważ na potrzeby zrozumienia zastosowanych DSL[3] sama informacja o istnieniu tych konwersji powinna być wystarczająca.

B.6. Scala Parser Combinators

W tym rozdziale zostaną pokrótce przedstawione Scala Parser Combinators, oraz ich składnia. Zostaną poruszone również tematy dotyczące wydajności kombinatorów oraz jak ustrzec się przez zanadto nawracającym się parserem.

Podstatową informacją którą należy sobie przyswoić podczas pracy z kombinatorami parserów, jest idea operowania na funkcjach wyższego rzędu - bo dokładnie tym są kombinatory. Najpierw jednak zdefiniujmy najprostszy możliwy parser (przedstawiony na Listingu B.5):

Listing B.5: Najprostszy możliwy parser

```
object SimplestParser extends JavaTokenParsers {
  def a: Parser[String] = "a"
}
```

Przedstawiona na listingu imponująca implementacja parsera jednego znaku, de facto dokonywana jest przez implicit konwersję (dostarczoną przez `JavaTokenParsers`), z typu `String` na typ `Parser[String]`. Zgodnie z oczekiwaniami „parsuje” on jedynie jeden znak - oraz zwraca samego siebie (wspomniane „a”). W

[3] DSL - Domain Specific Language

przypadku napisu dłuższego niż samo „a”, prymitywny parser możemy zdefiniować jako funkcję przyjmującą napis wejściowy a zwracającą krotkę przeprasowanego elementu, oraz pozostałej części napisu. W szczególności Parser zatem można zdefiniować jako:

Definicja 1
Parser zdefiniowany jest jako funkcja przyjmująca napis wejściowy, a zwracająca krotkę sparsowanego elementu oraz pozostałego napisu wejściowego.
[GH96]

Mając tak zdefiniowany parser, możemy przedstawić definicję kombinatora parserów:

Definicja 2
Kombinator parserów, jest *funkcją wyższego rzędu*, która przyjmuje jako argumenty funkcje parsujące, oraz zwraca nową funkcję utworzoną poprzez ich kombinację – zależną od jego implementacji.
[GH96]

W przypadku Scali, kombinatory implementowane są poprzez metody wypisane w Tabeli B.1, zdefiniowane w typie `Parser`.

Warto tutaj przytoczyć implementację dwóch kombinatorów, ponieważ pokazują jak olbrzymią elastyczność w definiowaniu ich uzyskać poprzez wykorzystywanie bardzo prostych kombinatorów (wewnętrznie zdefiniowanych `|||` oraz `&&&`, z których jako programista korzystający z Scala Parser Combinators raczej korzystać nie będziemy). Warte zauważenia na Listingu B.6 jest na przykład wywołanie `rep(p)`, otóż dzięki elastyczności Scali, możliwe jest zaimplementowanie Parserów w ten sposób że do wywołania rep(p) nie koniecznie musi dojść, pomimo że mamy do czynienia z zwyczajnymi wywołaniami metod. Umożliwia to tak zwany „pass-by-name parameter”, jednak znacznie wykracza to poza zakres konieczny do zrozumienia działania Parser Combinators, także pozostaniemy przy stwierdzeniu iż mamy tutaj do czynienia z leniwą ewaluacją tej metody.

Listing B.6: Implementacja parserów opt i rep

```
def opt(p: Parser): Parser = p ||| empty; // p ? p : empty
def rep(p: Parser): Parser = opt(rep1(p)); // p* = [p+]
def rep1(p: Parser): Parser = p &&& rep(p); // p+ = p ~ p*
```

Przy odpowiednim zastosowaniu tych prostych zasad, można bardzo łatwo tworzyć parsery nawet skomplikowanych gramatyk. Niestety z natury działania kombinatorów, możliwe jest wygenerowanie parsera który będzie wykonywał wielokrotne oraz głębokie nawroty. Celem zmniejszenia miejsc narażenia się na zbyt dużą klasę wygenerowanego parsera *LL(k)*, możemy zastosować kombinator `~!` który zgłosi błąd jeżeli dana kombinacja nie jest możliwa do rozwiązania w sposób generujący Parser klasy *LL(1)* w tym miejscu. Powinno się stosować ten operator jeżeli tylko możliwe, aby zagwarantować sobie wczesne ostrzeżenie o nie optymalnej formie gramatyki. Wówczas można przeprowadzić faktoryzację lewostronną, celem wyciągnięcia wspólnego symbolu dla dwóch lub więcej produkcji „przed nawias”, zmniejszając klasę parsera.

Listing B.7: Zastosowanie kombinatora transformującego

```
def TRUE: Parser[Boolean] = "true" ^^ { b => true }
```

```
//                              without ^^,
// def TRUE: Parser[String] = "true" // value is String!
```

Ostatnim elementem koniecznym do wprowadzenia zanim przejdziemy do przykładów pełnych implementacji prostych parserów, jest kombinator `^^`, oznaczający „*transformację*". Kombinator ten służyć nam będzie do produkowania typów które chcemy produkować z danego sparsowanego elementu, nie natomiast prymitówów takich jak listy i typy podstawowe, które zostałyby zwrócone przez parser bez zastosowania tego kombinatora. Prostym przykładem jest parser przedstawiony na Listingu B.7, który transformuje napis „true" na wartość typu `Boolean`.

Metoda	Nazwa	Przykład
`~`	sekwencja	`"a" ~ "b"` parsuje ciąg „ab", oraz zwraca „ab"
`~>`	sekwencja, odrzucając lewą stronę	`,,a'' ~> ,,b''` parsuje ciąg „ab", oraz zwraca „b"
`<~`	sekwencja, odrzucając prawą stronę	`,,a'' <~ ,,b''` parsuje ciąg „ab", oraz zwraca „a"
`|`	alternatywa	„a" | „b" parsuje ciąg „a" lub „b", oraz zwraca „a" lub „b"
`^^`	transformacja	`,,a'' ^^ () => Any` parsuje ciąg „a", oraz zwraca do, co zwróci funkcja przekazana jako argument tego kombinatora. Zostanie opisany szczegółowo poniżej.
`opt()`	sekwencja	opt(„b") parsuje ciąg „" lub „b", oraz zwraca Option[String], mogące zawierać „b" lub nic
`rep()`	powtórzenie	rep(„b") parsuje ciąg składający się z powtórzeń ciągu „b", oraz zwraca List(„b", „b", „b", ...)]
`repsep()`	powtórzenie, z separatorami	repsep(„b", „,") parsuje ciąg składający się z powtórzeń ciągu „b" oddzielonych znakiem „,", oraz zwraca List(„b", „b", „b", ...)]
`rep()`	powtórzenie	rep(„b") parsuje ciąg składający się z powtórzeń ciągu „b", oraz zwraca List(„b", „b", „b", ...)]

Tablica B.1: Dostępne podstawowe kombinatory parserów

B.6.1. Przykłady parserów

Przedstawione zostaną dwa proste przykłady parserów, mieszczące się na 1 stronie A4, oraz realizujące przydatne operacje, takie jak parsowanie formatu JSON, oraz obliczenie prostego wyrażenia matematycznego. Przykłady te mogą posłyżuć jako pomoc w unaocznieniu elegancji oraz potęgi Parser Combinators.

Jako pierwszy przykład parsera chciałbym w tym miejscu zacytować implementację parsera JSON [4] umieszczonego w książce *Programming in Scala* autorstwa Martina Oderskiego [Ode07]:

```
import scala.util.parsing.combinator._

class JSON extends JavaTokenParsers {

  def obj: Parser[Map[String, Any]] =
    "{"~> repsep(member, ",") <~"}" ^^ (Map() ++ _)

  def arr: Parser[List[Any]] =
    "["~> repsep(value, ",") <~"]"

  def member: Parser[(String, Any)] =
    stringLiteral~":"~value ^^
      { case name~":"~value => (name, value) }

  def value: Parser[Any] = (
    obj
  | arr
  | stringLiteral
  | floatingPointNumber ^^ (_.toDouble)
  | "null" ^^ (x => null)
  | "true" ^^ (x => true)
  | "false" ^^ (x => false)
  )
}
```

Kod nie dość że jest elegancki, to okazuje się, gramatyka wykorzystana w tym przykładzie (a zatem i wygenerowany parser) jest klasy *LL(1)*. Oznacza to, że kosztowny mechanizm nawrotów nigdy nie zostanie wykorzystany podczas parsowania JSONa przy pomocy powyższego parsera. Jest to ciekawy argument przeciwko zarzutowi stawianemu kombinatorom parserów, iż generują bardzo powolne (nawracające się) parsery – jest to oczywiście nie prawda, ponieważ to od zastosowanej gramatyki zależy jakiej klasy otrzymamy w efekcie parser.

Drugim przykładem parsera jest zaimplementowany przeze mnie celem pokazania kombinatorów parserów prosty parser, który w efekcie działania oblicza b. proste wyrażenia matematyczne. Parser ten można raczej traktować jako ciekawostkę niż pełną implementację Idei, jednak obrazuje on bardzo dobrze jak wykorzystywane są operacje `^^` celem wpasowania parsowanego drzewa do własnego modelu – analogicznie jak miało to miejsce podczas implementowania parsera ProtoDoc.

[4] JSON - Java Script Object Notation

Listing B.8: Przykładowy parser obliczający podczas parsowania proste zadania arytmetyczne

```
object MathParser extends RegexParsers with ImplicitConversions {
  def number: Parser[Long] = "[0-9]+".r ^^ { _.toLong }
  def num: Parser[Long] = lp ~> number <~ rp

  def add: Parser[AOperation] = ("+" | "-" ) ^^ { AOperation(_) }
  def mult: Parser[MOperation] = ("*" | "/" ) ^^ { MOperation(_) }

  def op: Parser[Long] = lp ~> num ~! (add | mult) ~! num <~ rp ^^ {
    case num1 ~ op ~ num2 => op.perform(num1, num2)
  }

  def lp: Parser[Option[String]] = opt("(")
  def rp: Parser[Option[String]] = opt(")")
  def stop: Parser[Option[String]] = opt(";")

  def parse(string: String): Any = parseAll(op, string) match {
    case Success(res, _) => res
    case e => throw new RuntimeException(e.toString)
  }
}

object MathParserTest extends App {
  override def main(args: Array[String]) {
    import MathParser._
    println(parse("10+15")) // 25
    println(parse("10*2")) // 20
  }
}
```

Kod ponownie jest zwięzły, czytelny oraz scentralizowany w jednym miejscu. W ramach przypomnienia jak podobny parser wyglądałby w GNU Bison, można spojrzeć tutaj: http://www.usna.edu/Users/cs/lmcdowel/courses/si413/F10/labs/L04/calc1/ex1.html gdzie umieszczony został przykład podobnego parsera. Wadami rozwiązania problemy przy pomocy Bison oraz Flex jest korzystanie z wielu narzędzi, wielu plików, oraz wielokrokowy proces budowania projektu – co w kontraście do 1 pliku, 1 języka, oraz braku dodatkowych kroków, faktycznie daje Scali pewną przewagę.

Bibliografia

[Ant11] Antlr - strona domowa projektu: http://www.antlr.org/, 2011.

[Apa12] Apache 2.0 license
http://www.apache.org/licenses/LICENSE-2.0, 2012.

[Bis11] Gnu bison - strona domowa projektu:
http://www.gnu.org/software/bison, 2011.

[BSD12] New bsd license, znana rowniez jako 2-clause bsd license
http://www.opensource.org/licenses/bsd-license.php, 2012.

[GH96] Erik Meijer Graham Hutton. *Monadic Parser Combinators*. 1996.

[Goo11] Google. Dokumentacja google protocol buffers
http://code.google.com/intl/pl-PL/apis/protocolbuffers, 2011.

[Gre08] Jerzy Grebosz. *Symfonia C ++ Standard*. 2008.

[Jav12] Strona domowa i dokumentacja narzedzia oracle javadoc
http://bit.ly/javadochome, 2012.

[Mav12] Strona domowa oraz dokumentacja narzedzia apache maven
http://maven.apache.org, 2012.

[MO] Frank Sommers Martin Odersky, Bill Venners. Wywiad z na temat korzeni jezyka scala
www.artima.com/.../origins_of_scala.

[Mus12] Dokumentacja projektu mustache
http://mustache.github.com/mustache.5.html, 2012.

[Ode07] Martin Odersky. *Programming in Scala*. 2007.

[Pul08] Sam Pullara. Using jax-rs with protocol buffers for high-performance rest apis
http://www.javarants.com/2008/12/27, 2008.

[Sca12a] Scalate - strona domowa projektu: http://scalate.fusesource.org/, 2012.

[Sca12b] Strona domowa oraz dokumentacja biblioteki scalatest
http://www.scalatest.org/, 2012.

[Smi11] Eishay Smith. Jvm serializers benchmark
https://github.com/eishay/jvm-serializers, 2011.

[Thr12] Strona domowa oraz dokumentacja projektu apache thrift
http://thrift.apache.org, 2012.