Plan van Aanpak Course 8a PG6

Richard Jansen, Rick Medemblik, Stan Wehkamp, Alex Staritsky

Mei 2017

Contents

1 Introductie/Projectopdracht

We hebben een opdracht van Ruud Heshof van het HAN BioCentre gekregen om een webapplicatie te ontwikkelen voor zijn onderzoek. Hij doet onderzoek naar Lipoxygenases (LOX's) naar de mogelijke (industriële) applicaties. Hij heeft een paper geschreven over een aantal bekende applicaties van LOX's zoals het bleken van verschillende producten met LOX's.

Ruud is geïnteresseerd in de potentiële verschillende (industriële) applicaties die nog niet bekend zijn van lipoxygenases. Hij wilt dit graag weten omdat hij al veel onderzoek naar lipoxygenases gedaan heeft en denkt dat ze veel potentiëel hebben. Om ze te vinden schakelt hij de hulp van bio-informatici in. Het doel is om met gebruik van text mining van vele wetenschappelijke artikelen over lipoxygenases nieuwe applicaties te vinden. Dit gebeurt door het extraheren van "keywords" (belangerijke (zoek)termen in een tekst) uit artikelen en deze vervolgens te koppelen aan andere keywords van andere artikelen. Door de koppeling van de artikelen kunnen er mogelijk nieuwe verbanden gevonden worden die door regulier onderzoek niet gevonden worden.

Wij hebben de opdracht van Ruud gekregen om een webapplicatie te ontwikkelen voor het HAN BioCentre om die text mining uit te voeren. De gebruiker moet via een web formulier een zoekopdracht opgeven. Vervolgens wordt relevante informatie van databases als NCBI opgezocht, verwerkt en gekoppeld aan elkaar en op een visuele manier weergeven aan de gebruiker. Dit wordt gebruikt door de onderzoekers van het HAN BioCentre om de vraagstelling over nieuwe applicaties te beantwoorden.

1.1 Onderzoeksvragen

Om een goede applicatie te ontwikkelen en een antwoord te geven op de uiteindelijke vraag hebben we een aantal bio-informatica vraagstellingen opgesteld:

1. Welke database(s) gebruiken we om de data en artikelen op te zoeken?
2. Hoe koppelen we verschillende artikelen aan elkaar?
3. Op welke manier kunnen we overzichtelijk relaties tussen artikelen weergeven?
4. Hoe kunnen we alle functionaliteit efficiënt verwerken in een makkelijk te gebruiken webapplicatie?
5. Op welke manier slaan we (tijdelijk) informatie op over de zoekopdrachten en de resultaten op de webserver?

2 Materiaal en methode

Om het project te voltooien zijn een aantal middelen benodigd om bepaalde stappen succesvol te voltooien.

2.1 Materiaal

Python 2.7[6] zal worden gebruikt als primaire scripting software vanwege gebruiksvriendelijkheid, volledige documentatie, brede implementatie onder bioinformatici en adequate functionaliteit. De functionaliteit van Python zal worden uitgebreid door middel van de BioPython en Flask Packages. BioPython heeft beschikking over entrez, een module om makkelijk PubMed artikelen te minen, en Flask wordt gebruikt als webserver die de applicatie web-based beschikbaar zal maken.

HTML4 en HTML5[2] zullen worden gebruikt om een grafische omgeving ter beschikking te stellen voor de klant. Dit helpt de klant om de resultaten overzichtelijk te interpreteten. HTML is de wereldwijde standaard en de functionaliteit is voldoende voor onze doeleinden.

Ubuntu server 14.04[4] De applicatie zal op meerdere platforms draaien maar de Proof of Concept zal worden gehost op een Ubuntu server omdat Ubuntu servers de standaard zijn voor webhosting en andere operating systems vaak compatibel zijn met software die geschikt is voor Ubuntu.

MySQL 5.5[7] zal worden gebruikt om data overzichtelijk op te slaan in een database. Dit zal waarschijnlijk een optionele feature zijn.

GIT en Gitlab[5] worden gebruikt voor distributie en gemakkelijke installatie van de applicatie. Gitlab biedt ook mogelijkheden om Wiki-paginas te maken voor de documentatie.

PubMed[3] zal worden gebruikt als bron van de data gezien het feit dat PubMed de wetenschappelijke standaard is. Er kan worden gekeken naar alternatieven mits de klant dat wenst.

2.2 Methode

2.2.1 Text-mining

Tekstmining verwijst naar het proces om waardevolle informatie te halen uit grote hoeveelheden tekstmateriaal. Met deze techniek wordt geprobeerd om patronen te vinden in de grote bulk aan data. Uit onderzoek bleek dat NCBI met de database ENTREZ [1] werkt. Om artikelen uit NCBI te halen door middel van tekstmining is besloten om queries te gebruiken met de ENTREZ

programming utility om zo de resultaten te bemachtigen. Reden waarom voor deze methode gekozen is, is omdat dit de meest betrouwbare resultaten levert.

2.2.1.1 Data retrieval

De data wordt uit verschillende databases via ENTREZ van NCBI gehaald. We zijn voornamelijk geïnteresseerd in PubMed artikelen en meta-informatie van eiwitsequenties met referenties naar artikelen. De data wordt in verschillende formaten geretourneerd afhankelijk van het type data.

2.2.1.2 Text-preprocessing

In de stap preprocessing worden de binnenkomende teksten uit de stap "data retrieval" verwerkt tot teksten die beter leesbaar zijn voor computers. De teksten die binnen komen zijn door de verschillende formatteringen en de gebruikte spreektaal in de artikelen niet altijd even duidelijk voor een computer. Om dit te verwerken voor de computer wordt gebruik gemaakt van verschillende technieken zoals Natural Language Processing (NLP), Part-of-Speech (POS) en Porter's Stemming.

2.2.1.3 Data analyse

Uit de pre-processed tekst kunnen we keywords uithalen. De keywords zijn volgens een algoritme de belangrijkste woorden uit de tekst. Dit proces wordt voor veel artikelen gedaan. Ook worden de belangrijkste keywords gebruikt om andere artikelen op te zoeken die met die keywords te maken hebben. Uiteindelijk ontstaat er door een algoritme een netwerk van keywords en artikelen die aan elkaar gekoppeld zijn om nieuwe verbanden te vinden.

2.2.1.4 Visualisatie

Het netwerk (waarschijnlijk in de vorm van een grafennetwerk) dat ontstaan is door de data analyse moet gevisualiseerd worden. Als dit niet gebeurt, wordt het moeilijk of niet te doen voor de gebruiker om de data te interpreteren. De visualisatie vind plaats in een webbrowser. De visualisatie moet dus gemaakt zijn in HTML5 of JavaScript om een interactieve visualisatie te krijgen. Een alternatieve mogeljkheid zou zijn dat er een figuur via Python op de server wordt gegenereerd die vervolgens naar de webbrowser van de gebruiker gestuurd wordt.

2.2.1.5 Evaluatie

Als laatste stap is de evaluatie, feitelijk de interpretatie van de data. Dit is ook meteen de belangrijkste stap voor ons omdat hier als alles goed gaat de medische/biologische betekenis van de gezochte data duidelijk weergegeven aan de eindgebruiker geserveerd wordt.

2.2.2 Visualisatie

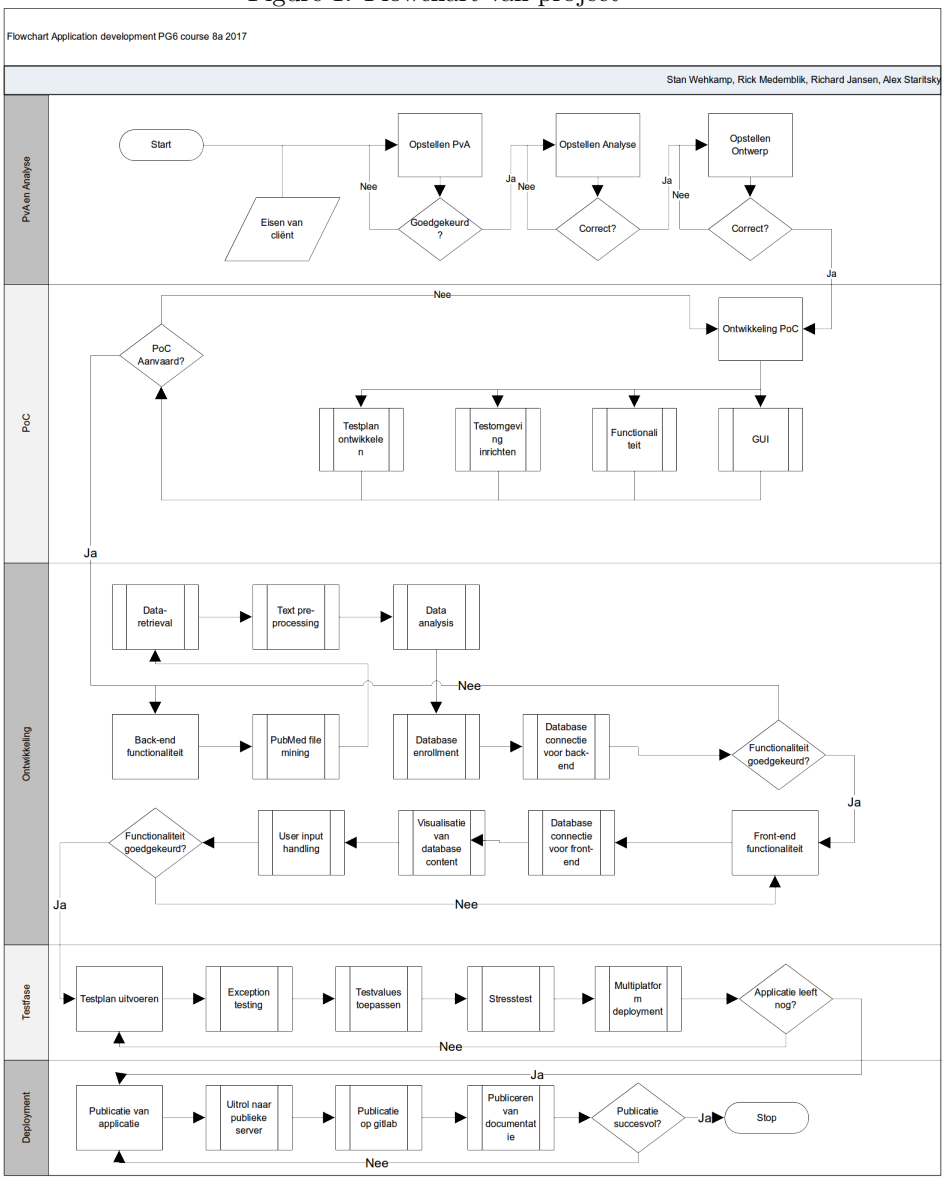
Er is gekozen om de resultaten van de tekstmining te visualiseren via een tabel die wordt weergegeven op een webpagina. Dit houdt in dat het bereikbaar is via het internet en bereikbaar is voor iedereen.

2.2.3 Resultaten

De resultaten afkomstig van het tekstmining worden opgeslagen in een database. De database wordt ontworpen door middel van SQL. Na het opslaan van de gegevens is de volgende stap om de gegevens te kunnen visualiseren door middel van een webpagina. Door middel van een script, de database en een server worden de gegevens geüpload en weergegeven op een webpagina.

3 Flowchart

Figure 1: Flowchart van project



4 Planning en Taakverdeling

4.1 Week 1

Table 1: Tijdsplanning week 1, plan van aanpak

Taak	Wie	Tijdsduur
Casus bijwonen	Iedereen	1 uur
Artikel van Ruud lezen	Iedereen	1 uur
Planning	Alex	1 uur
Taakverdeling	Stan	1 uur
Flowchart	Richard	1 uur

4.2 Week 2

Table 2: Tijdsplanning week 2, analyse

Taak	Wie	Tijdsduur
Inleiding	Rick	1 uur
(Niet) Functionele eisen	Stan	1 uur
Use cases	Alex, Stan	2 uur
Use case diagram	Alex, Stan	1 uur
Test plan	Richard	1 uur
Architectuur	Richard	2 uur
Bronnen	Rick	1 uur

4.3 Week 3

Table 3: Tijdverdeling week 3, ontwerp

Taak	Wie	Tijdsduur
Inleiding	Stan	1 uur
Systeemarchitectuur	Richard	1 uur
Softwarearchitectuur	Richard	2 uur
UML class diagram	Alex	2 uur
ERD diagram	Richard	1 uur
Technische gegevensstructuur	Rick	2 uur
Bronnen	Richard	1 uur

4.4 Week 4 t/m 6

Table 4: Tijdsverdeling week 4 t/m 6, programmeren

Taak	Wie	Tijdsduur
Dataflow	Allemaal	5 uur
Pagina ontwerp HTML	Stan, Rick en Alex	10 uur
Pagina ontwerp CSS	Rick	5 uur
Visualisatie in webpagina	Stan	6 uur
Front-end scripting	Alex	10 - 20 uur
Back-end scripting	Richard	20 - 30 uur
Database opzetten	Allemaal	5 - 10 uur

4.5 Week 7

Table 5: Tijdsverdeling week 7, afronden en presentatie

Taak	Wie	Tijdsduur
Afronden applicatie voor presentatie	Allemaal	5 uur
Presentatie maken	Allemaal	2 uur
Presenteren	Allemaal	15 minuten

References

[1] National Center for Biotechnology Information (US). Entrez programming utilities help, 2010.

[2] Dave Raggett, Arnaud Le Hors, Ian Jacobs, et al. Html 4.01 specification. *W3C recommendation*, 24, 1999.

[3] Richard J Roberts. Pubmed central: The genbank of the published literature, 2001.

[4] Mark G Sobell. *A practical guide to Ubuntu Linux*. Pearson Education, 2014.

[5] Linus Torvalds and Junio Hamano. Git: Fast version control system. *URL http://git-scm. com*, 2010.

[6] Guido Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, volume 41, page 36, 2007.

[7] Michael Widenius and David Axmark. *MySQL reference manual: documentation from the source*. " O'Reilly Media, Inc.", 2002.