Universiteit van Amsterdam

Projectplan

Het maken van een persoonlijke zoekmachine

4 november 2022

Student:
Jorit Prins
12862789

Supervisor:
Maarten Marx

Cursus:
Afstudeerproject

1 context

Sinds 1990 zat nog geen procent van de wereldpopulatie op het internet. In het jaar 2000 haalde al de helft van de Amerikaanse populatie informatie van het internet en in het jaar 2016 was dit al 80 procent. In hetzelfde jaar zit 90 procent van de Nederlandse populatie op het internet (Roser, Ritchie en Ortiz-Ospina 2015). Met deze groei van internetgebruikers groeit ook de informatie die beschikbaar is op het internet. Deze groei van beschikbare informatie is erg belangrijk en is coherent aan bijvoorbeeld economische groei (Vu 2019). Het is hierin belangrijk dat de gebruikers relevante informatie krijgen uit deze alsmaar groeiende berg informatie. Zoekmachines zijn hierin cruciaal. Of de zoekmachine 'goed' is, is afhankelijk van verschillende factoren. Zo is het belangrijk om op de *search engine result pages* (SERP) een *surrogate* (relevante informatie over de gevonden resultaten) te laten zien, zodat de gebruiker beter kan bepalen welk resultaat voldoet aan zijn wensen. Deze factoren maken de creatie van een zoekmachine een langdradig en gecompliceerd proces.

Net als publieke informatie groeien ook onze persoonlijke documenten. Ze hebben verschillende formaten, zijn van verschillende bestandstypes (pdf, mail, word, etc.) en bevatten verschillende informatie. Deze persoonlijke database groeit enorm in grootte. Ook bijvoorbeeld Journalisten hebben vaak een enorme database aan documenten. Het snel en goed kunnen zoeken in deze databases is cruciaal voor hun werk. Ook de vertrouwelijkheid van de documenten is belangrijk om te waarborgen, waardoor ze geen commerciele zoekmachine kunnen gebruiken. Deze en andere situaties zorgen voor de behoefte van een *open source* privé zoekmachine.

Omdat in derde wereldlanden technologie lang niet zo goed beschikbaar is, terwijl technologie hier net zo belangrijk is voor economische groei (Shahidullah 2019), en persoonlijke computers nooit zullen evenaren aan de commerciële rekenkracht van grote bedrijven moet de privé gemaakt zijn voor computers met gelimiteerde RAM en rekenkracht.

2 De onderzoeksvraag

Het doel van dit project is om een *open source* privé zoekmachine te maken die werkt met gelimiteerde rekenkracht en RAM op een grote dataset. De zoekmachine moet eenvoudig zijn (met installatie en gebruik). De onderzoeksvraag is als volgt:

Is het mogelijk om een open source zoekmachine te maken die

- gewenste resultaten geeft,
- eenvoudig te installeren is,

- eenvoudig te gebruiken is (vergelijkbaar met publieke zoekmachines, als bijvoorbeeld Google),
- schaalbaar is
- terwijl deze gelimiteerde rekenkracht heeft (als bijvoorbeeld een verouderde persoonlijke desktop)

Het product zal getest worden op een oude laptop door verschillende mensen een opdracht te geven om een specifiek document of informatie in een specifiek document op te zoeken in een database met de gemaakte zoekmachine.

3 Methoden

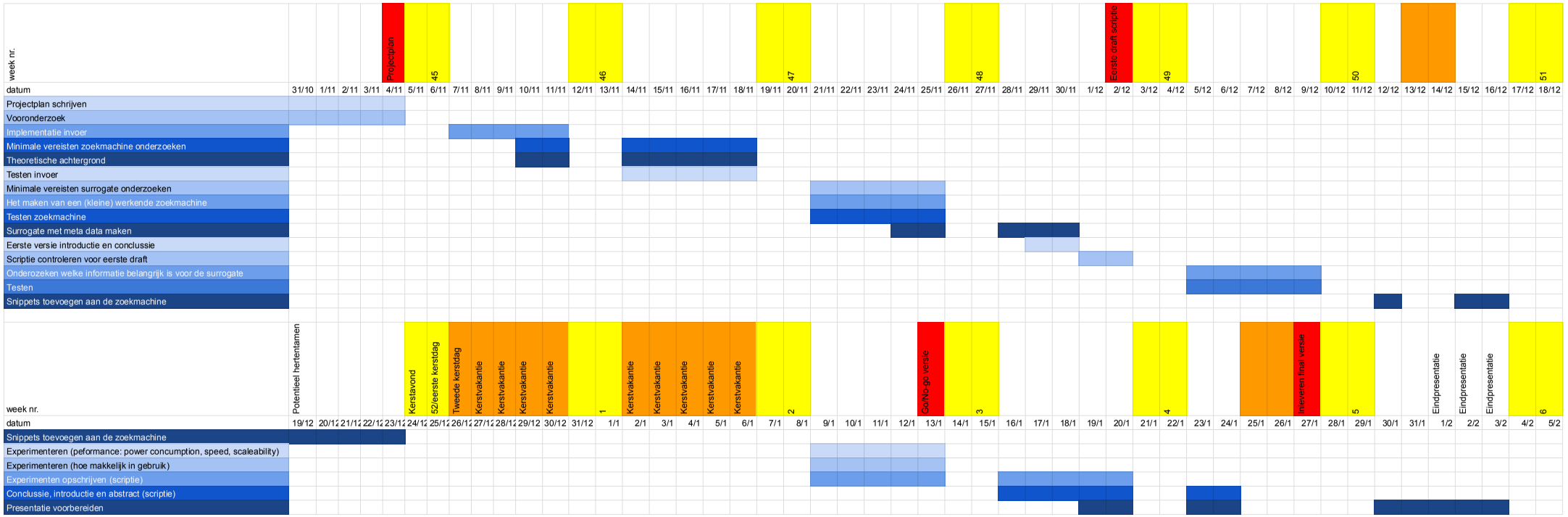
Aan het begin zal er literatuuronderzoek gedaan worden om te kijken wat de minimale eisen zijn om een zoekmachine gewenste resultaten te geven om de rekenkracht te beperken. Ook is het belangrijk om prioriteiten te stellen om wat wel, en wat niet te implementeren (denk aan tekst highlighting in de *surrogate*. Daarnaast moeten we onderzoeken hoeveel informatie de gebruiker wenst te krijgen om een goed oordeel te vellen of de informatie relevant is of niet.

Wanneer we dit weten kunnen we beginnen met de implementatie. De zoekmachine moet als invoer verschillende documenten aankunnen (mail, pdf, word, etc.) en deze op kunnen slaan in een database. Vervolgens moet de zoekmachine in deze database zoeken aan de hand van termen. De resultaten van deze *query* moet genoeg informatie bevatten voor de gebruiker om te bepalen of de zoekopdracht relevante informatie geeft.

Als de zoekmachine af is en getest op bugs kunnen we hem testen. We zullen verschillende groottes aan invoer geven en kijken hoe lang het duurt om informatie en documenten op te halen. Ook zullen we testen of mensen relevante informatie op kunnen halen door ze informatie te laten opzoeken in de database.

4 De planning

(zie volgende pagina)


week nr.
Projectplan
45
46
47
48
Eerste draft scriptie
49
50
51
datum
31/10 1/11 2/11 3/11 4/11 5/11 6/11 7/11 8/11 9/11 10/11 11/11 12/11 13/11 14/11 15/11 16/11 17/11 18/11 19/11 20/11 21/11 22/11 23/11 24/11 25/11 26/11 27/11 28/11 29/11 30/11 1/12 2/12 3/12 4/12 5/12 6/12 7/12 8/12 9/12 10/12 11/12 12/12 13/12 14/12 15/12 16/12 17/12 18/12
Projectplan schrijven
Vooronderzoek
Implementatie invoer
Minimale vereisten zoekmachine onderzoeken
Theoretische achtergrond
Testen invoer
Minimale vereisten surrogate onderzoeken
Het maken van een (kleine) werkende zoekmachine
Testen zoekmachine
Surrogate met meta data maken
Eerste versie introductie en conclussie
Scriptie controleren voor eerste draft
Onderozeken welke informatie belangrijk is voor de surrogate
Testen
Snippets toevoegen aan de zoekmachine
Potentieel hertentamen
Kerstavond
52/eerste kerstdag
Tweede kerstdag
Kerstvakantie
1
2
Go/No-go versie
3
4
Inleveren final versie
5
Eindpresentatie
Eindpresentatie
Eindpresentatie
6
week nr.
datum
19/12 20/12 21/12 22/12 23/12 24/12 25/12 26/12 27/12 28/12 29/12 30/12 31/12 1/1 2/1 3/1 4/1 5/1 6/1 7/1 8/1 9/1 10/1 11/1 12/1 13/1 14/1 15/1 16/1 17/1 18/1 19/1 20/1 21/1 22/1 23/1 24/1 25/1 26/1 27/1 28/1 29/1 30/1 31/1 1/2 2/2 3/2 4/2 5/2
Snippets toevoegen aan de zoekmachine
Experimenteren (peformance: power concumption, speed, scaleability)
Experimenteren (hoe makkelijk in gebruik)
Experimenten opschrijven (scriptie)
Conclussie, introductie en abstract (scriptie)
Presentatie voorbereiden

Referenties

Roser, Max, Hannah Ritchie en Esteban Ortiz-Ospina (2015). „Internet". In: *Our World in Data*. https://ourworldindata.org/internet.

Shahidullah, Shahid M. (feb 2019). *Capacity-Building in Science and Technology in the Third World*. Routledge. DOI: 10.4324/9780429044618. URL: https://doi.org/10.4324/9780429044618.

Vu, Khuong M. (jun 2019). „The internet-growth link: An examination of studies with conflicting results and new evidence on the network effect". In: *Telecommunications Policy* 43.5, p. 474–483. DOI: 10.1016/j.telpol.2019.04.002.