Rationale Project Taal

Hogeschool van Amsterdam - Master Applied Artificial Intelligence

Team 3

Auteurs:	*Studentnummers:*
Kristo Wind	500790749
Oscar Oosterling	500775970
Sander Steur	500809756

Amsterdam
Januari 2023

1 Voorwoord

Voor u ligt het rationale betreffende team 3: SKORT (Sander Kristo Oscar Reading Tool). Deze rationale is een verplicht en tevens afsluitend onderdeel van het project Natural Language Processing uit blok 2 van de master Applied Artificial Intelligence te Amsterdam.

In deze rationale zal het doorlopen proces, de gemaakte keuzes en de onderbouwing voor deze gemaakte keuzes worden behandeld. Het rationale wordt afgesloten met een conclusie, waarin wordt teruggekeken op de behaalde resultaten en mogelijke verbeterpunten voor in de toekomst.

Een dankwoord zouden wij ook nog willen uitspreken naar een aantal personen. Ten eerste naar onze coach, de heer P. Wiggers, voor zijn onafgebroken advies op maat en hulp tijdens het gehele project. Daarnaast willen wij ook de andere coaches/docenten bedanken voor hun hulp en diensten. Tenslotte nog een dankwoord naar de overige studenten van de master MAAI, voor het dienen als hulpvaardige testpersonen en het behouden van de gezelligheid binnen de klas.

'SKORT maar krachtig'

Inhoudsopgave

2 Inleiding

Als HBO-studenten academische artikelen willen of zelfs moeten lezen, kan dat voor veel problemen zorgen. Zo staat er vaak relatief veel jargon en complexe taal in de papers, wat het lastig maakt om de tekst volledig te begrijpen en misschien zelfs tot demotivatie leidt. Het doel van dit project betreft het ontwerpen van een AI oplossing die het lezen, bestuderen of begrijpen van academische teksten makkelijker maakt voor HBO-studenten. Veel studenten lezen voor hun opleiding veel teksten waarbij de betekenis of de relevantie van sommige woorden nog onbekend is.

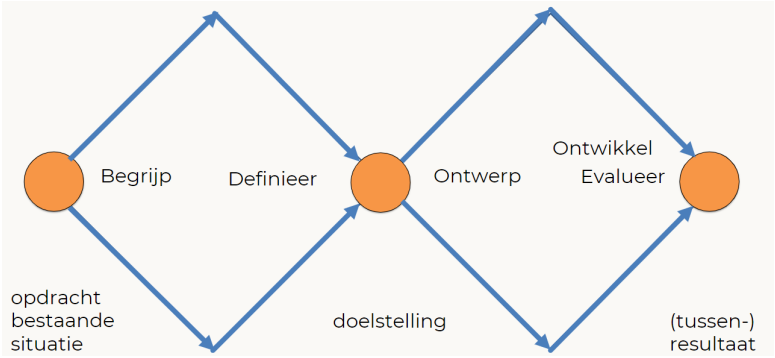
In Hoofdstuk 3 wordt het theoretische onderzoek van dit project blootgelegd. Vervolgens wordt in Hoofdstuk 4 het concept van de te ontwikkelen tool beschreven. Hoofdstuk 5 beschrijft de verschillende iteraties die doorlopen zijn om het concept te realiseren. De technische implementatie van de onderdelen van de te ontwikkelen tool worden in Hoofdstuk 6 aangehaald. Ten slotte zullen de discussie en conclusie in respectievelijk Hoofdstuk 7 en 8 beschreven worden.

3 Theoretisch onderzoek

In dit hoofdstuk wordt het gehele theoretische onderzoeksproces van team 3 doorlopen betreffende het project Natuurlijke Taalverwerking.

3.1 Procesflow

Gedurende dit project is een iteratief ontwerpproces doorlopen om tot een gewenst eindresultaat te kunnen komen. Dit proces wordt weergegeven in Figuur 1 en wordt ook wel het *double diamond* design proces [1] genoemd, welke werd geïntroduceerd door Design Council in 2004.



Figuur 1: Double Diamond design proces

De opdracht voor dit project luidt: ‘Ontwerp een AI oplossing om het lezen en begrijpen van teksten te vereenvoudigen voor studenten’, met de oorspronkelijke eisen dat de oorspronkelijke betekenis van de tekst behouden moet blijven en de student niet dommer mag worden door de oplossing. Het double diamond design proces is twee maal doorlopen dit project. Na het doorlopen van de eerste iteratie, resulteerde dit in een paper prototype. Na het doorlopen van de tweede iteratie, resulteerde dit in een uitgewerkt digitaal prototype in de vorm van een applicatie.

3.2 Brainstorm

Er zijn tijdens de eerste brainstormsessies verschillende oplossingsrichtingen en ideeën aan het licht gebracht die behandeld hadden kunnen worden in dit project. Hiervoor zijn eerst diverse

ideeën doordacht, waarvan enkele in een tool verwerkt konden worden, gebaseerd op de haalbaarheid binnen het beschikbare tijdsbestek en het belang van de doelgroep. De doelgroep werd vastgesteld op HBO-studenten die wetenschappelijke artikelen lezen. De mogelijke ideeën zijn hier destijds opgesomd.

3.3 Domeinkennis

Domeinkennis is nodig voor de verdieping betreffende iteratie één uit Figuur 1: 'Begrijp'. Dus dient er begrepen te worden wat er speelt bij de doelgroep, om te bepalen welke ideeën relevant kunnen zijn als onderdeel van de tool.

Markeren en herlezen van stukken tekst biedt minimaal effect bij HBO studenten [2]. Zelfoverhoring, oefenvragen en het over langere tijd verspreiden van leersessies zijn wel effectief. Hoe effectief studenten leren hangt niet enkel af van de manier waarop ze een tekst benaderen: 'Door het bevorderen van levensvaardigheden gaan studenten beter in hun vel zitten en worden zij sociaal en emotioneel vaardiger. Hierdoor zijn zij in staat om zichzelf te motiveren en verbeteren hun leerprestaties' [5]. Studenten hebben vaak de nodige stress, omdat ze over het algemeen veel werken, een druk sociaal leven hebben en bezig zijn met sociale media [7]. Daarnaast vinden studenten het lastig om hulp te vragen, hun gedrag te sturen, zich betrokken te voelen bij onderwijsactiviteiten en (kritisch) te communiceren in de collegezaal [8]. Studenten hebben verder vaak ondersteuning nodig om gemotiveerd te blijven en ze kunnen extra vatbaar zijn voor depressieve klachten. Levensvaardigheden dragen bij aan studiesucces en motivatie om te leren. Zie Hoofdstuk 19 van Hoe Leren Studenten [5] voor meer informatie over het investeren in studiesucces met behulp van levensvaardigheden als zelfbewustzijn, zelfmanagement, sociaal bewustzijn, relatievaardigheden en verantwoordelijke besluitvorming [6].

Om erachter te komen hoe het wel moet, is onderzoek gedaan naar de leerwijze van honoursstudenten. Honoursstudenten in het HBO houden zich tijdens het leren bezig met of de theorieën, interpretaties en conclusies goed onderbouwd zijn met bewijs [9]. Daarnaast stellen deze studenten gerichte doelen voor hunzelf. Honoursstudenten onderscheiden zich van reguliere studenten door een hogere mate van kritisch denken en zelfregulerend leren. Om reguliere studenten slimmer te maken is vanuit hunzelf een grote mate van interesse en motivatie nodig. Door informatie bij lastige woorden te plaatsen die de student kan bestuderen en bekritiseren, wordt de kans op interesse vergroot. Bij interesse komt vaak motivatie kijken, waardoor de docent vooral kan optreden als begeleider, zodat de student meer zelfgereguleerd kan studeren.

3.4 Behoeftes doelgroep

De behoeftes van de doelgroep zijn vastgesteld door middel van kwalitatief- en kwantitatief onderzoek. Er zijn 3 interviews afgenomen met studenten van de Master Applied Artificial Intelligence die binnen de doelgroep van dit project vallen, en er is een enquête gepubliceerd. De enquête is beantwoord door 34 studenten (MBO (3), HBO-bachelor (18), HBO-master (8), WO-bachelor (3) en afgestudeerd (2)) van verschillende opleidingen binnen en buiten de Hogeschool van Amsterdam. Voor de interviews zijn er 4 verschillende vragen gesteld aan 3 verschillende personen. De exacte antwoorden van de interviews zijn in ons logboek onder het kopje Interviews met doelgroep te vinden.

Uit de interviews is naar voren gekomen dat studenten relatief veel tijd kwijt zijn aan het achterhalen wat een woord betekent als dit niet duidelijk te herleiden is uit de tekst. Daarnaast zijn de geïnterviewde studenten geïnteresseerd in een gepersonaliseerde leestijd die aangeeft hoe lang de student doet over de tekst, gegeven zijn/haar leessnelheid. En het meest fundamentele aspect voor ons onderzoek: de bevraagde studenten geven aan meer interesse te hebben in een

bepaald onderwerp als ze goed begrijpen waar de tekst over gaat.

Uit de enquête kwam onder andere voort dat digitale hulpmiddelen als verwante artikelen en boeken aanbevelen op basis van de categorie en onderwerp, markering van de kernzin van een alinea en markeren van belangrijke afkortingen en jargon relatief gewild zijn. De resultaten van de hele enquête zijn hier te vinden.

3.5 Bepalen opdrachten

Duidelijk hebben welke onderdelen in een tool verwerkt zullen worden, is nodig voor het convergeren van iteratie één uit Figuur 1: 'Definieer'. Na een brainstormsessie over de mogelijke opdrachten en het opdoen van domeinkennis is bepaald welke ideeën uitgewerkt zullen worden in dit project. Gezamenlijk met de coach is besloten om de volgende zes ideeën te ontwikkelen en in een tool te verwerken:

- Leestijd persoonlijk bepalen
- Definities vinden voor vaktermen
- Kernwoorden/kernzinnen uit de tekst halen
- Automatische quiz generator
- Moeilijkheidsgraad van de tekst bepalen
- Tekst summarizer

Door bovenstaande ideeën uit te werken worden de teksten die de HBO-studenten dienen te lezen vereenvoudigd, zonder dat er belangrijke informatie verloren gaat. Door automatisch quizzen te genereren kan de student zichzelf overhoren, wat effectief werkt [2]. Ook overzichtelijke informatie als een leestijd of moeilijkheidsgraad van een tekst vinden studenten prettig (zie §3.4). Wetenschappelijke artikelen bevatten over het algemeen veel lastige vaktermen en lastige woorden, waardoor de tekst ingewikkeld en lastig te begrijpen kan zijn voor HBO-studenten. Daarom zullen kernwoorden en kernzinnen uit de tekst gehaald worden en zullen definities van vaktermen weergegeven worden. Zodoende zijn de teksten beter te begrijpen en wordt er meer interesse voor het vak opgewekt bij de student.

3.6 Doelstelling

Met het definiëren van een doelstelling is de eerste iteratie van het ontwerpproces uit Figuur 1 voltooid.

Het doel van de te ontwikkelen tool is om studenten te helpen teksten beter te begrijpen en interesse op te wekken in het eigen vakgebied bij studenten die minder motivatie hebben, waardoor ze beter gaan presteren. Dit wordt gerealiseerd door de teksten te vereenvoudigen met eerdergenoemde ideeën, wat de kennis en dus ook interesse en motivatie vergroot in het vakgebied, volgens de afgenomen enquête (zie §3.4). Met behulp van de tool wordt de stap kleiner om de betekenis en interpretatie van belangrijke en lastige woorden uit te zoeken. Daarnaast is het doel om de concentratie van de student te waarborgen. Als de student elk woord dat hij/zij niet begrijpt moet opzoeken in de browser verliest de student continu de concentratie. Door definities van belangrijke/lastige woorden in de tool weer te geven leest de student sneller, makkelijker en met bewaarde concentratie door de tekst.

3.7 Pakket van eisen

In Tabel 1 worden alle eisen waar SKORT aan moet voldoen opgesomd om de doelstelling uit §3.6 te realiseren.

Nummer	Categorie	Eis	Afkomstig van
1	Technisch	Het prototype van SKORT kan kernwoorden uit een tekst halen en deze markeren.	Doelgroep
2	Technisch	De leestijd voor de paper wordt berekend en is personaliseerbaar.	Doelgroep
3	Technisch	De moeilijkheidsgraad voor de paper wordt berekend en aan de gebruiker getoond.	Doelgroep
4	Technisch	Definities en Synoniemen van termen moeten in de tool op te zoeken of terug te vinden zijn.	Doelgroep
5	Technisch	Gebruiker moet zelf woorden kunnen selecteren om definities over te krijgen.	Doelgroep/Team
6	Technisch	De student mag niet dommer worden van de tool; De tekst mag geen informatie verliezen.	Opdracht
7	Technisch	Tool moet papers kunnen inlezen (en kernwoorden extracten) van elke website (scalability).	Team
8	Technisch	Gevonden kernwoorden moeten dicht, niet meer dan drie fout, bij de kernwoorden liggen die een gemiddeld persoon uit de doelgroep zou kiezen.	Team
9	Technisch	Het SKORT model, voor kernwoordextractie, moet op gebied van performance beter presteren dan het vastgestelde nulmodel.	Team
10	Etisch	Het SKORT model, voor kernwoordextractie, moet dezelfde prestatie kunnen leveren op verschillende talen.	Team
12	Etisch	Gevonden kernwoorden mogen niet een bias hebben richting een bepaalde doelgroep.	Team
13	Etisch	Het SKORT model werkt voor verschillende soorten teksten, zodat het toegankelijk is voor een breed publiek.	Team
14	Etisch	Het SKORT model geeft aan hoe de tool werkt en wat de beperkingen zijn, zodat de gebruiker weet wat hij/zij kan verwachten (transparantie).	Team
15	Juridisch	Er mag geen informatie behouden worden over de ingeladen paper.	Auteursrecht
16	Juridisch	De gebruiker hoeft geen account aan te maken en persoonsdata (IP-adres, naam etc.) wordt niet opgeslagen.	AVG
17	Juridisch	De gebruiker heeft het recht om zijn/haar gepersonaliseerde data (leestijd, moeilijkheidsgraad) te laten verwijderen.	Recht op wissing
18	Organisatorisch	De gebruiker moet feedback kunnen teruggeven/delen aan de ontwikkelaars.	Doelgroep
19	Organisatorisch	Tool moet overzichtelijk zijn ingericht, zodat onboarding niet nodig hoeft te zijn (explainability).	Doelgroep
20	Organisatorisch	De SKORT tool wordt gemonitord en onderhouden door het team en zal niet worden uitbesteed.	Team

Tabel 1: Pakket van eisen

3.7.1 MoSCoW prioritering

Om de doelstellingen onder te kunnen verdelen op basis van hun belang, is een MoSCoW prioritering [3] opgesteld. Deze is aan de hand van de requirements uit Tabel 1 opgesteld en geeft de significantie van de doelstellingen weer voor dit project.

Must

1. Het prototype van SKORT kan kernwoorden uit een tekst halen en deze markeren.
2. De leestijd voor de paper wordt berekend en is personaliseerbaar.
3. De moeilijkheidsgraad voor de paper wordt berekend en aan de gebruiker getoond.
4. Definities van termen moeten in de tool op te zoeken of terug te vinden zijn.

5. De student mag niet dommer worden van de tool; De tekst mag geen informatie verliezen.
6. Het SKORT model, voor kernwoordextractie, moet op gebied van performance beter presteren dan het vastgestelde nulmodel.
7. De SKORT tool wordt gemonitord en onderhouden door het team en zal niet worden uitbesteed.
8. Gebruiker moet zelf woorden kunnen selecteren om definities over te krijgen
9. De lay-out van SKORT moet zo dicht mogelijk bij e-ink (i.e. e-paper of elektronisch papier) [4] in de buurt komen om de leesbaarheid te optimaliseren (readability).

Should

1. SKORT moet papers kunnen inlezen (en kernwoorden extracten) van elke website (scalability).
2. Er mag geen informatie behouden worden over het ingeladen paper.
3. De gebruiker moet feedback kunnen teruggeven/delen aan de ontwikkelaars.
4. SKORT moet overzichtelijk zijn ingericht, zodat onboarding niet nodig hoeft te zijn (explainability).

Could

1. Gevonden kernwoorden moeten dicht, niet meer dan drie fout, bij de kernwoorden liggen die een gemiddeld persoon uit de doelgroep zou kiezen.
2. De gebruiker hoeft geen account aan te maken en persoonsdata (IP-adres, naam etc.) wordt niet opgeslagen.
3. Gevonden kernwoorden mogen niet een bias hebben richting een bepaalde doelgroep.

Won't

1. Dataset voor kernwoordextractie moet bestaan uit teksten geschreven door een even verhouding aan mannen en vrouwen.
2. Het SKORT model, voor kernwoordextractie, moet dezelfde prestatie kunnen leveren op verschillende talen.

3.8 Bias

Het is altijd mogelijk dat er bias aanwezig is in het SKORT model. Bias dient zoveel mogelijk gereduceerd te worden. Om dat te kunnen realiseren, zijn alle soorten bias waar SKORT mee in aanraking kan komen in onderstaande lijst opgesomd.

- **Selectiebias**: dit komt voor als SKORT getraind is op teksten van bijvoorbeeld blanke mannen, mensen uit een bepaalde cultuur of voertaal [10]. Dan kunnen andere teksten mogelijk anders geïnterpreteerd worden en handelt het model niet eerlijk. Daarom is het van belang dat SKORT getraind is op teksten van mensen met verschillende achtergronden.
- **Historische bias**: dit komt voor als SKORT is getraind op teksten die niet de werkelijkheid weerspiegelen, doordat het geen recente teksten zijn [11]. In dat geval kan het moeilijk zijn om teksten geschreven in een ander tijdperk goed te begrijpen. Daarom is het van belang dat SKORT getraind is op teksten van diverse tijdperken.
- **Taalbias**: dit komt voor als SKORT enkel is getraind op teksten met Engels als voertaal. In dit geval is SKORT weinig tot niet bruikbaar in andere talen, afhankelijk van de optie die de gebruiker wilt benutten. Om dit te voorkomen is het van belang dat er alleen Engelstalige teksten gebruikt kunnen worden in de tool.

3.9 Evaluatiemethode

SKORT zal geëvalueerd worden door

- **de betrouwbaarheid te controleren**: het is belangrijk dat SKORT consistent presteert op allerlei soorten teksten, rekening houdend met de soorten bias, en doet wat wij zeggen dat hij moet doen.
- **de nauwkeurigheid te controleren**: dit gaan we controleren door de tool te testen met verschillende soorten teksten en te kijken of de kernwoorden en definities die SKORT genereert juist zijn.
- **het gebruiksgemak te evalueren**: dit gaan we controleren door te kijken of de user interface (UI) duidelijk en logisch is aan de hand van testsessies en of SKORT niet te langzaam werkt.

In §6.7 zullen de uitkomsten van de evaluatie van het prototype besproken worden.

3.10 Confusiematrix

De confusiematrix uit Tabel 2 laat de mogelijke uitkomsten voor de kernwoordextractie uit de applicatie zien. De True Positive en de True Negative zijn de toestanden waar we naar toe streven (de groene cellen). We focussen ons zo veel mogelijk op het voorkomen van False Negatives. Bij deze uitkomst missen we mogelijke kernwoorden die cruciaal kunnen zijn voor de tekst. False Positives worden minder zwaar geschat in dit geval, doordat de daadwerkelijke kernwoorden nog steeds gemarkeerd kunnen worden. Bovengenoemde keuze had geverifieerd moeten worden bij potentiële gebruikers van de tool, maar daar is geen tijd voor geweest.

Een overvloed van kernwoorden zou later nog gefilterd kunnen worden. Kernwoorden zijn in dit project gedefinieerd als de belangrijkste woorden uit de tekst die essentieel zijn voor het begrijpen van de tekst. Met deze kernwoorden zou je iemand die de tekst niet heeft gelezen, zoveel mogelijk informatie over de tekst kunnen verstrekken.

Confusiematrix	**Positieve voorspelling**	**Negatieve voorspelling**
Positieve waarheid	**True Positive**: Woorden die in de tekst zijn gemarkeerd zijn daadwerkelijk belangrijke kernwoorden.	**False Negative**: Een woord dat daadwerkelijk een belangrijk kernwoord is in de tekst wordt niet gemarkeerd.
Negatieve waarheid	**False Positive**: Er wordt een woord gemarkeerd, maar dit is geen kernwoord in de tekst.	**True Negative**: woorden die geen kernwoorden zijn in de tekst worden niet gemarkeerd.

Tabel 2: Confusiematrix SKORT

3.11 Impact op de maatschappij

Hoewel het hier gaat om een simpele reading tool, kan de tool toch impact hebben op de maatschappij. Een positieve impact zou hierbij het doel zijn, waarbij studenten in staat zijn om de volledig functionerende tool te kunnen gebruiken om het lezen/begrijpen van papers te kunnen vereenvoudigen. Daarnaast is het de bedoeling hiermee overige ontwikkelaars aan te sporen om ook taalverwerkingsprojecten op te starten om studenten te kunnen helpen.

De tool kan ook een (theoretische) negatieve impact op de maatschappij hebben. Denk hierbij aan het gebruik van papers die op onreglementaire wijze zijn verkregen en in de tool worden geladen (privacy).

Daarentegen de kans klein op deze negatieve impact klein geacht, gezien de staat van het huidige prototype en de nog lange weg van ontwikkelen voor ons. Er wordt verder ook geen persoonlijke informatie opgeslagen van de gebruikers. Hierom wordt de impact op de maatschappij op laag ingesteld.

3.12 Mate van automatisering

In Tabel 3 worden de verschillende maten van automatisering gekwantificeerd.

Laag	1	De computer assisteert, de gebruiker neemt alle beslissingen en acties,
	2	De computer geeft een lijst aan alternatieve beslissingen/acties, of
	3	Beperkt de selectie tot een paar, of
	4	Stelt een alternatief voor, en
	5	Voert suggestie uit bij toestemming van de gebruiker, of
	6	Geeft de mens een beperkte veto tijd voor automatische uitvoering
	7	Voert automatisch uit, informeert dan noodzakelijkerwijs de gebruiker, en
	8	Informeert de gebruiker als daar naar gevraagd wordt, of
	9	Informeert de gebruiker alleen als de computer het beslist
Hoog	10	De computer beslist alles, handelt autonoom, negeert de gebruiker

Tabel 3: Mate van automatisering. Aangepast overgenomen van [15].

Ons prototype valt onder level 5 uit Tabel 3. Het prototype voert automatisch het AI gedeelte in de achtergrond uit wanneer de tekst wordt ingeladen, maar het is vervolgens aan de gebruiker om te selecteren welke opties te gebruiken of in te zien. Daarbij kunnen gebruikers feedback geven als zij liever iets anders zien.

3.13 Ongewenste situaties

Er kunnen zich verschillende ongewenste situaties voordoen bij het inzetten van de tool. Deze zijn in onderstaande lijst opgesomd.

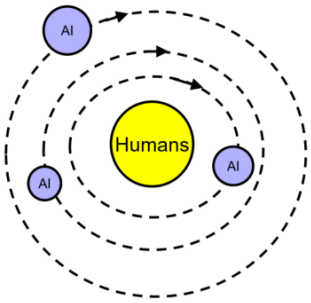
- **Onnauwkeurigheid**: als SKORT onnauwkeurig is kan dit leiden tot verkeerde definities of een verkeerd begrip van de tekst (zie §3.10). Verkeerde informatie kan leiden tot miscommunicatie of verwarring bij de gebruiker.
- **Bias**: als SKORT niet goed is afgestemd op verschillende soorten teksten kan dit leiden tot discriminatie in toegang tot informatie (zie §3.8).
- **Privacy-inbreuk**: als SKORT onbedoeld persoonlijke gegevens opslaat of deelt zonder toestemming van de gebruiker, kan dit leiden tot privacy-inbreuk [12].
- **Verwarring**: als SKORT onduidelijk is of als de uitleg van het model niet goed is, kan dit leiden tot verwarring bij gebruikers.
- **Reputatieschade**: als SKORT fouten maakt of als er problemen optreden met de tool (slechte prestaties), kan dit leiden tot een negatief effect op de reputatie van SKORT [13].
- **Sociale uitsluiting**: als SKORT duur is om te gebruiken of te onderhouden, kan dit leiden tot financiële kosten voor gebruikers of organisaties, wat op den duur kan leiden tot sociale uitsluiting. Dit omdat de rijken dan wel gebruik kunnen maken van SKORT, waar de armen het geld niet (over) hebben voor een abonnement [14].
- **Incompatibiliteit**: als SKORT niet goed samenwerkt met andere systemen of programma's, kan dit leiden tot problemen met het gebruik van SKORT of het uitvoeren van taken.

4 Concept

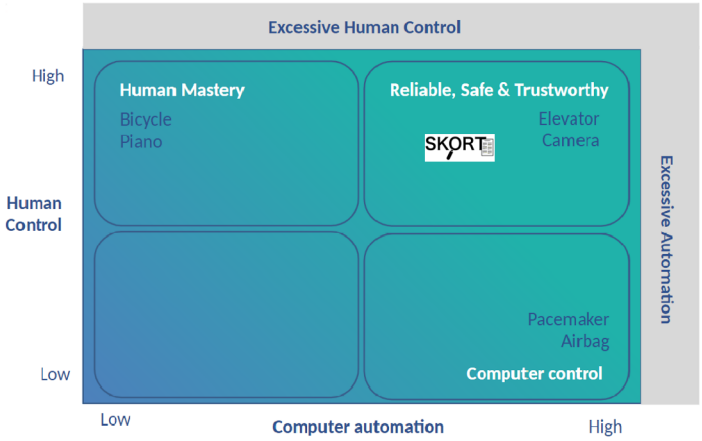
Vanuit het onderzoek is er een concept tool ontworpen. Om tot een concept te komen zijn een (1) level of control, (2) user problem statement, (3) flowchart, (4) role of AI concept en een (5) storyboard gerealiseerd. In onderstaande paragrafen worden alle eerdergenoemde punten aangehaald.

4.1 Level of control

Het doel is om betrouwbaar en veilig te zijn. Dit wordt behaald door een goede balans tussen human-control en computer-automation te vinden [16]. Human-control houdt in dat de gebruiker de controle heeft over het AI-systeem, en computer-automation houdt in dat de computer veel geautomatiseerd doet. Dit is lang niet altijd het geval binnen de AI tools. Ons doel is om de mens het middelpunt te maken van de tool, waar AI om draait (zoals te zien in Figuur 2). SKORT beschikt over zowel human-control als computer-automation (zie Figuur 3), zoals bepaald in §3.12. Zo is snelle actie nodig in de tool door de computer, maar ook de controle van de gebruiker. Omdat de gebruiker zelf kan selecteren welke tools hij/zij wilt dat toegepast worden op de tekst, beschikt SKORT over beide. De gebruiker kan daarnaast ook feedback geven op de tool door bijvoorbeeld aan te geven welk woord de gebruiker belangrijk/lastig vind waar geen verdere informatie bij staat.



Figuur 2: Human centered-AI. Aangepast overgenomen uit [17]



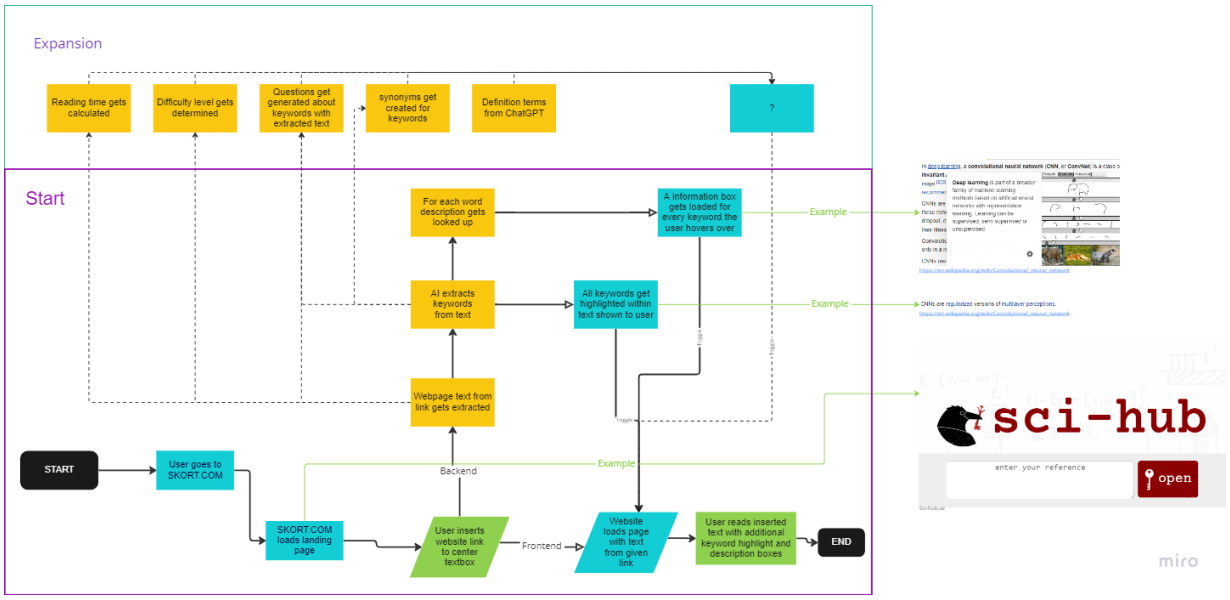
Figuur 3: Twee-dimensionaal framework: Level of control

4.2 User problem statement

Om een beter idee te krijgen van wat het concept oplost zijn er twee user problem statements opgesteld. Het user problem statement is hier te vinden op GitLab. Hier is naar voren gekomen dat het voor een student lastig is om een tekst te begrijpen wanneer bepaalde kernwoorden niet begrepen wordt.

4.3 Flowchart

De flowchart van het project is hieronder weergegeven (zie Figuur 4). Bij de flowchart is het proces voor de gebruiker uitgewerkt, waarbij de stappen van de gebruiker en het systeem worden weergegeven.



Figuur 4: Flowchart

4.4 Role of AI concept

Om inzicht te kunnen krijgen in de taak van AI binnen het concept zijn er vragen beantwoord die dit duidelijk maken. Die zijn hier terug te vinden in het logboek. Hierbij is naar voren gekomen dat de AI efficiëntie voor beter begrip moet zorgen en dat de AI pas waardevol is wanneer dit het begrip van een tekst helpt verbeteren en de juiste kernwoorden markeren.

4.5 Storyboard

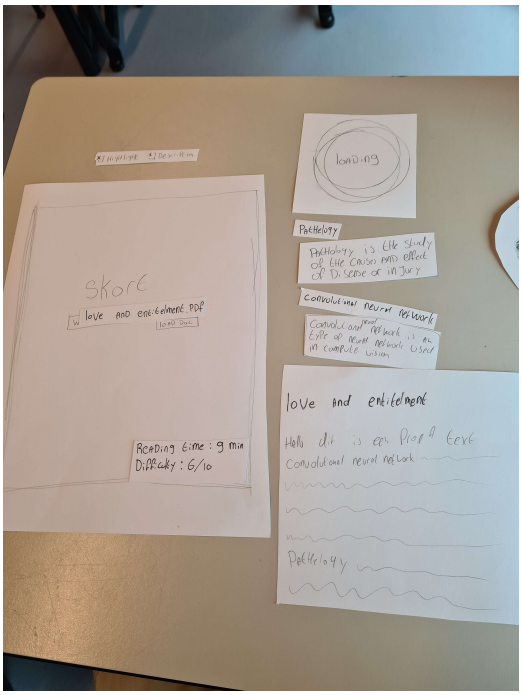
Het storyboard van SKORT wordt op onze GitLab-pagina onder het kopje Storyboard weergegeven. De behoeften van de doelgroep is het vereenvoudigen van het vinden van de essentie van een tekst en het eenvoudig vinden van definities/omschrijvingen van vaktermen en lastige woorden in een paper. SKORT zal daarom eenvoudig definities/omschrijvingen van jargon kunnen geven, daarnaast levert het algemene nuttige informatie over de paper (moeilijkheidsgraad en (persoonlijke) leestijd). Ook geeft SKORT een korte samenvatting van het artikel (maximaal 30 woorden) en kan de ChatGPT API gebruikt worden in de tool voor overige vragen over de tekst.

5 Concept iteraties

Om een concept te realiseren is als eerste stap een paper prototype gemaakt (zie Figuur 5). Deze is getest op verschillende testpersonen. Aan de hand van de feedback van de testpersonen is een digitaal prototype gerealiseerd.

5.1 Paper prototype

Het paper prototype van SKORT is onder het kopje Paper Prototype te vinden op GitLab. Het betreft een schets die laat zien hoe SKORT in zijn werk gaat. Er is een paper prototype gemaakt, omdat er creatieve ideeën ontstaan tijdens het maken van een paper prototype. Over elk detail dient grondig nagedacht te worden. Daarnaast is een paper prototype nuttig, omdat er nog weinig tot geen tijd is gestoken in het technische aspect van de tool. Het zou erg veel tijd kosten om telkens verbeterpunten aan te passen in de techniek. Het eerst vragen van feedback op een paper prototype en vervolgens programmeren, scheelt veel programmeerwerk.



Figuur 5: Paper prototype

5.1.1 Gebruikers testen

Verschillende testpersonen hebben het paper prototype getest. Dit is verlopen op de volgende werkwijze:

1. De testpersoon krijgt uitleg over wat hem gevraagd wordt, maar krijgt geen informatie over hoe de tool werkt.
2. De testpersoon dient zelf uit de voeten te kunnen met de tool door fysiek te drukken op geschetste knoppen.
3. Na drukken op een knop verandert de samenstelling van het paper prototype.
4. Herhaal vanaf stap 2 tot het prototype volledig is doorlopen.
5. Na het voltooien van het prototype is een feedbackformulier ingevuld door de testpersoon voor tips en tops.

De feedback van de testen binnen de klas zijn hier te vinden en de feedback van de testen buiten de klas zijn hier te vinden.

5.2 Digitale prototype

Na het opnemen van feedback aan de hand van het paper prototype, is feedback verwerkt. Enkele doorgevoerde aanpassingen zijn:

- Informatieknoppen zijn toegevoegd bij de leestijd en moeilijkheidsgraad.
- Terug- en homeknoppen zijn toegevoegd.
- Kernwoorden poppen niet allemaal meer op na een klik op de knop.
- De lay-out is overzichtelijker gemaakt. De paper is nu te lezen op de linkerkant van de pagina. Buttons bevinden zich op de rechterkant van de pagina.

- Sci-Hub [18] links zijn toegevoegd als optie voor inladen, zodat de gebruiker niet eerst nog de paper hoeft te downloaden als PDF-file.

Het digitale prototype na verwerkte feedback is op de GitLab-pagina te vinden in de vorm van een video (.mp4), waar alle stappen van het prototype doorlopen worden.

6 Technische implementatie

SKORT beschikt over diverse specificaties. Zo kan SKORT kernwoorden extraheren, vragen genereren, de moeilijkheidsgraad bepalen, een samenvatting maken en definities genereren van jargon. Alle eerdergenoemde specificaties worden in de volgende paragrafen besproken.

6.1 Kernwoordextractie

Kernwoord extractie modellen vinden kernwoorden in documenten. Er zijn diverse kernwoord extractors of kernwoord extraheermethoden die gebruikt zullen worden in dit project. Naast Spacy [19], RAKE [20], en KeyBERT [21] zijn er 10 PKE (Pyramid Keyword Extraction) modellen getest. Waaronder het supervised model PKE Kea ([22] getraind op de SemEval-2010 dataset [23]), maar voornamelijk unsupervised modellen die onderverdeeld kunnen worden in statistische modellen (TfIdf, KPMiner [24] en YAKE [25]) en graph-based modellen (TextRank, SingleRank [26], TopicRank [27], TopicalPageRank [28], PositionRank [29] en MultipartiteRank [30]). Deze alinea is gebaseerd op het Medium artikel 'Keyword Extraction Methods - The Overview' van Godec [31].

Statistische kernwoordextractiemethoden zijn het minst complex. Deze statistische modellen gebruiken statistiek om kernwoorden te berekenen en een score toe te kennen. Er zit verder niks diepers achter. Graph-based modellen zijn wat complexer en genereren grafen van gerelateerde termen van het document. Deze modellen maken een graaf van de tekst, met de woorden als knopen en verbindingen tussen woorden die samen voorkomen. Met een bepaalde metriek, welke beschreven wordt in het Medium artikel 'Exploring Different Keyword Extractors - Graph Based Approaches' van Shrivastava [32], kan bepaald worden welke woorden als kernwoorden beschouwd kunnen worden. De supervised modellen zijn nog complexer, omdat dit modellen zijn die getraind moeten worden.

De diverse extractors zijn getest op drie test-tekstjes met bijbehorende kernwoorden, welke onder het kopje Kernwoord Extraction Comparison te vinden zijn op de GitLab-pagina van team 3. Een extractor wordt al beste gekroond, als de cosine similarity score [33] van het desbetreffende model het hoogst is. De cosine similarity score geeft weer wat de afstand is tussen twee woorden (in dit geval echte kernwoorden en voorspelde kernwoorden). De resultaten van de testen zijn weergegeven in de tabel die tevens op dezelfde GitLab-pagina te vinden is.

Uit de testresultaten blijkt dat de performance van PKE SingleRank gemiddeld het beste is op de gebruikte testdata. Bovendien bleek de PKE methodiek in zijn algemeen sneller te werken dan modellen als KeyBERT of YAKE, gezien de lagere resource demands. In termen van explainability en model complexity heeft PKE Singlerank een specifieke methode om kernwoorden te extracten die gemakkelijker te begrijpen is dan het gebruik van een voorgetraind model als KeyBERT of supervised modellen als PKE Kea. PKE Singlerank werkt minder goed als het grote hoeveelheden tekst moet verwerken. Academische papers zijn relatief klein, dus PKE Singlerank is scalable genoeg voor ons onderzoek. Hierom is uiteindelijk besloten om PKE SingleRank toe te passen in het prototype als definitieve kernwoordextractiemethode.

6.2 Gepersonaliseerde leestijd

Een wat kleiner en eenvoudiger onderdeel van SKORT is het bepalen van de leestijd van een paper en mocht de gebruiker dit wensen, deze naar zijn/haar leessnelheid te kunnen personaliseren. De leestijd, in seconden, voor een paper wordt in eerste instantie berekend aan de hand van de volgende formule:

$$\text{leestijd} = \frac{\text{woorden}}{\text{woorden/min.}} \times 60 \text{ seconden} = \frac{\text{woorden}}{265 \text{ woorden/min}} \times 60 \text{ seconden}$$

Waarbij:

- woorden = totale hoeveelheid woorden in een tekst
- woorden/min = aantal woorden dat een persoon gemiddeld in een minuut leest = 265 [34]

Zoals eerder vermeld biedt SKORT ook de optie om deze leestijd te personaliseren. Dit wordt gedaan aan de hand van een eenvoudige tekst die een vastgestelde 156 woorden bevat. De gebruiker wordt gevraagd de timer te starten en vervolgens de tekst volledig door te lezen. Wanneer de tekst volledig is gelezen, stopt de gebruiker de timer en wordt de gepersonaliseerde leessnelheid van de gebruiker bepaald. Deze leessnelheid vervangt vervolgens de initiële 265 woorden per minuut en de leestijd wordt opnieuw bepaald. De leessnelheid wordt tevens opgeslagen binnen het prototype, zodat bij later gebruik van SKORT de gebruiker niet nogmaals zijn/haar leessnelheid hoeft te bepalen.

6.3 Question generation

Het vraag-generatie onderdeel van SKORT is als volgt gemaakt: eerst worden de gevonden kernwoorden gekozen als antwoorden, waarna deze antwoorden samen met de aangeleverde tekst van de gebruiker in een model worden gestopt als input tekst. De input tekst heeft de volgende structuur:

```
input_text = "answer: %s  context: %s </s>" % (answer, context)
```

Hierbij is 'context' de aangeleverde tekst en 'answer' het gevonden kernwoord. Deze invoer tekst wordt getokenized voor een T5 base model [35]. De output van de tokenizer wordt in een aangepast T5 model (die afgestemd is op het genereren van vragen) gestopt. het resultaat dat uit het model komt wordt weer gedetokenized om uiteindelijk een vraag te genereren.

Het T5 model is gemaakt door Google Research met als doeleinde een algemeen text-to-text model te creëren die gefinetuned kan worden op, volgens de auteurs, ieder NLP probleem. Hierbij is door Manuel Romero een T5 model gefinetuned op de SQuAD v1.1 dataset [36] voor het genereren van vragen. Verdere uitleg van de implementatie is in het Question Generation notebook-bestand te vinden.

6.4 Moeilijkheidsgraad

Een ander onderdeel van SKORT omvat het bepalen van de moeilijkheidsgraad van de, door de gebruiker geleverde, tekst. Om deze moeilijkheidsgraad te kunnen bepalen zijn een aantal methodieken onderzocht en geëxperimenteerd, die onderverdeeld kunnen worden in zogenaamde Tokenized Measures (TM) en Syllables Measures (SM):

Tokenized Measures:

- MATTR (Moving Average Type Token Ratio) [37]
- HD-D (Hypergeometric Distribution Diversity Index) [38]

- MTLD (Measure of Textual Lexical Diversity) [38]

Syllables Measures:

- Flesh-reading index [39]
- SMOG-index (Simple Measure of Gobbledygook) [40]

De experimenten zijn uitgevoerd aan de hand van de Python library `lexicalrichness`, die allerlei *measures* bevat voor het bepalen van de moeilijkheidsgraad van teksten. Tijdens de experimenten toonden de Flesh-Reading index en de SMOG-index aan het meest gebruiksvriendelijk te zijn. Bij deze twee methoden bleek preprocessing (i.e. voorbewerken) van tekst, zoals het verwijderen van leestekens en witregels, niet nodig. Wat betreft performance behaalden MTLD, Flesh-Reading index en SMOG-index de hoogste score. Deze performance is gemeten door te kijken in hoeverre de drie methodes de verschillende niveau's van de papers kon onderscheiden. Een van de papers was namelijk zelfgeschreven door een student en dus van een wat lager tekstniveau. Daarnaast was een van de papers een gepubliceerd en hoogstaande paper over AI. Uiteindelijk is ervoor gekozen om de Flesh-Reading index toe te passen in het prototype, gezien deze op beide vlakken goed presteerde. Exacte resultaten zijn te vinden op de GitLab-pagina.

6.5 Summarizer

Een ander hoofdonderdeel wat is geïmplementeerd in SKORT, is een zogenaamde summarizer, een model dat in staat is een (korte) samenvatting te genereren van een tekst.

Voor het bepalen van de summarizer dataset zijn een aantal technische en ethische requirements opgesteld, weergegeven in Tabel 4.

Nummer	Categorie	Eis	Afkomstig van
1	Technisch	De dataset moet teksten en samenvattingen van deze teksten bevatten geschreven in de Engelse taal	Groep
2	Technisch	De teksten in de dataset moeten minstens 200 woorden bevatten (niet geldend voor de samenvattingen)	Groep
3	Technisch	Zowel de teksten als de samenvattingen moeten grammaticaal kloppen	Groep
4	Technisch	De dataset moet minstens 1000 teksten met bijhorende samenvattingen bevatten	Groep
5	Ethisch	De dataset moet geen privacygevoelige gegevens bevatten (privacy)	AVG
6	Ethisch	De teksten (en samenvattingen) moeten geen opinies bevatten van mogelijke biased auteurs (fairness)	Groep
7	Ethisch	Er moet duidelijk informatie over de herkomst van data in de dataset te vinden zijn (transparency)	Groep
8	Ethisch	De teksten aanwezig in de dataset hebben goedkeuring gekregen van de originele auteurs voor gebruik (consent)	Auteursrecht
9	Ethisch	De dataset moet voldoen aan de wetgeving omtrent het gebruik van online verkrijgbare datasets (compliance)	Groep
10	Ethisch	De dataset moet goed beveiligd zijn tegen misbruik (security)	Groep

Tabel 4: Dataset requirements

Aan de hand van de bovenstaande lijst van eisen zijn een aantal datasets onder de loep genomen:

- **BillSum** [41]: een dataset die samenvattingen bevat van Amerikaanse (specifiek Californië) wetgeving.

- **SAMSum** [42]: een dataset die samenvattingen bevat van langdurige chatberichten tussen personen.
- **BigPatent** [43]: een dataset die handgeschreven samenvattingen bevat van patenten.
- **CNN/Daily Mail** (par. 4.3 in [44]): een dataset die samenvattingen van nieuwsartikelen bevat afkomstig van de CNN en Daily Mail.
- **Multi-Xscience** [45]: een dataset die samenvattingen bevat van wetenschappelijke artikelen.

Hieruit kwam naar voren dat de CNN/Daily Mail dataset aan de meeste opgestelde eisen voldeed, afgezien van de Content en Security Policy (CSP). De dataset bevat zo'n 300.000 (objectieve) nieuwsartikelen met bijbehorende samenvattingen geschreven in het Engels. De inhoud van de dataset is vervolgens onderzocht, bijvoorbeeld door in kaart te brengen wat de hoeveelheid woorden per artikel is en welke woorden het vaakst voorkomen in de artikelen. Uiteindelijk zijn de eerste 20.000 (en later 40.000) artikelen uit de dataset geselecteerd, wegens geheugen limitaties van het operating system.

Voordat de artikelen getokenized konden worden, werden alle artikelen eerst gepreprocessed totdat alle teksten alleen nog uit cijfers en letters bestonden. Daarnaast werden twee stopwoorden <START> en <STOP> aan de samenvattingen toegevoegd, benodigd voor het gebruik in een encoder-decoder model [46].

Voor het model in kwestie is er dus gekozen voor een encoder-decoder model, opgesteld uit LSTM (long short-term memory) [47] units. De eerste iteratie van dit model bestond uit 500 LSTM units voor zowel de encoder als de decoder, wat het model op meer dan 77.000.000 modelparameters bracht. Echter bleek dit computationeel te zwaar te zijn voor de GPU (graphics processing unit) van het operationele systeem, waarnaar het model is verkleind naar 250 LTSM units. Vervolgens zijn er meerdere iteraties betreffende het model uitgevoerd, om de performance van het model te kunnen verbeteren. Deze iteraties en de trainingsresultaten hiervan zijn onder het kopje Summarizer terug te vinden. De testresultaten van het zelfgetrainde LSTM model zijn in Tabel 5 weergegeven, naast testresultaten van enkele pre-trained summarizer modellen, die ontwikkeld zijn door Sam Shleifer. Het gaat in dit geval om distilBART modellen (kleinere versie van BART [48]). Voor de distilBART modellen is maar één iteratie uitgevoerd, aangezien de vastgestelde waarden van HuggingFace voor de toegepaste modellen gebruikt worden.

ROUGE-L (F1)	Iteratie 1	Iteratie 2	Iteratie 3	Iteratie 4
LSTM model	0.04	0.03	0.04	0.06
distilbart-cnn-12-6	0.31	-	-	-
distilbart-xsum-12-1	0.33	-	-	-
distilbart-xsum-9-6	0.36	-	-	-

Tabel 5: Testresultaten summarizers

De testresultaten laten zien dat het model qua performance, gemeten aan de hand van de ROUGE-L score [49], niet goed presteert. Zeker als er wordt vergeleken met voorgetrainde summarizer modellen van BART. Hierom is er uiteindelijk besloten het zelfgetrainde model niet toe te passen in het prototype, maar te kiezen voor een voorgetraind distilBART model, omdat BART te veel parameters bevat en niet gerund kan worden op de computers die tot onze beschikking zijn. Dit voorgetrainde model in kwestie is distilbart-cnn-12-6 geworden.

6.6 Term definition generation

Het laatste onderdeel waar SKORT over beschikt is het genereren van definities bij vaktermen en jargon die voorkomen in de geüploade paper. In dit onderdeel van het project worden kernwoorden dus gedefinieerd als vaktermen en jargon. Dit omdat belangrijke woorden niet altijd lastig zijn en dus geen definitie nodig hebben. Om deze definities te genereren is de ChatGPT [50] API gebruikt.

Er zijn verschillende aanbieders van NLP API's, zoals IBM Watson, Google Cloud Natural Language, of Microsoft Azure Text Analytics. Deze API's bieden functies voor het verwerken van menselijke taal, zoals analyseren hoe vaak bepaalde woorden voorkomen, sentimentanalyse of zinssegmentatie. In dit project wordt de ChatGPT API toegepast om definities van vaktermen te genereren, gezien de hype die momenteel rond de chatbot heerst.

Wetenschappelijke artikelen bevatten relatief moeilijke vaktermen om te beschrijven waar het artikel over gaat. Voor lezers uit hetzelfde onderzoeksgebied zijn de vaktermen eenvoudig te begrijpen, maar voor anderen niet. Daarom heeft SKORT de optie geïmplementeerd om definities van vaktermen te genereren die lezers van andere domeinen kunnen helpen de basisconcepten van het onderzoek te begrijpen.

Om het genereren van definities te realiseren in SKORT, waren een aantal stappen nodig. Allereerst is de tekst ge-extract uit de PDF-file met behulp van de `PyPDF2`-library in Python. Vervolgens is de tekst voorbewerkt. Dit was nodig, omdat de ChatGPT API vanuit de cloud maar 2048 tokens (1000 tokens $\approx$ 750 woorden) kan verwerken. Een wetenschappelijk artikel bevat uiteraard gemiddeld meer dan 2048 tokens. Om het aantal tokens te verlagen, zijn eerst de referenties verwijderd. Vervolgens zijn alle stopwoorden en tekens verwijderd met de `nltk`-library. Daarna zijn alle duplicaties in woorden verwijderd, waardoor er alleen nog unieke woorden overbleven. Deze unieke woorden waren de invoer voor de ChatGPT API en alle vaktermen worden geëxtraheerd. De model engine `text-davinci-003` presteerde het beste na het testen van de rekentijd en output van elke model engine. De resultaten van het kleine experiment zijn onder het kopje Resource Demand (Model engines ChatGPT API) te vinden. De vaktermen waren opnieuw de invoer voor de ChatGPT API, maar dit keer heeft ChatGPT de definities van de vaktermen gegeven. De vaktermen en bijbehorende definities zijn in een dictionary geplaatst. Tot slot zijn de vaktermen opgezocht in de originele tekst en is de definitie bij de vaktermen geplaatst in een tooltip. De complete code is op deze pagina te vinden en te downloaden als notebook bestand (.ipynb).

6.7 Prototype evaluatie

Aan het einde van de tweede iteratie is het definitieve prototype geëvalueerd op de punten die opgesteld zijn in §3.9. De resultaten uit deze evaluatie waren als volgt:

1. **Betrouwbaarheid**: om de betrouwbaarheid van SKORT te kunnen evalueren zijn tien verschillende papers op het SKORT prototype getest. Deze set bevatte zowel door studenten geschreven papers als gepubliceerde papers, ieder geschreven in de Engelse taal. De gehele set papers is specifiek getest op de kernwoordextractie en de summarizer. Zowel bij de kernwoordextractie als de summarizer deden zich geen problemen voor en het prototype bleef naar behoren functioneren. Verder kon er, door middel van menselijke evaluatie binnen de projectgroep, geen duidelijke bias terug gevonden worden in de resultaten. De betrouwbaarheid van het prototype wordt daarom op niveau geacht.

2. **Nauwkeurigheid**: om de nauwkeurigheid van SKORT te kunnen evalueren zijn dezelfde tien papers als bij het evalueren van de betrouwbaarheid gebruikt. Ditmaal werd het on-

derdeel kernwoordextractie specifiek onder de loep genomen. Hierbij werden de gevonden kernwoorden (maximaal tien kernwoorden) door SKORT vergeleken met de kernwoorden die door het team zijn opgesteld, uitgevoerd op meerdere (korte) teksten. De resultaten hierbij waren positief, de kernwoorden gevonden door SKORT kwamen in zeven van de tien gevallen overeen met de door het team opgestelde kernwoorden. Echter viel wel op dat de kernwoorden wat minder goed overeen kwamen bij de door studenten geschreven papers dan bij de gepubliceerde papers.

3. **Gebruiksgemak**: om het gebruiksgemak van SKORT te kunnen evalueren is een testsessie gehouden met testpersonen, afkomstig van verschillende studies/disciplines. De testmethodiek uit de eerdere gebruikerstesten met het paper prototype werd hier wederom gevolgd. Waarbij testpersonen zelf uit de voeten moeten kunnen met het prototype met alleen de informatie dat wordt aangeleverd in het prototype. De feedback gegeven door de testpersonen betreffende het gebruiksgemak was over het algemeen positief (zie het logboek). Pluspunten van de testpersonen waren vooral met betrekking tot de eenvoud/overzichtelijkheid van het prototype en de snelheid van het prototype in het algemeen. De snelheid van de summarizer werd wat negatiever geacht, gezien het prototype voor een ogenblik bevroor bij het opstellen van een samenvatting. Desondanks wordt het gebruiksgemak van het SKORT prototype op gewenst niveau gewaardeerd.

7 Discussie

Tijdens het ontwikkelen van het prototype en de modellen is er zo goed mogelijk nagedacht over en gewerkt aan het voldoen aan de opgestelde eisen. Echter zijn er alsnog een aantal punten naar voren gekomen, waar meer aandacht aan besteed had kunnen worden. Deze punten waren als volgt:

1. **Sci-Hub**: in het huidige prototype is er een mogelijkheid om papers van Sci-Hub in te kunnen laden. Echter is het gebruik van Sci-Hub ethisch gezien vrij discutabel, gezien het gebruik van papers op Sci-Hub als piraterij geacht kan worden. Hierom is er besloten de Sci-Hub optie in het huidige prototype voor nu te laten, maar zal dit uit het prototype verwijderd worden indien SKORT online vrijgegeven zal worden.

2. **Kernwoord extractie test:** in Tabel 1 in §3.7 wordt bij eis 8 vermeld dat de gevonden kernwoorden dicht bij de kernwoorden moeten liggen die een gemiddeld persoon uit de doelgroep zou kiezen. Echter is er tijdens het evalueren alleen met onze eigen gekozen kernwoorden vergeleken. Idealiter was er nog een testsessie gehouden waarbij de gegenereerde kernwoorden van SKORT vergeleken konden worden met die van testpersonen.

3. **Kernwoord extractie onderzoek:** in §6.1 wordt het onderzoek naar de kernwoordextractiemethoden besproken. Deze methoden zijn getest op slechts drie korte teksten met bijbehorende kernwoorden. Om een betrouwbaarder resultaat te kunnen bewerkstelligen hadden langere en ook meer teksten gebruikt kunnen worden ter evaluatie.

4. **Website uitbreiding:** in het huidige prototype is er de mogelijkheid om een paper van Sci-Hub te halen of een eigen paper als PDF-file aan te leveren. Naast het feit dat Sci-Hub niet de meest ethische website is, had de mogelijkheid om papers van elke website te kunnen gebruiken in de SKORT Tool mogelijk verwerkt kunnen worden indien meer tijd beschikbaar was.

5. **Summarizer snelheid:** het huidige prototype bevriest voor een ogenblik tijdens het gebruik van de summarizer, gezien deze een volledige tekst ter plaatse moet gaan samenvatten. Idealiter had dit eerder moeten gebeuren tijdens het laadscherm, immers daar is het laadscherm voor bedoeld.

6. **Visuele weergave prototype:** in het huidige prototype ontbraken nog een aantal functies die de visuele weergave binnen het prototype hadden kunnen verbeteren. Zo had er een zoom functie voor het lezen van de papers toegevoegd kunnen worden en had de scrollbar in het prototype op een handigere plek geplaatst kunnen worden.
7. **Bias:** in §3.8 zijn een aantal mogelijke vormen van bias opgesteld die in het SKORT prototype kunnen voorkomen. Echter zijn bij het ontwikkelen van het prototype niet voor iedere mogelijke bias maatregelen genomen. Met name demografische bias en historische bias hadden beter behandeld kunnen worden in het ontwikkelen van het prototype.

7.1 Vervolg onderzoek

Tijdens dit project is er gebruikt gemaakt van meerdere state-of-the-art Transformer modellen. Deze modellen zijn op dit moment de beste keuzes voor de toepassing binnen dit project. In de toekomst zullen er betere en mogelijk nieuwe technieken en modellen worden uitgebracht. Wanneer dit product werkelijk op de markt zou komen zal er met deze toekomstige modellen rekening gehouden moeten worden.

Verder is het mogelijk om specifieker op een bepaalde doelgroep te richten zoals bijvoorbeeld Artificial Intelligence artikelen en hierbij de state-of-the-art modellen te finetunen wanneer er genoeg training data aanwezig is.

8 Conclusie

In §3.7 zijn eisen voor dit project opgesteld in de vorm van een MoSCoW prioritering. Onderstaande doelstellingen zijn behaald.

1. Het prototype van SKORT kan kernwoorden uit een tekst halen en deze markeren.
2. De leestijd voor de paper wordt berekend en is personaliseerbaar.
3. De moeilijkheidsgraad voor de paper wordt berekend en aan de gebruiker getoond.
4. Definities van termen moeten in de tool op te zoeken of terug te vinden zijn.
5. De student mag niet dommer worden van de tool; de tekst mag geen informatie verliezen.
6. Het SKORT model, voor kernwoordextractie, moet op gebied van performance beter presteren dan het vastgestelde nulmodel.
7. De SKORT tool wordt gemonitord en onderhouden door het team en zal niet worden uitbesteed.
8. De lay-out van SKORT moet zo dicht mogelijk bij e-ink (i.e. e-paper of elektronisch papier) [4] in de buurt komen om de leesbaarheid te optimaliseren (readability).

Het ontworpen systeem kan gezien worden als een functionerende proof-of-concept dat als basis kan worden gebruikt voor het ontwikkelen van een (web)applicatie.

Referenties

[1] J. Ball, 2019, 'The Double Diamond: A universally accepted depiction of the design process', Design Council. https://www.designcouncil.org.uk/our-work/news-opinion/double-diamond-universally-accepted-depiction-design-process/

[2] Nationale Onderwijsgids, 2013, 'Markeren van studiestof geen optimale leermethode'. https://www.nationaleonderwijsgids.nl/hbo/nieuws/15139-markeren-van-studiestof-geen-optimale-leermethode.html

[3] ProductPlan, z.d., 'MoSCoW prioritization', https://www.productplan.com/glossary/moscow-prioritization/

[4] Wikipedia, z.d., 'Electronic paper', https://en.wikipedia.org/wiki/Electronic_paper

[5] Hogeschool Leiden, 2021, 'Hoe leren studenten in het hoger beroepsonderwijs'. https://www.hsleiden.nl/binaries/content/assets/hsl/over-hl/hl-canon-hoe-leren-studenten.pdf

[6] Hogeschool Leiden, z.d., 'Levensvaardigheden (VO)'. https://www.hsleiden.nl/ouderschap-en-ouderbegeleiding/onderzoek/levensvaardigheden-vo

[7] J.M. Dopmeijer, 2021, 'Running on empty. The impact of challenging student life on wellbeing and academic performance.', Proefschrift, Amsterdam: UVA.

[8] Hogeschool Leiden, 2019, 'Levensvaardigheden voor studenten.', Lectoraat Ouderschap & Ouderbegeleiding, Utrecht University Library (202441), Boom Lemma uitgevers.

[9] E. Kazemier, J. Offringa, L. Eggens & M. Wolfensberger, 2014, 'Motivatie en leerstrategieën van honoursstudenten'. https://oadoi.org/10.5553/tvho/016810952014032001009

[10] P. Jüni, D.G. Altman, M. Egger, 'Assessing the quality of controlled clinical trials', BMJ 2001; 323 :42 doi:10.1136/bmj.323.7303.42

[11] Australian Human Rights Commission, 2020, 'INFOGRAPHIC: Historical bias in AI systems', https://humanrights.gov.au/about/news/media-releases/infographic-historical-bias-ai-systems

[12] De Nederlandse Grondwet, z.d., 'Artikel 10: Privacy', https://www.denederlandsegrondwet.nl/id/via0hb5jcjzv/artikel_10_privacy

[13] Redactie KVK, 2022, '5 tips om reputatieschade te voorkomen', Kamer van Koophandel, https://www.kvk.nl/advies-en-informatie/veiligzakendoen/5-tips-om-reputatieschade-te-voorkomen/#:~:text=Reputatieschade%20is%20meestal%20het%20gevolg,oorzaken%20zijn%20cybercrime%20en%20datalekken.

[14] Nederlands Jeugdinstituut, z.d., 'Armoede en welbevinden', https://www.nji.nl/armoede/welbevinden#:~:text=Sociale%20uitsluiting,beperkte%20ontwikkeling%20en%20sociaal%20isolement.

[15] T. B. Sheridan & W. L. Verplank, 1978, 'Human and Computer Control of Undersea Teleoperators', MIT, Technical rept. 15 Mar 1977-14.

[16] B. Shneiderman, 2020, 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy', arXiv.org. https://arxiv.org/abs/2002.04087v1

[17] B. Shneiderman, 2020, 'Human-Centered Artificial Intelligence: Three Fresh Ideas', AIS Transactions on Human-Computer Interaction, 12(3), 109-124. https://doi.org/10.17705/1thci.00131 DOI: 10.17705/1thci.00131

[18] A. Elbakyan, z.d., 'Sci-Hub'. https://sci-hub.ru/about

[19] M. Honnibal & I. Montani, 2017, 'spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing'

[20] S. Rose, D. Engel, N. Cramer & W. Cowley, 2010, 'Automatic Keyword Extraction from Individual Documents', In Text Mining (eds M.W. Berry and J. Kogan). https://doi.org/10.1002/9780470689646.ch1

[21] Grootendorst, 2020, 'KeyBERT: Minimal keywords extraction with BERT', Zenodo. https://doi.org/10.5281/zenodo.4461265

[22] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin & C.G. Nevill-Manning, 1999, 'Kea: Practical automatic keyphrase extraction', In Proceedings of the Fourth ACM Conference on Digital Libraries, pages 254–255, ACM.

[23] S.N. Kim, O. Medelyan, M.-Y. Kan & T. Baldwin, 2010, 'SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles', Association for Computational Linguistics, Proceedings of the 5th International Workshop on Semantic Evaluation. https://aclanthology.org/S10-1004/

[24] S.R. El-Betagy & A. Rafea, 2010, 'KP-Miner: Participation in SemEval-2', Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 190–193. https://aclanthology.org/S10-1041.pdf

[25] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes & A. Jatowt, 2020, 'YAKE! Keyword extraction from single documents using multiple local features', Information Sciences, Volume 509, Pages 257-289, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.09.013

[26] X. Wan & J. Xiao, 2008, 'CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction', Institute of Computer Science and Technology, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 969–976. https://aclanthology.org/C08-1122.pdf

[27] A. Bougouin, F. Boudin & B. Daille, 2013, 'TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction', International Joint Conference on Natural Language Processing, pages 543–551. https://aclanthology.org/I13-1062.pdf

[28] L. Sterckx, T. Demeester, J. Deleu, & C. Develd, 2015, 'Topical Word Importance for Fast Keyphrase Extraction', WWW 2015 Companion, May 18–22, ACM 978-1-4503-3473-0/15/05. http://users.intec.ugent.be/cdvelder/papers/2015/sterckx2015wwwb.pdf

[29] C. Florescu & C. Caragea, 2017, 'PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents', Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1105–1115. https://aclanthology.org/P17-1102.pdf

[30] F. Boudin, 2018, 'Unsupervised Keyphrase Extraction with Multipartite Graphs'. https://arxiv.org/abs/1803.08721

[31] P. Godec, 2021, 'Keyword Extraction Methods - The Overview', Towards Data Science. https://towardsdatascience.com/keyword-extraction-methods-the-overview-35557350f8bb

[32] I. Shrivastava, 2020, 'Exploring Different Keyword Extractors - Graph Based Approaches', Medium. https://medium.com/gumgum-tech/exploring-different-keyword-extractors-graph-based-approaches-c46ec6c12c34

[33] A.R. Lahitani, A.E. Permanasari & N.A. Setiawan, 2016, 'Cosine similarity to determine similarity measure: Study case in online essay assessment'. 2016 4th International Conference on Cyber and IT Service Management. doi:10.1109/citsm.2016.7577578

[34] Medium, z.d. 'Read Time' https://help.medium.com/hc/en-us/articles/214991667-Read-time

[35] A. Roberts, 2020, 'Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer', Google Research.

[36] Wolfram Research, 2018, 'SQuAD v1.1', the Wolfram Data Repository https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

[37] M.A. Covington & J.D. McFall, 2010, 'Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)', Journal of Quantitative Linguistics, 17:2, 94-100, DOI: 10.1080/09296171003643098

[38] P.M. McCarthy & S. Jarvis, 2010, 'MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment', Behavior Research Methods 42, 381–392. https://doi.org/10.3758/BRM.42.2.381

[39] Readable, z.d., 'Flesch Reading Ease and the Flesch Kincaid Grade Level' https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

[40] Readable, z.d., 'The SMOG Index' https://readable.com/readability/smog-index/

[41] A. Kornilova & V. Eidelman, 2019, 'BillSum: A Corpus for Automatic Summarization of US Legislation', arXiv preprint: 1910.00523

[42] B. Gliwa, I. Mochol, M. Biesek & A. Wawer, 2019, 'SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization', In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

[43] E. Sharma, C. Li & L. Wang, 2019, 'BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization', arXiv preprint: 1906.03741

[44] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre & B. Xiang, 2016, 'Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond', arXiv prePrint: 1602.06023v5

[45] Y. Lu, Y. Dong & L. Charlin, 2020, 'Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles', arXiv prePrint:2010.14235

[46] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk & Y. Bengio, 2014, 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation', arXiv prePrint: 1406.1078, Computation and Language (cs.CL), Machine Learning (cs.LG), Neural and Evolutionary Computing (cs.NE), Machine Learning (stat.ML), FOS: Computer and information sciences, FOS: Computer and information sciences, DOI: 10.48550/ARXIV.1406.1078, https://arxiv.org/abs/1406.1078

[47] S. Hochreiter, & J. Schmidhuber, 1997, 'Long short-term memory', Neural computation, 9(8), 1735–1780.

[48] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov & L. Zettlemoyer, 2019, 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', arXiv prePrint: 1910.13461, Computation and Language (cs.CL), Machine Learning (cs.LG), Machine Learning (stat.ML), FOS: Computer and information sciences, FOS: Computer and information sciences https://doi.org/10.48550/arxiv.1910.13461

[49] C.-Y. Lin, 2004,'ROUGE: A Package for Automatic Evaluation of Summaries' In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[50] OpenAI, 2022, 'ChatGPT: Optimizing Language Models for Dialogue', https://openai.com/blog/chatgpt/