24-6-2020

Tekst classificatie als automatisering bij de analyse van open begrotingen

Consultants in organisatie en techniek

K & R

UNIVERSITY OF TWENTE.

Mats Tavenier

Dit rapport is geschreven voor K&R Consultants en de examinatoren van de Universiteit Twente in het kader van de bachelor scriptie voor de bachelor opleiding Technische Bedrijfskunde.

Universiteit Twente
BSc Technische Bedrijfskunde
Postbus 217
7500 AE Enschede

K&R Consultants
Gladsaxe 26
7327 JZ, Apeldoorn

Auteur:

Mats Tavenier
S1865617

Begeleider K&R Consultants
R.J. (Robert Jan) Zwama

Begeleiders Universiteit Twente
Dr. M.C. (Matthieu) van der Heijden
Dr. A.B.J.M. (Fons) Wijnhoven

Publicatie datum: 24-6-2020

Aantal pagina's zonder bijlagen: 36
Aantal pagina's met bijlagen: 46
Aantal bijlagen: 7

Voorwoord

Beste lezer,

Voor u ligt mijn scriptie waarin ik onderzoek heb gedaan bij K&R Consultants in Apeldoorn. Deze thesis is geschreven in het kader van het afronden van de bachelor studie Technische Bedrijfskunde aan de Universiteit Twente.

Er zijn een aantal mensen die ik graag wil bedanken voor hun bijdrage aan het onderzoek. Ten eerste wil ik de werknemers van K&R Consultants bedanken voor hun open en gastvriendelijke houding. Ik heb mij de afgelopen periode zeer thuis gevoeld en nooit bezwaard gevoeld om vragen te stellen of onderwerpen aan te kaarten. In het bijzonder wil ik Robert Jan bedanken voor de prettige samenwerking, waarin je altijd beschikbaar was om de richting en inhoud van mijn onderzoek te bespreken.

Verder wil ik Matthieu hartelijk bedanken voor je tijd en inzet bij het begeleiden van mijn onderzoek. Ik heb je feedback als erg nuttig beschouwt, waardoor mijn scriptie tot een hoger niveau werd gebracht.

Tenslotte wil ik mijn ouders, vrienden en huisgenoten bedanken voor hun ondersteuning gedurende het onderzoek.

Dan wens ik u verder veel leesplezier.

Management samenvatting

Aanleiding

Dit onderzoek vindt plaats bij K&R Consultants (K&R), een installatietechnisch adviesbureau in Apeldoorn. Een van de kerntaken is het maken van kostenanalyses op de open begrotingen van verschillende bouwprojecten. Deze kostenanalyses worden uitgevoerd met de hulp van prijzen in een zelfontwikkelde database van ongeveer 2500 bouwelementen. K&R's manier van opstellen van kostprijzen draait om een 'optimum level of detail', waarbij de kosten op een detailniveau worden besproken waardoor een duidelijk kostenbeeld ontstaat voor klanten. Deze wijze verschilt van de calculatieschema's waar aannemers gebruik van maken voor het opstellen van open begrotingen. Hierdoor wordt een kostenanalyse bemoeilijkt vanwege de beperkte vergelijkbaarheid tussen de twee manieren van prijs opstellingen. Het vergelijkbaar maken van de aannemers prijzen met de K&R database is een repetitieve en tijdsintensieve werkzaamheid, waarin mogelijkheden voor automatisering zijn.

Aanpak

In de huidige situatie worden de materiaalomschrijvingen van aannemers in open begrotingen handmatig gelabeld, waarbij elk label een bouwelement voorstelt in de database van K&R. In de literatuur is binnen de tak van 'text mining' een onderzoeksveld genaamd tekst classificatie. Hierbij wordt op basis van woorden in tekstdocumenten, zoals artikelen of online berichten, een label toegewezen aan een document. Tekst classificatie lijkt toepasbaar op de situatie van K&R, waarbij er tekstdocumenten (omschrijvingen) zijn waaraan labels moeten worden gegeven.

De Naive Bayes theorie is een theorie die veel wordt gebruikt voor het classificeren van tekst. De theorie wordt geroemd vanwege de eenvoudige en snelle aanpak, waarbij desondanks goede resultaten worden geboekt. Voor K&R is er met behulp van programmeertaal Python een basismodel gemaakt, waarmee onderzocht wordt of deze theorie K&R in de toekomst kan helpen richting automatisering.

Resultaten

De resultaten waren op een nu nog beperkte trainingsset van 20.000 omschrijvingen uit 30 analyses niet erg goed, waarbij zo'n 15 tot 20% van de labels maar goed werden voorspeld. Wanneer er gekeken word naar de top 3 voorspelde labels, bleek er ongeveer 10% aan extra labels goed voorspeld te zijn. De combinatie van beperkte trainingsdata, korte omschrijvingen en het grote aantal labels (2500) is de belangrijkste reden voor de slechte prestaties van het model.

Conclusies

Uit dit verkennende onderzoek naar tekstclassificatie als verbetering voor het analyseproces van K&R kunnen ondanks de ogenschijnlijk lage nauwkeurigheidspercentages, enigszins hoopvolle conclusies getrokken worden. De trainingsdata kan relatief makkelijk worden opgeschaald waardoor de prestaties van het model zich zullen verbeteren. De relatief korte omschrijvingen en de grote aantallen labels (2500) maken het een moeilijk tekst classificatie project. De omschrijvingen kunnen niet worden aangepast, maar de labels kunnen wel verminderd worden. Vervolgonderzoek naar de mogelijkheid om het aantal labels te verlagen, waarbij de prijsdatabase van K&R zijn waarde wel behoudt, is zeer interessant.

Dit onderzoek levert een basismodel waarmee bewezen wordt dat tekstclassificatie mogelijk uitkomsten biedt aan K&R. Ook is dit model een eerste stap richting automatisering waarop doorgebouwd kan worden. Ten slotte kan een nieuwe inrichting van de prijsdatabase de automatisering door middel van tekstclassificatie tot een succes maken.

Inhoudsopgave

Definities

Analysebestand = Een speciaal ingericht Excel-bestand waarin op een begroting een kostenanalyse wordt losgelaten.

Calculaties schema's = Een schema waarmee aannemers de kosten van een bouwproject calculeren.

Kosten analyseproces (binnen K&R) = Het werkproces waarbij de prijzen van open begrotingen worden vergeleken met marktconforme prijzen in K&R's database.

Machine Learning (ML) = Het onderzoeksveld binnen kunstmatige intelligentie waarbij wordt ingegaan op de ontwikkeling van algoritmes en computertechnieken met behulp van data.

Natural Language Processing (NLP) = de poging om een volledigere betekenis af te leiden uit vrije tekst.

Naïve Bayes (NB) theorie = Een theorie veel gebruikt in tekst classificatie, waarbij op basis van de Bayes theorie en een aanname van onafhankelijkheid tussen 'features' classificaties kunnen worden voorspeld.

Omschrijvingscel = Een Excel-cel waarin een omschrijving van een bouwelement staat.

Open begroting = Een open begroting is een uitgewerkte offerte waarin alle bouwmaterialen, hoeveelheden en werkuren zijn meegenomen.

Text mining = Het proces waarbij waardevolle informatie en patronen uit tekstbestanden worden onttrokken.

Vectoriseren = Het omzetten van woorden in vectoren.

1 Probleemidentificatie en probleemaanpak

In dit hoofdstuk is de context neergezet, waarna daarbinnen de verschillende problemen en het kernprobleem geïdentificeerd zijn. Met die identificatie is er vervolgens een probleemaanpak opgesteld die wordt uitgewerkt in meerdere onderzoeksvragen die in deze scriptie zullen worden beantwoord.

1.1 Bedrijfscontext

K&R Consultants (K&R) is een installatietechnisch adviesbureau gevestigd in Apeldoorn. Een van de activiteiten waar zij zich bezig mee houden, is het analyseren van de prijzen in open begrotingen van aannemers. In deze open begrotingen vermelden aannemers al hun verwachte kosten voor een bouwproject, zoals bouwmaterialen, hoeveelheden en werkuren. K&R wordt vaak ingehuurd door bedrijven en instellingen, zoals universiteiten, om hen te helpen in het analyseren van open begrotingen van verschillende aannemers. Deze bedrijven of instellingen hebben vaak niet de kennis in huis om de juiste keuze te maken met de betrekking tot aannemers, omdat zij niet op de hoogte zijn van de prijzen in de bouwsector. Om toch tot de juiste beslissing te komen, huren zij K&R in om hen te helpen bij het maken van een beslissing tussen de open begrotingen.

Voor het analyseren van de open begrotingen heeft K&R een database gemaakt, waarin marktconforme prijzen van bijna alle materialen van een bouwproject zijn opgenomen. Deze databaseprijzen zijn gebaseerd op de open begrotingen die zij de afgelopen jaren hebben geanalyseerd. Samen met de expertise van de werknemers kan K&R met deze database een betrouwbare conclusie trekken over de volledigheid en eerlijkheid van een open begroting. Dit helpt de bedrijven en instellingen die K&R inhuren bij het maken van een beslissing tussen verschillende aannemers.

1.2 Probleemcontext

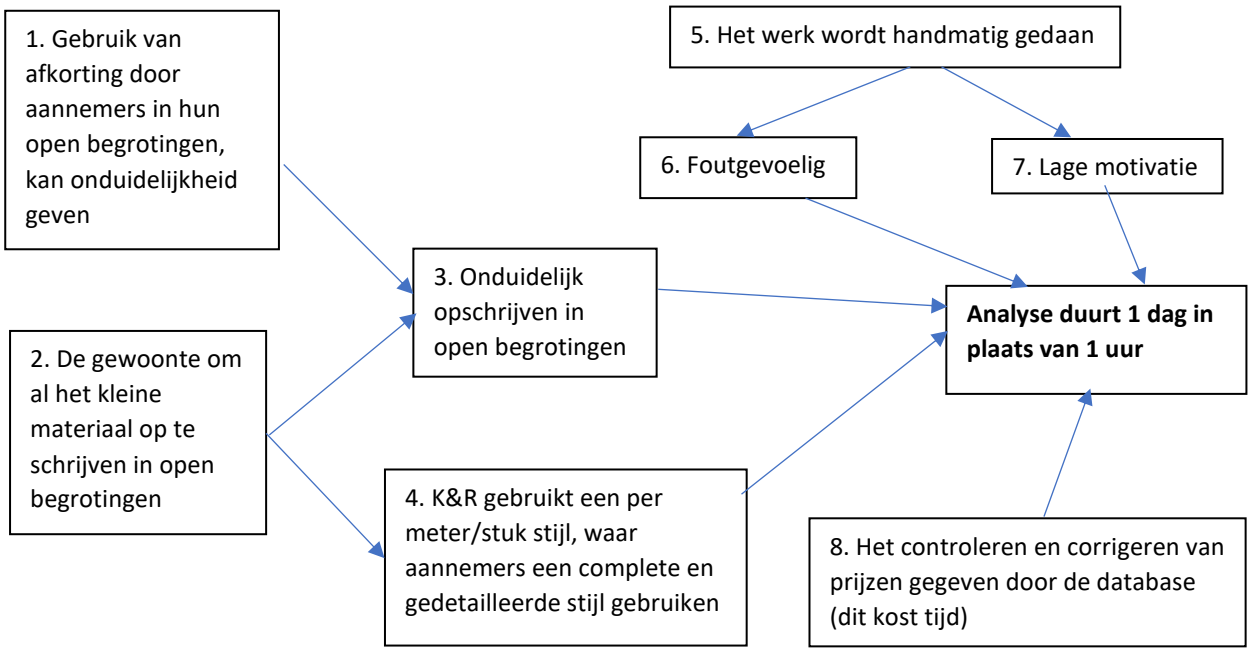
Het probleem dat K&R heeft, is dat het vergelijkbaar maken van de open begrotingen van de aannemers met hun database veel tijd kost. K&R moet handmatig elke materiaalomschrijving in een open begroting toewijzen aan een prijsnaam in hun eigen database. Aannemers schrijven hun open begrotingen in een hele uitgebreide manier op, waarbij ze elke schroef, bout en bochtje vermelden dat ze nodig hebben voor het bouwproject. Dit is verschillend van de manier van K&R, aangezien zij de kosten in een per meter of per stuk techniek opschrijven. Deze techniek staat binnen K&R bekend als de ‘optimum level of detail’, omdat dit detailniveau een veel duidelijker kostenbeeld kan geven aan de klanten.

Om het verschil tussen de twee manieren te verduidelijken, gebruiken we een voorbeeld. De Universiteit Twente laat een nieuw collegegebouw neerzetten en zoals in elk gebouw zijn er ook toiletvoorzieningen. Voor deze voorzieningen hebben de aannemers buizen aangelegd om water bij de toiletten te krijgen. Aannemers hebben de gewoonte om in de calculatie schema’s vervolgens elke schroef, bout en bocht nodig voor die buizen op te schrijven. Deze onderdelen worden elk als aparte kostenpost op de open begroting geplaatst. Dit maakt de open begrotingen onduidelijk en lastig om te begrijpen voor de klanten. Om een duidelijker en overzichtelijker beeld van de kosten te geven, gebruikt K&R een andere manier van het kosten opschrijven. Als we het voorbeeld blijven gebruiken, betekent dat dat K&R een prijs per meter voor die buizen heeft, waarin alle schroeven, bouten en bochten zijn meegenomen. Op die manier ontstaat er een kostenpost op het eindblad waarin de prijzen van buizen en de hoeveelheid meter staan. Dit is een veel duidelijker manier van het opschrijven van kosten, die door alle klanten wordt begrepen ook al hebben ze geen ervaring met bouwprojecten.

Kortom, K&R gebruikt een per meter of per stuk stijl voor de prijzen in hun database, wat betekent dat ze de open begrotingen van aannemers moeten herschrijven in diezelfde stijl om een goede vergelijking te maken. Het herschrijven van die kostenposten wordt nu handmatig gedaan, wat veel tijd kost en foutgevoelig is. Dit is het probleem waar K&R mee zit.

1.3 Probleemkluwen en kernprobleem

Deze sectie behandelt de probleemkluwen en de motivatie voor het kernprobleem. Voor een uitgebreidere uitleg van de problemen in de probleemkluwen kan appendix 7.1 (pagina 37) geraadpleegd worden.



Figuur 1 - Probleemkluwen K&R Consultants

In de probleemkluwen in Figuur 1 komt duidelijk naar voor dat het probleem dat 'de analyse 1 dag duurt in plaats van 1 uur' het belangrijkste effect van alle andere problemen is. Er zijn een aantal problemen waarvoor er een directe, onderliggende oorzaak is gevonden. Dit zijn de problemen waarbij er geen pijl meer in het vak gaat, alleen maar een pijl uit het vak. De problemen 1 en 2 zijn niet te beïnvloeden, aangezien deze problemen branchewijd zijn en niet door K&R veroorzaakt worden. De problemen 6 en 7 worden vooral veroorzaakt doordat het proces handmatig, repetitief en tijdsintensief is. Omdat het repetitief is, zou het proces zeker voor een deel geautomatiseerd kunnen worden, wat ook gedeeltelijk de problemen aanpakt dat het handmatig en tijdsintensief is. Probleem 8 wordt veroorzaakt doordat er te weinig analyses per jaar worden gedaan om een valide database met markconforme prijzen te creëren.

Het kernprobleem is dat het proces handmatig is. Wanneer dit wordt opgelost, zal een aantal andere problemen ook opgelost worden, zoals de foutgevoeligheid, de lage motivatie en het controleren en corrigeren van de kostprijzen. Voor dit probleem is er tot nu toe nauwelijks naar mogelijke oplossingen gezocht. Het probleem is namelijk complex, waarbij een oplossing niet gemakkelijk gevonden is. Dat zijn de redenen waarom het probleem 'het werk wordt handmatig gedaan' als kernprobleem wordt gezien.

1.4 Onderzoeksrichting

In de vorige sectie is het kernprobleem ‘het werk wordt handmatig gedaan’ vastgesteld aan de hand van het actieprobleem dat het analyseproces langer duurt dan K&R zou willen. In deze sectie worden de onderzoeksrichting en het onderzoeksdoel die bij deze problemen horen uitgelegd.

De onderzoeksrichting van deze scriptie zijn alle technieken, methodes en programma’s beschikbaar in de literatuur dat K&R kan helpen richting een geautomatiseerd analyse proces. De velden van ‘text mining’, tekst herkenning, databases en software programma’s zijn nuttige richtingen om het lange termijn doel van K&R te bereiken. Door het onderzoeken van de beschikbare informatie, kan er een theorie worden uitgekozen dat toepasbaar is op de situatie van K&R. Aan de hand van deze theorie zal er een praktisch voorbeeld worden uitgewerkt waarmee K&R verdere stappen kan zetten in de toekomst.

De belangrijkste hulpbron tot onze beschikking zijn de labels die de werknemers van K&R de afgelopen twee jaar hebben gegeven aan de geanalyseerde open begrotingen. Deze labels zijn gebaseerd op de database van K&R, die gemaakt is in Excel. De afgelopen twee jaar hebben de werknemers van K&R bij elke analyse die ze gemaakt hebben de open begrotingen en database vergelijkbaar moeten maken. Terwijl zij dit deden, hebben ze labels gehangen aan alle kostposten die ze hebben overgeschreven. Deze labels zijn zeer nuttige informatie, omdat het toont hoe K&R de materiaalomschrijvingen geschreven door aannemers omschrijft naar hun eigen labelnamen die ze in de database gebruiken. Al die labels zijn bekend van de analyses van de afgelopen twee jaar, wat betekent dat er terug te vinden is hoe K&R normaal gesproken een bepaalde kostenpost van een aannemer overschrijft naar hun eigen database naam. De gekozen onderzoeksrichting en de uitkomsten van die richting zullen waarschijnlijk goed aansluiten op de probleemstelling van K&R door de beschikbaarheid van deze labels.

1.5 Perspectief

Deze sectie behandelt vanuit welk perspectief deze scriptie is geschreven en waarom. Verschillende perspectieven komen aan bod en degene die het best past bij deze scriptie wordt naar voren geschoven.

Een logisch perspectief om naar elk probleem te kijken, is door het probleem aan te pakken bij de wortels. In dit geval is dat de manier waarop aannemers hun open begrotingen opschrijven, omdat een groot deel van het probleem voorkomt uit de onduidelijke kosten opschrijfstijl die zij hanteren. Het is echter zeer lastig om uit dit perspectief het probleem aan te pakken, omdat de aannemers een financieel voordeel hebben bij het onduidelijk opschrijven van open begrotingen. Dit wordt gezien als een branchewijd probleem, aangezien bijna elke aannemer eenzelfde soort stijl gebruikt. Voor aannemers is het heel lastig om hun beschrijving van kosten aan te passen, omdat de codenamen in een calculatieschema vaak intern voor de inkoop, projectorganisatie en planning worden gebruikt. Deze onderdelen binnen een aannemersbedrijf zijn erg nauw met elkaar verbonden, waardoor een aannemer met hun eigen, constante stijl blijft werken. Een aannemer heeft dus geen baat bij het aanpassen van opschrijfstijl, omdat dit intern voor veel extra werkt zorgt. De externe begrotingen die naar partijen als K&R Consultants worden gestuurd, zijn daarom gebaseerd op de specifieke stijl van die ene aannemer. Helaas verschillen de stijlen onderling per aannemer, maar het verschilt vaak ook met de stijl van K&R. Dit gat tussen de verschillende stijlen is lastig te overbruggen.

Een ander perspectief dat K&R wel kan helpen, is het verbeteren van de huidige situatie door uit te zoeken hoe het analyseproces geautomatiseerd kan worden. Dit perspectief gaat in op het onderzoeken van de beschikbare mogelijkheden en hoe die toegepast kunnen worden in de situatie van K&R Consultants. In het veld van automatisering is veel mogelijk, maar het is relatief nieuw en

onduidelijk voor K&R wat de precieze mogelijkheden voor hen zijn. Dit perspectief sluit ook mooi aan op de lange termijn doelen van K&R, waarbij ze graag naar een volledig geautomatiseerd portaal voor open begrotingen zouden willen.

Deze scriptie richt ze daarom op de stappen die K&R zou moeten zetten om zich richting dit lange termijn doel te bewegen. Het eindproduct van deze scriptie is dan ook niet een volledig geautomatiseerd systeem, maar een praktijkvoorbeeld en aanbevelingen waarmee K&R stappen kan gaan zetten in de toekomst.

De volgende sectie bespreekt hoe deze scriptie richting dit eindproduct zal toewerken met behulp van verschillende onderzoeksvragen.

1.6 Onderzoeksvragen

Er zijn vier onderzoeksvragen, met twee sub onderzoeksvragen die deze sectie benoemt. Alle onderzoeksvragen samen zullen de hoofd onderzoeksvraag beantwoorden, die als volgt luidt:

Welke soort technieken, methodes en programma's binnen de literatuur kunnen worden toegepast in het analyseproces van K&R Consultants om de bestede tijd per analyse te verlagen?

Deze hoofdonderzoeksvragen geven de structuur van de scriptie aan, waarbij elk hoofdstuk een van de onderzoeksvragen beantwoordt.

1. **Hoe wordt het analyseproces bij K&R Consultants in de huidige situatie uitgevoerd?**
   1.1 Welk deel zou verbeterd kunnen worden en waar zou op gefocust moeten worden?

Om een duidelijk beeld van de huidige situatie waarin de analyse wordt uitgevoerd te krijgen, wordt deze uitgebreid onderzocht. Dit is beschreven in hoofdstuk 2, waarbij er vooral is gelet op het schrijven van een compleet beeld, aangezien een oplossing voor het kernprobleem alleen praktisch nut heeft wanneer het aansluit op de huidige situatie. Met de sub onderzoeksvraag is er een onderdeel van de huidige situatie uitgelicht waarop de rest van de scriptie zich focust.

2. **Welke literatuur kan worden gebruikt om verbetering te realiseren?**
   2.1 Welk deel van de gevonden literatuur is het beste toepasbaar op K&R's analyseproces?

Hoofdstuk 3 beschrijft hoe wij een literatuurstudie hebben uitgevoerd, waarbij verschillende onderwerpen zijn onderzocht zoals tekst classificatie en 'machine learning'. Verder is er een conclusie geschreven waarin de meest belangrijke en relevante informatie is meegenomen, zodat deze informatie makkelijk gebruikt kan worden in volgende onderzoeksvraag.

3. **Wat voor oplossingsontwerp kan er worden opgesteld voor het automatiseren van het analyseproces?**

In hoofdstuk 4 is de informatie van de eerste twee onderzoeksvragen gebruikt om tot een oplossingsontwerp te komen voor de situatie bij K&R Consultants. Met de hulp van programmeertaal Python is er een concept ontwikkeld, waarmee is aangetoond dat de gevonden literatuur ook werkelijk toegepast kan worden in de praktijk. In hoofdstuk 5 worden de resultaten van het oplossingsontwerp besproken.

4. **Welke stappen moet K&R zetten om de bevindingen van deze scriptie in de praktijk te brengen?**

In hoofdstuk 6 zijn er conclusies getrokken op basis van de volledige scriptie, waarna er ook een aantal aanbevelingen zijn gedaan. Met deze conclusies en aanbevelingen kan K&R verder gebracht worden in de praktijk. Daarnaast worden er ook suggesties gedaan voor vervolgonderzoek.

2 Huidige situatie

Om ervoor te zorgen dat de resultaten van deze scriptie goed aansluiten op de situatie bij K&R, is er eerst een uitgebreide analyse gemaakt van de huidige situatie. Dit hoofdstuk zal ingaan op alle aspecten van het analyseproces van open begrotingen door K&R zoals dat nu wordt gedaan. Het proces is in fases verdeeld en elke fase is apart behandelt in een sectie.

2.1 Binnenkrijgen gegevens

Het analyseproces van K&R Consultants begint met het verzamelen van alle informatie. Dit gaat vaak om het opvragen van bestanden bij aannemers en klanten zoals open begrotingen en het bestek. Deze gegevens worden vaak opgestuurd in PDF, zodat de opmaak hetzelfde blijft en lezers het bestand niet zomaar kunnen aanpassen. Een voorbeeld van een open begroting is te vinden in Figuur 2.

Calculatie: 10470201/06112 Offertepost: 010 Heijmans Utiliteit 14.11.2019 Blad : 1

Offertepost : MW 112; WANDCONTACTDOZEN 230V
Valuta : EUR
Werk: HUIZE WESTERLICHT, ALKMAAR
Aanvrager:
Bestek: HUIZE WESTERLICHT ALKMAAR

Reg	Mat-kode	Fabr.	Omschrijving	Aantal	Eh	Netto	Netto-tot	Norm	Norm-tot
1			AANPASSEN WANDCONTACTDOZEN 230V						
2									
3			** ALGEMENE AANSLUITINGEN **						
4	S580800000		INB.DOOS U50	186,00	STK	1,70	316,20	0,3100	57,66
5	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18,00	STK	7,86	141,48	0,2100	3,78
6	S481700000		LASDOOS VOOR WANDGOOT	6,00	STK	3,93	23,58	0,1700	1,02
7	S610321720		WKD 1-V MET R.A. INBOUW	1,00	STK	3,43	3,43	0,1700	0,17
8	S610323720		WKD 2-V MET R.A. INBOUW	105,00	STK	6,86	720,30	0,3400	35,70
9	S610323722		WKD 2-V MET R.A. (WG) INBOUW	18,00	STK	6,86	123,48	0,3400	6,12
10	S681600000		AFDEKPLAAT 1-VOUDIG	1,00	STK	1,48	1,48	0,0000	0,00
11	S681700000		AFDEKPLAAT 2-VOUDIG	110,00	STK	2,53	278,30	0,0000	0,00
12	S481401000		KABELDOOS + 2-V WKD + RA	9,00	STK	12,19	109,71	0,5700	5,13
13	S481500031		SNELMONTAGEPLAAT P31-GOOT	9,00	STK	3,61	32,49	0,0300	0,27
14	9999999900		VERPLAATSING WCD, IDENTIEKE SITUATIE	12,00	STK	0,00	0,00	0,0000	0,00
15	9999999900		VERVALLEN WCD, REEDS EERDER VERREKEND	93,00	STK	0,00	0,00	0,0000	0,00
16	0000141610	LEGRAND	Energiezuil MZ-6+ L2880 aluminium R9010	2,00	STK	374,02	748,04	2,0000	4,00
17	9900000050		KLEIN MATERIAAL	1,00	POS	50,00	50,00	0,0000	0,00
18									
19			** KABEL- EN BUISAANLEG **						
20	S810305010		HULT CCA 3X2,5 AANSL.KAST EXCL.WARTEL	26,00	STK	2,00	52,00	0,3800	9,88
21	S810305111		HULT CCA 3X2,5 VERTIKAAL BAAN BANDJES	52,00	MTR	2,29	119,08	0,1200	6,24
22	S810305120		HULT CCA 3X2,5 IN GOOT	1425,00	MTR	2,06	2935,50	0,0400	57,00
23	S810305140		HULT CCA 3X2,5 IN WANDGOOT	24,00	MTR	2,06	49,44	0,0300	0,72
24	S810305210		HULT CCA 3X2,5 IN BUIS	585,00	MTR	2,06	1205,10	0,0600	35,10
25	S551800420		HALOVOLT 3/4 ZICHT ENKELVOUDIG	393,00	MTR	3,67	1442,31	0,2600	102,18
26	S551800720		HALOVOLT 3/4 IN (METSEL)WAND	178,00	MTR	2,92	519,76	0,1900	33,82
27	S481400000		KABELDOOS HAF 3640	87,00	STK	3,96	344,52	0,5700	49,59

Figuur 2 – PDF-voorbeeld van een pagina van een open begroting.

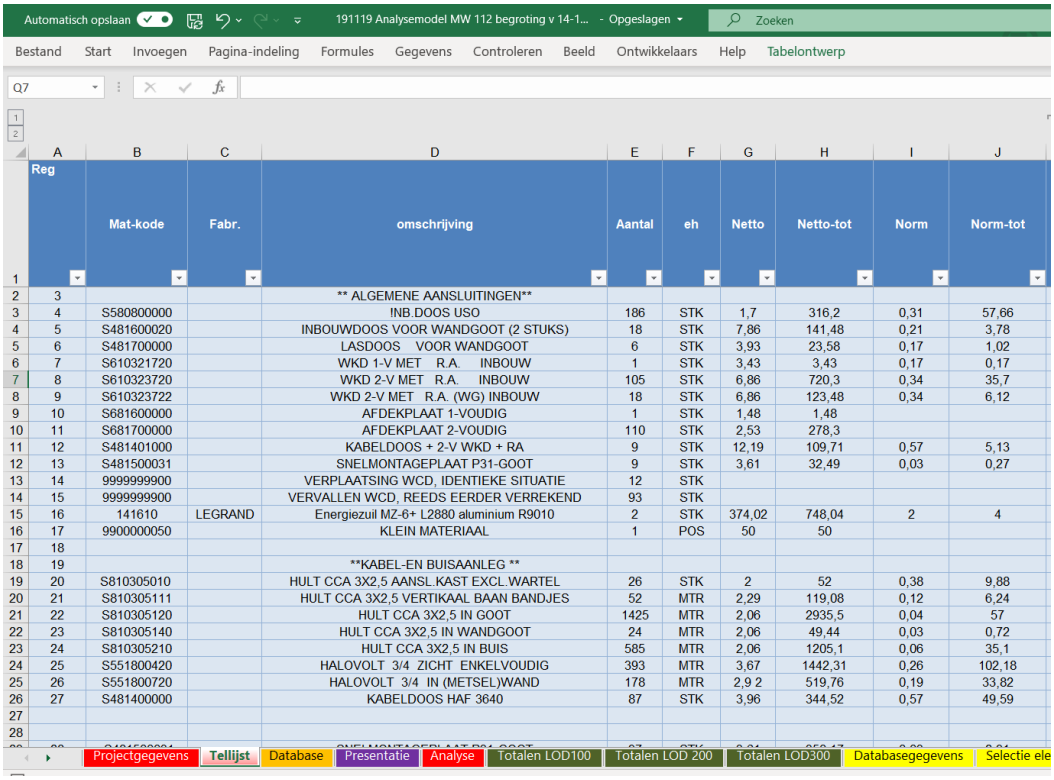
De tool waarmee K&R werkt om de analyses uit te voeren is in Excel. De PDF-bestanden dienen daarom te worden geconverteerd naar Excel, wat vaak leidt tot verschillende opmaak problemen. Het komt regelmatig voor dat het converteer-programma bepaalde rijen als één rij ziet, waardoor de informatie van meerdere rijen in een rij terecht komt. Dit zorgt voor problemen bij het toewijzen van eigen labels aan elke kostenpost van de open begroting. Een voorbeeld van een geconverteerde pagina van een open begroting kan gevonden worden in appendix 7.2 (pagina 39). Bij het labelen wordt er systematisch per rij aangegeven welk K&R label bij een kostenpost hoort, maar als rijen per ongeluk zijn samengevoegd, kan dit leiden tot problemen. Er kan over belangrijke informatie heen gelezen worden of kosten van verschillende posten worden samengevoegd terwijl die posten niet bij elkaar horen. Het verzamelen van deze gegevens en omzetten naar Excel is niet de fase die de meeste tijd kost, maar het is wel frustrerend werk. Opmaak problemen zijn niet de soort problemen waar de hoogopgeleide werknemers van K&R hun tijd aan willen besteden. Desondanks is het wel erg belangrijk dat hier geen fouten worden gemaakt. Fouten die in deze fase worden gemaakt kunnen de hele analyse nutteloos maken.

2.2 Invullen tool

Om uit te leggen hoe de tool werkt, gebruiken we de screenshot in Figuur 3. Met dit Excel-bestand, waarin ook de database met haar prijzen zijn verwerkt, kunnen de werknemers van K&R een groot deel van hun analyse uitvoeren. Het is belangrijk voor de lezer om dit bestand goed te begrijpen. De rest van deze scriptie is namelijk gebaseerd op hoe K&R hun analyse in deze tool uitvoert.

In Figuur 2 uit sectie 2.1 is een voorbeeld te zien van de opbouw van een open begroting. De kolomnamen worden overgenomen in de tool van Figuur 3 in de eerst kolommen. In dit deel van de tool staan dus de gegevens uit de open begrotingen zoals die zijn aangeleverd. Per kolom dient er maar een soort informatie in te staan, zoals 'aantal', 'omschrijving' of 'netto'. Dit maakt het later mogelijk om prijzen en aantallen makkelijk op te tellen met Excel. Hier komt dus ook al naar voren waarom het belangrijk is dat bij fase 1 alle informatie in de juiste rijen blijft staan. Als dit niet gebeurt zullen de totalen in het Excel-bestand ook niet kloppen.

Als de lezer geïnteresseerd is in hoe een nog niet ingevulde tool eruitziet, is dit terug te vinden in appendix 7.3 (pagina 40). Na het invullen van de tool, worden er nog een aantal gegevens in de tool aangepast die voor het proces waar deze scriptie over gaat niet essentieel zijn. Deze fase is terug te vinden in Appendix 7.4 (pagina 41).



Reg	Mat-kode	Fabr.	omschrijving	Aantal	eh	Netto	Netto-tot	Norm	Norm-tot
3			** ALGEMENE AANSLUITINGEN**						
4	S580800000		!NB.DOOS USO	186	STK	1,7	316,2	0,31	57,66
5	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18	STK	7,86	141,48	0,21	3,78
6	S481700000		LASDOOS VOOR WANDGOOT	6	STK	3,93	23,58	0,17	1,02
7	S610321720		WKD 1-V MET R.A. INBOUW	1	STK	3,43	3,43	0,17	0,17
8	S610323720		WKD 2-V MET R.A. INBOUW	105	STK	6,86	720,3	0,34	35,7
9	S610323722		WKD 2-V MET R.A. (WG) INBOUW	18	STK	6,86	123,48	0,34	6,12
10	S681600000		AFDEKPLAAT 1-VOUDIG	1	STK	1,48	1,48		
11	S681700000		AFDEKPLAAT 2-VOUDIG	110	STK	2,53	278,3		
12	S481401000		KABELDOOS + 2-V WKD + RA	9	STK	12,19	109,71	0,57	5,13
13	S481500031		SNELMONTAGEPLAAT P31-GOOT	9	STK	3,61	32,49	0,03	0,27
14	9999999900		VERPLAATSING WCD, IDENTIEKE SITUATIE	12	STK				
15	9999999900		VERVALLEN WCD, REEDS EERDER VERREKEND	93	STK				
16	141610	LEGRAND	Energiezuil MZ-6+ L2880 aluminium R9010	2	STK	374,02	748,04	2	4
17	9900000050		KLEIN MATERIAAL	1	POS	50	50		
18									
19			**KABEL-EN BUISAANLEG **						
20	S810305010		HULT CCA 3X2,5 AANSL.KAST EXCL.WARTEL	26	STK	2	52	0,38	9,88
21	S810305111		HULT CCA 3X2,5 VERTIKAAL BAAN BANDJES	52	MTR	2,29	119,08	0,12	6,24
22	S810305120		HULT CCA 3X2,5 IN GOOT	1425	MTR	2,06	2935,5	0,04	57
23	S810305140		HULT CCA 3X2,5 IN WANDGOOT	24	MTR	2,06	49,44	0,03	0,72
24	S810305210		HULT CCA 3X2,5 IN BUIS	585	MTR	2,06	1205,1	0,06	35,1
25	S551800420		HALOVOLT 3/4 ZICHT ENKELVOUDIG	393	MTR	3,67	1442,31	0,26	102,18
26	S551800720		HALOVOLT 3/4 IN (METSEL)WAND	178	MTR	2,9 2	519,76	0,19	33,82
27	S481400000		KABELDOOS HAF 3640	87	STK	3,96	344,52	0,57	49,59

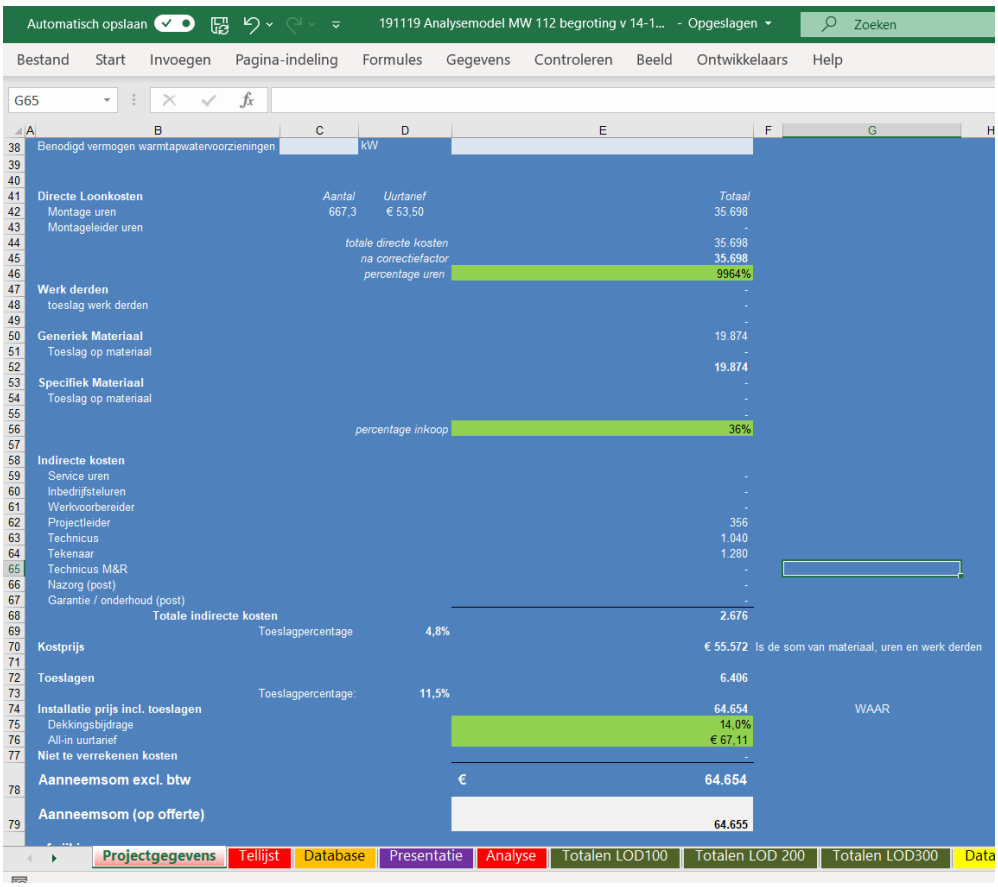
Figuur 3 - Voorbeeld van Excel-tool, tabblad Tellijst

2.3 Indirecte kosten

Elke open begroting heeft een totaalblad waarop de totale kosten, uren en toeslagen te vinden zijn. Voor het verwerken van de gegevens van dit totaalblad heeft K&R een eigen tabblad gemaakt binnen hun tool. Dit is het tabblad 'Projectgegevens' dat te vinden is in Figuur 4. In appendix 7.5 (pagina 42) is dit tabblad ook te vinden zonder ingevulde informatie. Het tabblad 'Projectgegevens' moet ook ingevuld worden, zodat alle kosten worden meegenomen. Daarnaast heeft het een functie om te controleren of alles in de vorige fases wel correct is gegaan. Hoe dit in zijn werk gaat is in de volgende alinea beschreven.

Op het totaalblad worden de uren van projectleider, tekenaar en andere specialisten vaak speciaal vermeld. Deze personen hebben een ander uurloon dan de werknemers die alles monteren en moeten daarom apart vermeld worden. Daarnaast worden er ook vaak toeslagen doorberekend op ‘materiaal generiek’, ‘materiaal specifiek’ en ‘derden’. Ook zijn toeslagen op ‘winst en risico’, ‘algemene kosten’ en ‘transportkosten’ niet ongebruikelijk. Deze toeslagen worden door K&R allemaal verwerkt in het tabblad ‘Projectgegevens’. Een voorbeeld van een ingevuld tabblad kan gevonden worden in Figuur 4. Een deel van dit tabblad moeten handmatig overgenomen worden vanuit de open begrotingen. De rest wordt automatisch opgeteld door wat er al is ingevuld in het andere tabblad ‘Tellijst’ in de eerdere fases van het analyseproces. Een voorbeeld van het tabblad ‘Tellijst’ is te vinden in Figuur 3.

Figuur 4 toont dat onderaan het tabblad de ‘aanneemsom (op offerte)’ wordt ingevuld. Deze wordt in elke begroting vermeld. De tool is zo gebouwd dat in het kopje daarboven de aanneemsom wordt berekend aan de hand van alle ingevulde gegevens in de tabbladen ‘Projectgegevens’ en ‘Tellijst’. Als alles goed is overgenomen en ingevuld, zal er nauwelijks tot geen verschil tussen de aanneemsommen zitten. Is er wel verschil, dan is er iets fout gegaan bij het overnemen van de gegevens en kan die fout nu nog gevonden worden. Ook is er in kolom ‘G’ in het tabblad ‘Projectgegevens’ een formule gebouwd die ‘WAAR’ of ‘ONWAAR’ teruggeeft. Deze formule checkt of er in de in sectie 2.2 genoemde kolommen alleen maar getallen staan en geen foutwaardes. Zijn er geen foutwaardes, geeft de formule ‘WAAR’ terug en weet de K&R-werknemer dat er waarschijnlijk niks fout gaat binnen de tool. Zo dient het tabblad ‘Projectgegevens’ ook als methode om te controleren of alles is goed gegaan bij het invullen van de tool.



Figuur 4 - Ingevuld 'Projectgegevens' tabblad uit de Excel-tool

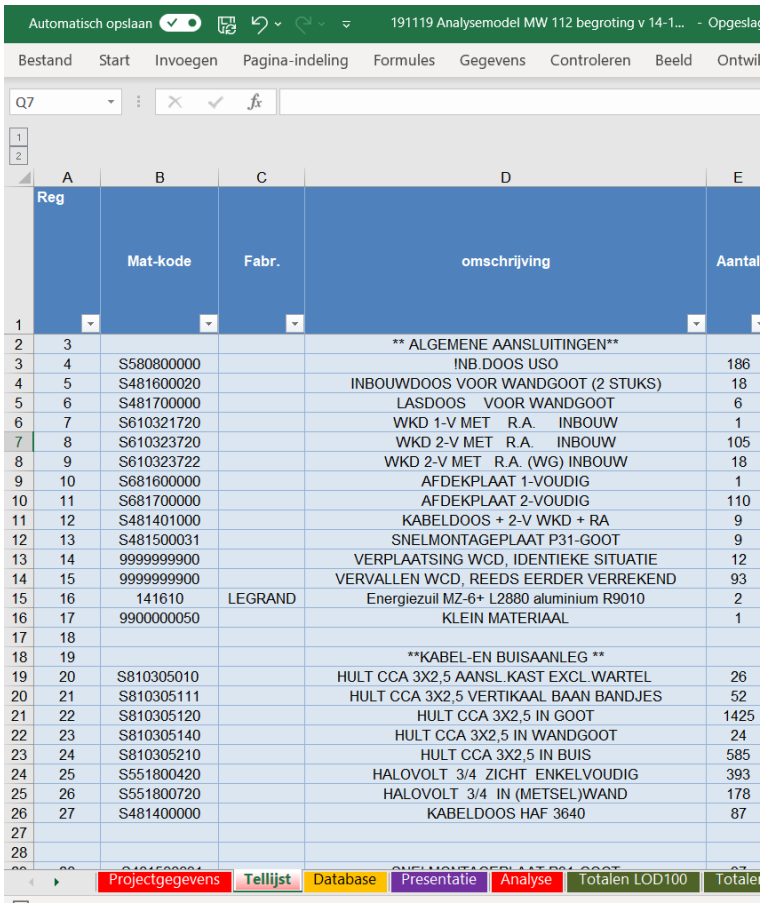
2.4 Labelen

De volgende fase van het analyseproces is het labelen van alle kostenposten van een open begroting van een aannemer. De gegevens die zijn aangeleverd door de aannemers worden dus in de eerste kolommen van de tool geplaatst. Het toewijzen van K&R-labels wordt gedaan in de kolommen 'hoofdcode', 'Kolom2' en 'Kolom3' (respectievelijk kolom 'R', 'T' en 'U'). In Figuur 6 is te zien hoe dat eruitziet. Alle elementen in de database van K&R zijn opgedeeld in codes, waarbij elke code staat voor een apart element. De eerste twee cijfers van een code geven aan bij welke hoofdcode het hoort. Voorbeelden van hoofdcodes zijn 'afvoeren', 'luchtbehandeling' en 'krachtstroom'. Binnen deze hoofdcodes komen dus logischerwijs alle elementen te maken met die hoofdcode voor.

Elke hoofdcode kan weer onderverdeeld worden in verschillende element varianten. Dit zijn de namen die te vinden zijn in de zogenaamde 'Kolom2' in het voorbeeld van Figuur 6. Als er gekeken wordt naar de hoofdcode van 'afvoeren', zijn daar bijvoorbeeld 'HWA buiten appendages' en 'VWA leidingwerk' te vinden. Binnen deze element variant zijn er weer andere elementen te vinden. Deze elementen zijn terug te vinden in 'Kolom3' van Figuur 6. Binnen 'VWA leidingwerk' zijn bijvoorbeeld de volgende elementen te vinden: 'Afvoerleiding PE 40 mm' en 'Vuilwater afvoer PVC'.

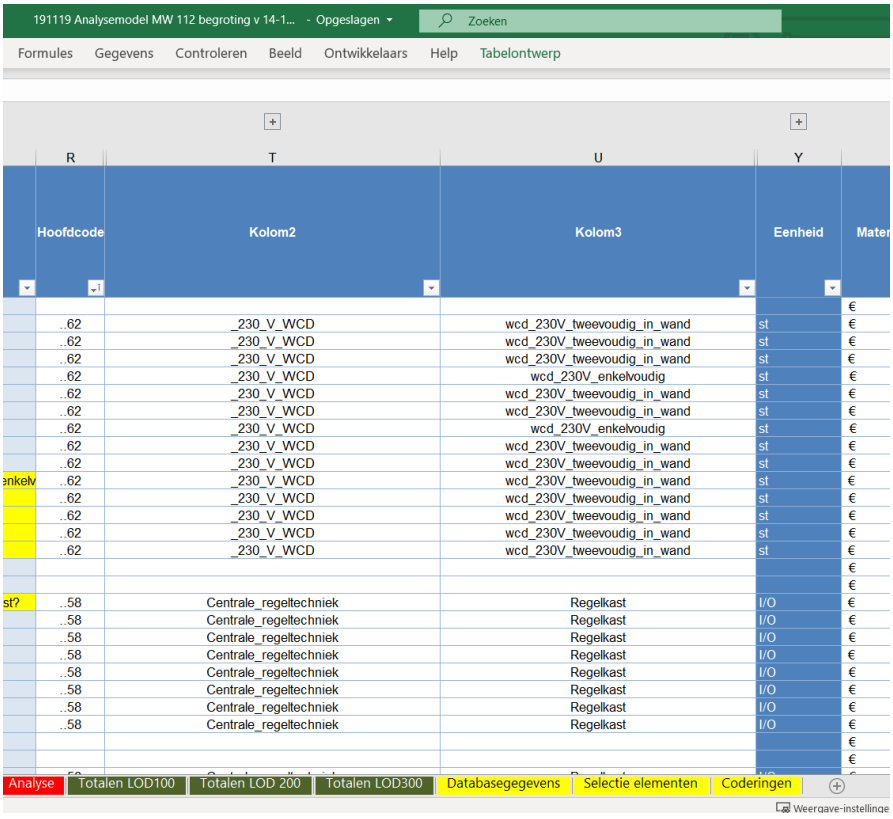
Elk element heeft een unieke code van zes cijfers, waarbij de eerste twee cijfers slaan op de hoofdcode, de middelste twee op het element variant en de laatste twee op het unieke element. Wanneer werknemers van K&R bezig gaan met labelen, werken ze elke rij af en schrijven ze per rij een hoofdcode, een element variant en een element op die overeenkomt met de materialen zoals de aannemer het heeft beschreven in hun open begroting. Deze omschrijving komt bijna nooit letterlijk overeen, waardoor K&R per omschrijving goed moet kijken wat er precies instaat en waar de nuttige informatie voor hen is verstopt. Aan de hand van het analyseren van die omschrijvingen en hun kennis van de database, kunnen zij een label geven aan een kostenpost. Wanneer zij dat label hebben gegeven, vult de tool automatisch de bijbehorende code in. Uiteindelijk worden de kostenposten met dezelfde code bij elkaar opgeteld, waardoor er een totaalbedrag ontstaat voor een bepaald soort bouwonderdeel.

Figuur 5 en Figuur 6 tonen screenshots van de tool waarin de elementen al gelabeld zijn. Figuur 5 laat de eerste kolommen (kolommen A t/m E) van de tool zien, waarin de omschrijving zoals de aannemers het hebben opgesteld te zien zijn. Figuur 6 toont de labels die verder naar rechts in de tool te vinden zijn (kolommen R t/m Y). Om een overzichtelijk idee te krijgen van hoe de elementen worden overgeschreven, zal aan de hand van Figuur 5 en Figuur 6 uitgelegd worden waarom bepaalde kostenposten bepaalde labels krijgen. Een van de belangrijkste kolommen bij het labelen is kolom 'D' in Figuur 5, oftewel de kolom waar de omschrijving is te vinden. In open begrotingen is dat vaak een van de eerste kolommen. Verder zijn de kolommen 'R', 'T' en 'U' in Figuur 6 ook belangrijk, want hier zijn de labels beschreven zoals K&R die heeft staan in hun database. Als we kijken naar regel 6, staat hier in kolom 'D' de omschrijving 'WKD 1-V MET R.A. INBOUW' wat staat voor een enkelvoudige wandcontactdoos. Deze omschrijving krijgt het K&R label 'wcd_230V_enkelvoudig' toegewezen. In regel 7 staat er 'WKD 2-V MET R.A. INBOUW' wat staat voor een tweevoudige wandcontactdoos. Zoals in Figuur 6 te zien is, wordt dit dan gelabeld als 'wcd_230V_tweevoudig_in_wand'. Alle onderdelen van een wandcontactdoos, zoals de bedrading en afdekplaatjes, worden in de regels ervoor en erna benoemd. Al deze onderdelen vallen bij K&R onder dezelfde label, zodat er één kostprijs ontstaat voor de wandcontactdozen. In Figuur 6 is te zien hoe de tool er dan uit komt te zien, waarbij de meeste regels dezelfde label krijgen toegewezen.



	A	B	C	D	E
1	Reg	Mat-kode	Fabr.	omschrijving	Aantal
2	3			** ALGEMENE AANSLUITINGEN**	
3	4	S580800000		!NB.DOOS USO	186
4	5	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18
5	6	S481700000		LASDOOS VOOR WANDGOOT	6
6	7	S610321720		WKD 1-V MET R.A. INBOUW	1
7	8	S610323720		WKD 2-V MET R.A. INBOUW	105
8	9	S610323722		WKD 2-V MET R.A. (WG) INBOUW	18
9	10	S681600000		AFDEKPLAAT 1-VOUDIG	1
10	11	S681700000		AFDEKPLAAT 2-VOUDIG	110
11	12	S481401000		KABELDOOS + 2-V WKD + RA	9
12	13	S481500031		SNELMONTAGEPLAAT P31-GOOT	9
13	14	9999999900		VERPLAATSING WCD, IDENTIEKE SITUATIE	12
14	15	9999999900		VERVALLEN WCD, REEDS EERDER VERREKEND	93
15	16	141610	LEGRAND	Energiezuil MZ-6+ L2880 aluminium R9010	2
16	17	9900000050		KLEIN MATERIAAL	1
17	18				
18	19			**KABEL-EN BUISAANLEG **	
19	20	S810305010		HULT CCA 3X2,5 AANSL.KAST EXCL.WARTEL	26
20	21	S810305111		HULT CCA 3X2,5 VERTIKAAL BAAN BANDJES	52
21	22	S810305120		HULT CCA 3X2,5 IN GOOT	1425
22	23	S810305140		HULT CCA 3X2,5 IN WANDGOOT	24
23	24	S810305210		HULT CCA 3X2,5 IN BUIS	585
24	25	S551800420		HALOVOLT 3/4 ZICHT ENKELVOUDIG	393
25	26	S551800720		HALOVOLT 3/4 IN (METSEL)WAND	178
26	27	S481400000		KABELDOOS HAF 3640	87
27					
28					

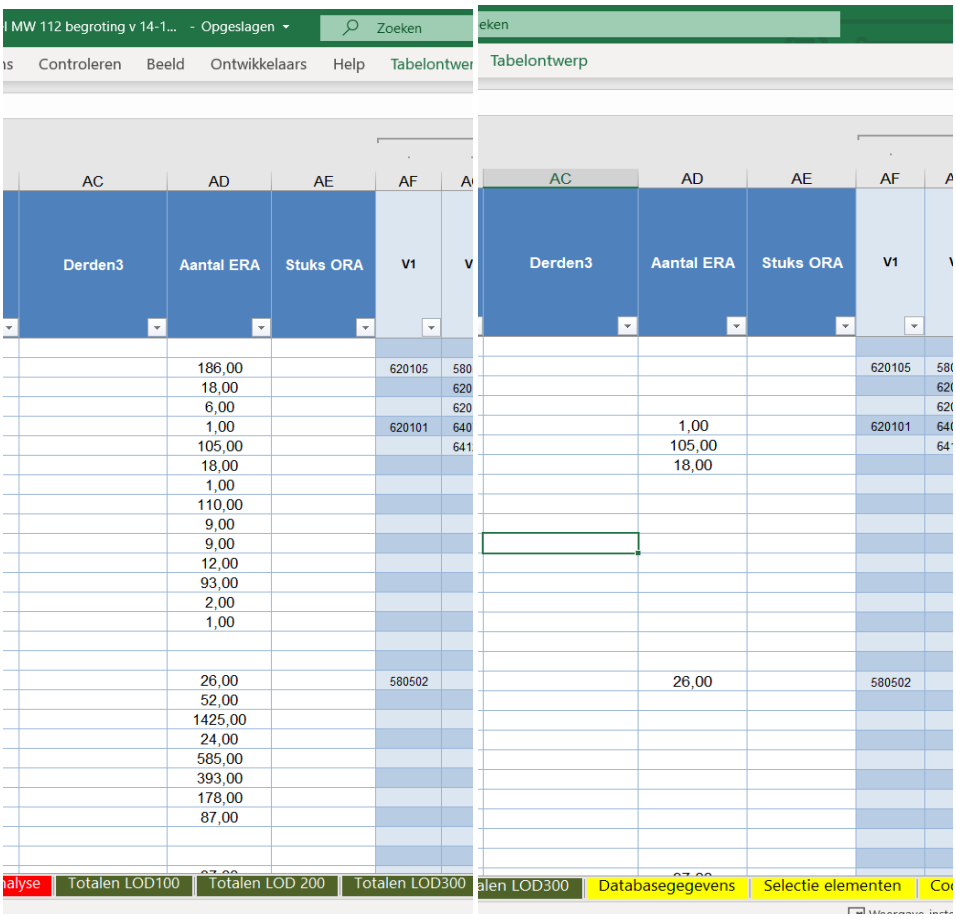
Figuur 5 - Kolommen 'A' t/m 'E' van het tabblad 'Tellijst' uit de Excel-tool



	R	T	U	Y	
	Hoofdcode	Kolom2	Kolom3	Eenheid	Mater
					€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_enkelvoudig	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_enkelvoudig	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
enkelv	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
	..62	_230_V_WCD	wcd_230V_tweevoudig_in_wand	st	€
					€
					€
st?	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
	..58	Centrale_regeltechniek	Regelkast	I/O	€
					€
					€

Figuur 6 - Kolommen 'R' t/m 'U' van het tabblad 'Tellijst' uit de Excel-tool

Wanneer alle regels zijn gelabeld, blijken er vaak meerdere regels bij één kostenpost te horen. Daardoor moeten de aantallen zoals die zijn overgenomen van de aannemer vaak worden aangepast. Voor de wandcontactdozen maakt de aannemer bijvoorbeeld onderscheid tussen het contact zelf en de afdekplaat. Daardoor staat ervoor de ene regel in dat er bijvoorbeeld 105 contacten nodig zijn en 110 afdekplaatjes, terwijl voor K&R dit allemaal valt onder dezelfde 105 wandcontactdozen. Om ervoor te zorgen dat niet elke post wordt gezien als nieuwe wandcontactdozen, moeten de aantallen worden aangepast. Dit is zichtbaar in Figuur 7, waarbij links eerst de oude situatie met het letterlijk overnemen van de aantallen te zien is en rechts de situatie aangepast door K&R te zien is. De aantallen van de kostdrivers (wandcontactdozen) blijft behouden, terwijl de bijbehorende elementen niet als aantallen worden meegenomen. Dit is meegenomen in de prijs van het label wandcontactdozen. Het aanpassen van deze aantallen kost tijd en kan ook snel resulteren in fouten. Deze actie is nodig omdat anders de tool elk onderdeel gaat optellen als een volledige wandcontactdoos, waardoor de analyse bij het prijzen vergeleken geen correct beeld geeft.



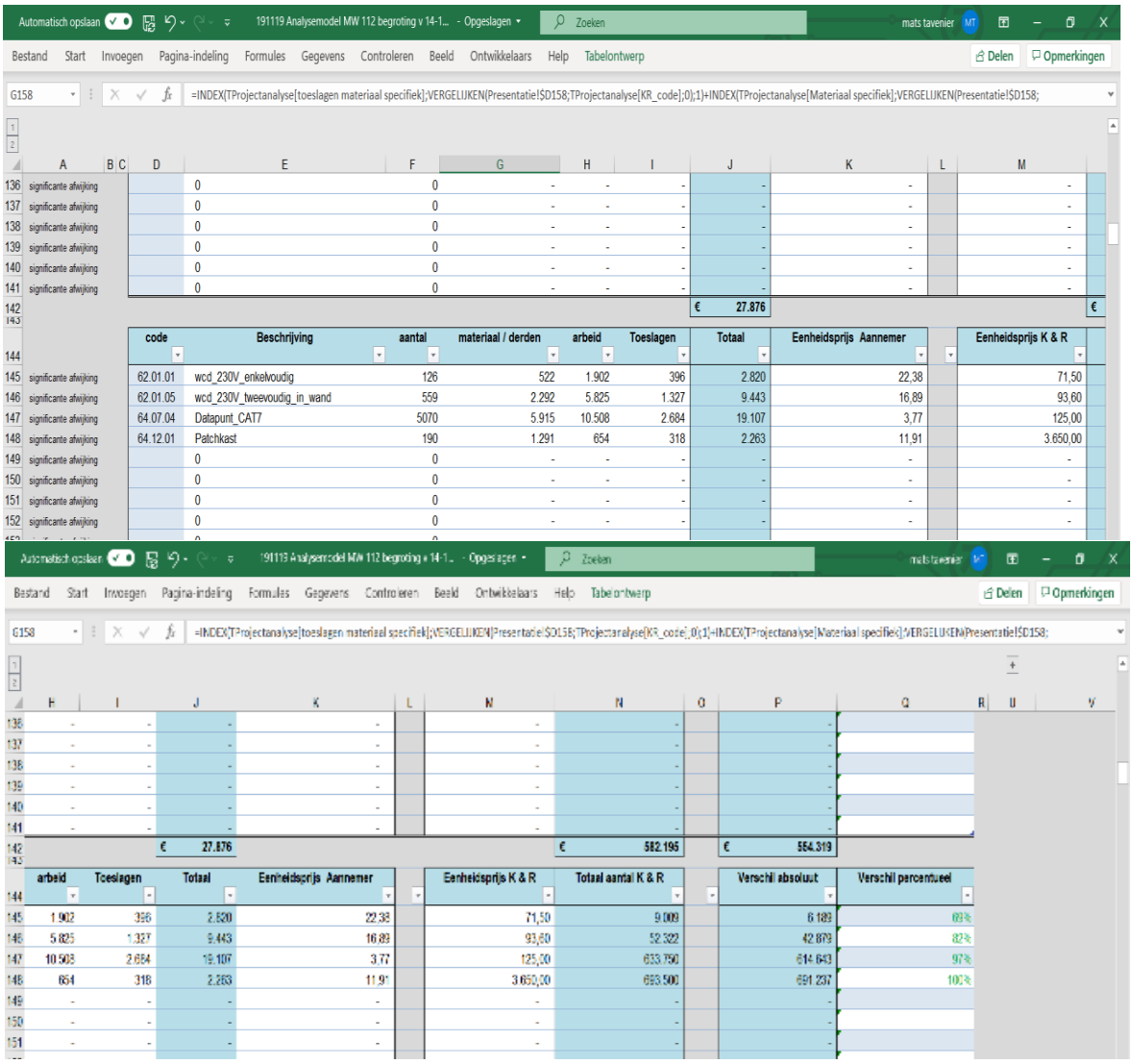
Figuur 7 -Voorbeeld van aanpassen van aantallen in de Excel-tool

2.5 Prijzen vergelijken

Nadat elke regel is gelabeld en dus elke kostenpost van de aannemer een label heeft, kunnen de prijzen van de aannemer worden vergeleken met de prijzen uit de database. Dit wordt gedaan in het tabblad 'Presentatie', waarin elke label terugkomt die in het tabblad 'Tellijst' is ingevuld. Figuur 8 toont een screenshot van het tabblad 'Presentatie'. Achter elke label staat onder andere het aantal, de toeslagen, het totaal en de uiteindelijk eenheidsprijs. Al deze getallen worden automatisch uitgerekend op basis van de informatie ingevuld in de tabblad 'Tellijst' en 'Projectgegevens'. In de kolommen daarnaast staan de eenheidsprijs en totaalprijs wat volgens K&R's database

marktconform is en in de meest rechtse kolommen staan het ‘verschil absoluut’ en ‘verschil percentueel’ tussen de database prijzen en de aannemers prijzen. Daardoor kan er in een oogopslag worden gezien waar de verschillen zitten en hoe groot die zijn.

De onderdelen waarvan het percentuele prijsverschil meer dan ongeveer 30% is, worden vaak geaccentueerd met een kleurtje en/of opmerkingen. Dit zijn de onderdelen waarop K&R moet gaan bespreken met de aannemer over hoe dat verschil zo groot kan zijn. Soms worden er in deze fase ook database prijzen aangepast omdat de werknemers niet altijd geloven dat de database de juiste, markconforme prijs teruggeeft. Hier komt dan de expertise, ervaring en kennis van de werknemers van K&R bij kijken om met de hulp van deze tool tot een conclusie te komen over de kosten opgegeven in een open begroting.



Figuur 8 - Voorbeeld van tabblad 'Presentatie' uit de Excel-tool. De onderste screenshot hoort rechts van de eerste screenshot door te lopen, maar vanwege leesbaarheid is de screenshot er onder geplakt.

2.6 Conclusie huidige situatie

In hoofdstuk 2 is een overzicht gemaakt van de verschillende fases bij het analyseproces van K&R. Per sectie is er een fase besproken en er hieruit kunnen de volgende conclusies getrokken worden.

In fase 1, waarin de gegevens worden verkregen en gedigitaliseerd, ontstaan veel opmaak problemen die als frustrerend kunnen worden ervaren. Fase 2 het invullen van de tool is na het goed uitvoeren van fase 1 geen lastige opgave meer. Tijdens fase 3 ‘indirecte kosten’ moeten er veel gegevens worden overgetypt, maar er wordt ook al veel automatisch opgeteld waardoor hier niet veel tijd wordt besteed. Fase 4 het labelen van alle kostenposten neemt het meeste tijd in beslag. Het per regel afgaan welk label erbij hoort is een tijdsintensieve en repetitieve bezigheid. In de laatste fase worden de prijzen vergeleken, waarbij vooral met de expertise van de K&R-werknemers de gegevens geanalyseerd moeten worden.

3 Literatuurstudie

In dit hoofdstuk wordt er ingegaan op de verschillende mogelijkheden in de literatuur beschikbaar voor K&R Consultants voor het verbeteren van het analyse proces. In het vorige hoofdstuk is beschreven waar de grootste problemen te vinden zijn. Met de literatuur uit dit hoofdstuk zal in theorie de problemen omtrent het labelen van kostenposten aangepakt kunnen worden.

3.1 Text mining

Zoals besproken in hoofdstuk 2, neemt het labelen in fase 4 van het analyseproces de meeste tijd in beslag. De omschrijving van materialen door aannemers moeten worden vertaald naar de omschrijving die K&R gebruikt. De verschillende omschrijvingen komen regelmatig deels met elkaar overeen, waardoor er een repetitief aspect in het werk verschijnt omdat bepaalde materialen in elk project voorkomen. Wanneer een label al eens eerder is toegewezen aan een omschrijving, zou in een nieuw project op basis van tekst herkenning automatisch dat zelfde label als suggestie gegeven moeten worden.

De theoretisch velden die voor K&R's situatie interessant zijn, zijn 'Text Mining', 'Natural Language Processing (NLP)' en 'Machine Learning (ML)'.

'Text Mining' wordt door Kao en Poteet (1) beschreven als 'de ontdekking en extractie van interessante, niet-triviale kennis uit vrije of ongestructureerde tekst'. Als voorbeeld van 'Text Mining' wordt tekst classificatie genoemd, wat specifiek betrekking heeft op deze scriptie. Uiteindelijk is het doel om de omschrijvingen van aannemers te classificeren met de labels van K&R.

Tekst classificatie wordt door Jo (2) uitgelegd op een manier waarbij het direct te linken is aan de situatie van K&R. In hoofdstuk 1.3.1 van het boek 'Text Mining' van Jo worden de voorafgaande activiteiten bij classificatie beschreven als het maken van een lijst van categorieën, wat dan weer een classificatiesysteem wordt genoemd. Hierna worden data items toegewezen aan elke categorie als test data. In principe is deze data bij K&R al bekend, aangezien de analyses van de afgelopen jaren gebruikt kunnen worden als classificatiesysteem. Hierna wordt er een test collectie gemaakt, wat bestaat uit al gelabelde data en wordt onderverdeeld in een training set en test set. De training set wordt gebruikt als voorbeeldgegevens, waar het algoritme van kan leren. Daarna wordt de test set, waarvan de labels al bekend zijn, door het algoritme ook geclassificeerd. Van de test set kunnen de verschillen tussen het werkelijke en toegewezen label worden geïdentificeerd. De verhouding van juist gelabelde items ten opzichte van het totale aantal items, geven een indicatie van de nauwkeurigheid van het classificatiesysteem.

Een term sterk gerelateerd aan 'Text Mining' is 'Natural Language Processing (NLP)'. NLP is de poging om een volledigere betekenis af te leiden uit vrije tekst. (1) Hierbij ligt de focus vooral op het verwerken van grotere stukken tekst waarin volledige zinnen te vinden zijn. Er wordt rekening gehouden met onder andere grammatica, zinsopbouw en synoniemen. De tekst bij K&R wordt over het algemeen minder vrij aangeleverd, aangezien er een relatief korte omschrijving wordt gegeven. Wel zijn er NLP-technieken die ook op het niveau van zinnen een classificatie kunnen geven. In hoofdstuk 4.1 van 'Natural Language Processing and Text Mining' (1) wordt ingegaan op zin- en alineaniveau om informatie te ontdekken. Deze technieken kunnen dus ook interessant zijn voor de situatie van K&R.

3.2 Machine learning

Zoals besproken in de vorige sectie, heeft deze thesis betrekking op onder andere 'Text Mining' en tekst classificatie. Door Sharda et al. (3) wordt in het boek 'Business Intelligence' er gesproken van

twee voornaamste manieren om tekst te classificeren. Deze zijn ‘Knowledge Engineering’ en ‘Machine Learning’ (ML). (4) Dit boek van Sharda wordt in de derde module van de bachelor Technische Bedrijfskunde behandeld.

Bij de ‘Knowledge Engineering’ aanpak wordt de kennis van de expert over de categorieën in het systeem verwerkt. De ML-aanpak daarentegen bouwt een classificatie systeem dat leert van eerder geclassificeerde tekst. Hierbij is het dus nodig om een al getrainde dataset te hebben waarbij de klassen of labels al zijn toegewezen. Een belangrijk voordeel van deze aanpak is dat het systeem zich blijft ontwikkelen naarmate er meer data wordt ingevoerd. Dit betekent dat hoe meer data er bekend is, hoe beter het systeem wordt.

Ook in andere literatuur wordt er gesproken over verschillende manieren van tekst analyses, waarbij ook tekst classificatie terug komt. Zo heeft Aggarwal (5) het in zijn boek ‘Machine Learning for text’ over het duidelijke verschil tussen tekst classificatie en tekst clustering. Bij beide technieken wordt de data in groepen verdeeld, maar bij tekst classificatie heeft het algoritme veel meer begeleiding. Bij tekst clustering is er niet een vooraf getrainde dataset en zijn de categorieën dus van tevoren ook nog niet bekend. Dit wordt dan ook wel ‘unsupervised learning’ genoemd. Bij tekst classificatie wordt het algoritme veel meer in de juiste richting gestuurd door de al bekende datasets en categorieën. Aggarwal (5) noemt daarom tekst classificatie ook wel ‘supervised learning’ en dit is ook de techniek die in dit onderzoek zal worden gebruikt.

Binnen dit ‘supervised learning’ zijn meerdere algoritmes, onder andere ‘rule-based’, ‘nearest neighbor’, ‘linear’ en ‘Naive Bayes classifiers’. In de volgende secties zijn een aantal algoritmes kort behandeld en wordt er dieper ingegaan op de ‘Naive Bayes’ theorie. Voor verdere uitleg over deze algoritmes, kan ook Fanny et al. (6) worden geraadpleegd.

3.3 Algoritmes in tekst classificatie

Er zijn meerdere bekende algoritmes in tekst classificatie, die door Aggarwal (5) duidelijk zijn uitgelegd. Een aantal van deze algoritmes worden hier ook kort behandeld.

Het eerste algoritme is ‘rule-based’ classificaties, waarbij er in een ‘rule’ bepaalde kenmerken van subsets worden gerelateerd aan een specifiek label. Zo’n ‘rule’ bestaat uit die voorwaarden en wanneer de aanwezigheid van een subset van woorden aan die voorwaarden voldoen, wordt er een label voorspeld. Een voorbeeld van een subset van de volgende materiaalomschrijving ‘Fatini inbouwdeel badkraan’ zou zijn ‘badkraan’. Aan de hand van een vooropgestelde regel kan het algoritme die materiaalomschrijving toewijzen aan het label ‘Bad’.

De tweede theorie die wordt besproken, is de ‘Naive Bayes (NB)’ theorie. De ‘Naive Bayes’ theorie is een van de meest simpele classificatiesystemen, maar wordt wel veel gebruikt. Het simpele schuilt in het naïeve van de theorie, omdat er wordt uitgegaan van volledige onafhankelijkheid tussen elk woord, ook wel ‘feature’ genoemd. Ondanks dat deze methode relatief simpel is, is het wel een erg effectieve methode. (7)

In de volgende secties wordt de NB-theorie uitgebreider uitgelegd en wordt besproken hoe deze theorie nuttig kan zijn voor K&R.

3.4 Naive Bayes

De verschillende termen die in de vorige secties zijn besproken, zijn redelijke algemene termen waarmee K&R in de praktijk niet verder komt. In deze sectie zal de ‘Naive Bayes’ (NB) theorie worden uitgelegd en hoe die theorie van waarde kan zijn voor K&R.

De NB-theorie bestaat uit twee delen, namelijk de Bayes theorie en het naïeve onderdeel. In dit artikel door Zhang (8) wordt de theorie op een duidelijk manier uitgelegd. Ook is deze manier van uitleggen gericht op toepassingen in de programmeertaal 'R', wat ook helpt bij het werken richting een werkend concept in een andere programmeertaal. Hier wordt later verder op ingegaan. De Bayes theorie kan worden gebruikt om een voorspelling te maken op basis van statistische voorkennis en huidige data. Met het vergaren van meer data, verandert de voorspelling. De voorkennis is de meest waarschijnlijke schatting van de uitkomst zonder aanvullend bewijs. De weergave hiervan wordt ook wel 'prior probability' genoemd. Het huidige bewijs wordt uitgedrukt als de waarschijnlijkheid dat de kans weerspiegelt dat een voorspeller een bepaalde uitkomst krijgt.

De Bayes theorie kan op de volgende manier als vergelijking worden opgeschreven:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Vergelijking 1 - Bayes theorie

Hierbij zijn *P(A)* en *P(B)* de kansen dat gebeurtenissen A en B plaatsvindt zonder met elkaar rekening te houden. *P(A/B)* is de voorwaardelijk kans dat gebeurtenis A plaatsvindt, gegeven dat gebeurtenis B plaatsvindt en *P(B/A)* is de voorwaardelijke kans dat gebeurtenis B plaatsvindt, gegeven dat gebeurtenis A plaatsvindt. Het naïeve van deze theorie is dat ervan uit wordt gegaan dat de 'features' van alle voorspellers onafhankelijk van elkaar zijn. Elke 'feature' is een woord binnen een zogenaamde materiaalomschrijving. Door de onafhankelijkheid van de 'features' worden de berekeningen veel makkelijker, omdat $P(B|A) = P(b1, b2, b3|A)$ als $P(b1|A) \times P(b2|A) \times P(b3|A)$ kan worden geschreven. De 'features' zijn dan geschreven in de formule als $b1$, $b2$ en $b3$. In het geval van tekst classificatie betekent dit dat voor elk woord onafhankelijk de kans wordt berekend dat dit woord in een bepaalde klasse/label hoort.

In praktijk zijn de woorden wel afhankelijk van elkaar, maar deze afhankelijkheid uitdrukken in een formule of algoritme is erg lastig. Door te veronderstellen dat de woorden onafhankelijk zijn, kan er makkelijker gebruik worden gemaakt van tekst classificatie methodes. Een eerste versie waarin onafhankelijkheid tussen de woorden wordt verondersteld kan altijd verbeterd en aangepast worden. Een werkend concept waarin er wordt uitgegaan van onafhankelijkheid tussen de 'features' kan een goed begin zijn voor K&R.

Deze theorie zou bij K&R in de praktijk als volgt worden toegepast. In de formule zoals beschreven in vergelijking 1 zou dan 'A' een label van K&R zijn en 'B' de tekst in een materiaalomschrijving. Vervolgens wordt elk woord in een materiaalomschrijving gezien als de eerdergenoemde onafhankelijke 'feature'. Bij elk woord wordt de kans berekend dat het woord bij een label hoort. Dit wordt gedaan in *P(B/A)* wat dus gelijk staat aan $P(b1, b2, \ldots b_n|A)$ als $P(b1|A) \times P(b2|A) \times \ldots \times P(b_n|A)$. Hierbij staat '*b*' gelijk aan een woord/'feature' in een materiaalomschrijving en *n* gelijk staat aan de hoeveelheid woorden in een materiaalomschrijving. De uitkomst hiervan wordt vermenigvuldigd met de algemene kans op het label, wat in de formule staat als *P(A). P(B)* blijft gelijk zolang de berekeningen worden gedaan bij een bepaalde materiaalomschrijving, aangezien de tekst van de materiaalomschrijving niet veranderd.

Het algoritme werkt vervolgens als volgt. Voor materiaalomschrijving *X* wordt de kans berekend van alle labels, wat betekent dat in de formule 'B' constant blijft, maar 'A' elke keer verandert (elke keer weer een nieuw label). Voor elke nieuwe label wordt de kans berekend dat die label bij materiaalomschrijving X hoort. Dit gebeurt door de kans van elke feature dat die bij dat label hoort te

vermenigvuldigen met de algemene kans op dat label. Zodoende wordt voor elk label de kans berekend dat het label bij materiaalomschrijving X hoort. Het label met de hoogste kans, wordt uiteindelijk door het algoritme toegewezen aan materiaalomschrijving X. Nadat dit voor materiaalomschrijving X is gebeurd, gaat de algoritme door naar materiaalomschrijving Y, waarbij opnieuw elke kans van elk label wordt uitgerekend. Dit gaat zo door totdat alle materiaalomschrijvingen een label hebben gekregen.

3.5 Voor- en nadelen NB-theorie

De voordelen van de NB tekst classificatie zit voornamelijk in de makkelijke implementatie, simpele aanpak en hoge snelheid.(7) De theorie is goed te gebruiken voor een eerste opzet van tekst classificatie. Daarnaast geeft de hoge snelheid waarmee berekeningen en classificaties kunnen worden uitgevoerd een groot voordeel, aangezien er tegenwoordig steeds grotere data sets beschikbaar zijn.

Volgens Kumar et al. (12) wordt NB vooral gebruikt vanwege de eenvoud, elegantie en robuustheid. In dit artikel wordt wel gemeld dat aanpassingen aan het model verbeterde prestaties kunnen geven, maar vaak wel ten koste gaat van de eenvoud van het model. Er is ook veel literatuur te vinden bij het verbeteren van NB-classificaties, waardoor mogelijke verbeteringen later ook mogelijk zijn. Zo gaat Kim (9) in op effectieve methodes om NB tekst classificaties te verbeteren, maar zijn er ook andere mogelijkheden om het classificatie systeem in de toekomst verder te ontwikkelen (10,11) (Zhang, Jiang).

Een van de nadelen van de NB-theorie zijn de prestaties wanneer er een onbalans in data is. Met een onbalans in data wordt bedoeld dat de test dataset voor de meerderheid bestaat uit bepaalde labels en dat andere labels veel minder terugkomen. Volgens Kumar et al. (12) heeft een test set waarbij de kenmerken van klassen goed zijn weergegeven, een grotere kans op een succesvolle classificatie.

Daarnaast presteert een NB-classificatie soms minder wanneer er geen rekening wordt gehouden met woorden die vaker voorkomen in een document en daarom dus ook meer waarde hebben.(7) In NB wordt hier vanwege de veronderstelling van onafhankelijkheid tussen de woorden geen rekening mee gehouden. Voor de situatie van K&R is dit niet een heel groot nadeel, omdat bij veel classificaties modellen wordt gewerkt met grotere documenten waarin op basis van veel zinnen en woorden een classificatie voor het document moet worden gegeven. Bij K&R wordt er een classificatie van omschrijvingscellen gegeven, die veel korter zijn dan volledige artikelen of documenten. Binnen een omschrijvingscel worden bijna geen woorden herhaald, waardoor dit nadeel van NB voor de situatie van K&R geen probleem geeft.

Een belangrijk punt om te benoemen is dat er weinig tot geen literatuur te vinden is over classificatie theorieën die zijn toegepast op datasets met veel klassen. Ook op het gebied van ‘multionomial Naïve bayes’ zijn er hooguit voorbeelden van enkele tientallen categorieën. Dit maakt het wel onduidelijk of de NB-theorie nog steeds goed werkt bij een keuze uit veel meer verschillende klassen, wat namelijk het geval is bij K&R. Dit is een moeilijkheid die wordt veroorzaakt door de situatie van K&R en zou bij andere theorieën ook onduidelijk zijn. Van de verschillende classificaties theorieën die zijn te onderzoeken, lijkt de NB-theorie wel het meest geschikt voor een situatie met veel verschillende klassen/labels.

Een tweede punt van onzekerheid is de toepasbaarheid op de relatief lage hoeveelheid tekst die in materiaalomschrijving te vinden is. Binnen bijna alle bronnen over tekst classificaties met de NB-theorie zijn tekstdocumenten, mails of nieuwsberichten als voorbeeld genomen. Bij K&R is het zo dat

de tekst uit Excel-bestanden gehaald moeten worden en daardoor een ander soort format heeft. Onduidelijk is hoe NB-theorie hier op zal reageren.

3.6 Conclusie literatuurstudie

In de eerste twee secties zijn de algemene begrippen ‘Text Mining’, ‘Natural Language Processing (NLP)’, ‘Machine Learning (ML)’ en tekst classificatie besproken. Daarna is er dieper ingegaan op de verschillende algoritmes en theorieën die binnen tekst classificatie bekend zijn. In secties 3.4 en 3.5 is de ‘Naive Bayes’ theorie uitgelegd en waarom deze theorie van waarde kan zijn voor K&R. Met name de relatief makkelijke implementatie en hoge snelheid bij grote tekstdata vraagstukken zijn grote voordelen van deze theorie. Daartegenover staat dat het onduidelijk is of deze theorie ook goed werkt op vraagstukken met veel verschillende klassen. Verder is veel theorie gebaseerd op tekstdocumenten, mails of nieuwsberichten, terwijl de tekst data van K&R in Excel-bestanden zijn verwerkt.

In hoofdstuk 4 is uitgelegd hoe de theoretische kennis beschreven in hoofdstuk 3 in praktijk is gebracht voor de situatie van K&R.

4 Oplossingsontwerp

In hoofdstuk 4 wordt de in hoofdstuk 3 besproken theorie toegepast op de situatie van K&R. Met behulp van programmeertaal Python is een oplossingsontwerp ontwikkeld, waarvan de werking in dit hoofdstuk wordt uitgelegd. Stapsgewijs worden de verschillende experimenten doorgenomen, waarvan de resultaten in hoofdstuk 5 worden besproken.

Uit de literatuurstudie van hoofdstuk 3 kwam de Naive Bayes algoritme voor het gebruik van tekst classificatie naar voren als beste optie voor K&R. Voor de werking van een tekst classificator dienen er trainings- en testdata te worden gekozen. Deze twee datasets worden in de eerste twee secties van dit hoofdstuk besproken. Daarna worden er twee experimenten uitgewerkt, waarbij er een zonder en een met voorbewerking van data wordt uitgevoerd.

4.1 Trainingsdata

In hoofdstuk 3 is de huidige situatie van K&R uitgelegd, waarbij is ingegaan op het analyseproces zoals dat nu wordt uitgevoerd in Excel. Dit analyseproces wordt al meerdere jaren op deze manier uitgevoerd, dus de meeste omschrijving aan label gekoppelde data is in deze Excel-bestanden te vinden. Deze verschillende Excel-bestanden zullen dus worden onderverdeeld in een trainingsset en een testset.

Voor het trainen van het algoritme is gekozen voor een trainingsset van 30 verschillende analyses. Samen bestaan deze analyses in totaal uit ongeveer 20.000 Excel-regels, waarbij in elke regel in een omschrijving en een label aan elkaar is gekoppeld. De trainingsdata bestaat uit 788 labels van de 2500 labels in totaal beschikbaar in de database van K&R. Afhankelijk van de mate van voorbewerking van data zijn er tussen de 8941 en 9289 unieke ‘features’ te onttrekken uit de trainingsdata.

Aan het prepareren van de trainingsdata hebben enkele haken en ogen gezeten. De problemen die zich voordeden zijn vooral op software-gebied en zijn besproken in appendix 7.6 (pagina 43).

4.2 Testdata

Om te testen hoe goed de classificator gebaseerd op de trainingsdata werkt, zijn er ook een aantal test bestanden gekozen. Er is gekozen voor twee keer drie verschillende analyses, zodat uitgesloten kan worden dat mogelijk betere resultaten tussen de verschillende experimenten op toeval berust. Dit betekent dus dat er in totaal zes verschillende analyse-bestanden zijn bestaande uit twee groepen van drie. Bij de testen op de ene groep zijn de analyse-bestanden van de andere groep toegevoegd aan de trainingsdata. De trainingsdata blijft op die manier bestaan uit 30 verschillende analyses. Door meerdere testbestanden te nemen, kan er met meer zekerheid conclusies getrokken worden uit de verschillende resultaten van de experimenten.

De namen van de drie analyse bestanden uit groep 1 zijn: ‘190118 ANA TP6 Loods’, ‘190515 Analysemodel Electro v2 (database hard)’ en ‘190621 ANA Landstation TenneT W-installaties’. De analyse bestanden uit groep 2 zijn: ‘190903 ANA E VU Onderzoeksgebouw V2’, ‘190903 Analysemodel W’ en ‘190905 ANA ASML gebouw 8 W’. Er is voor deze zes bestanden gekozen omdat er een goede variatie tussen de bestanden zit. Zo bestaan sommige analyses uit 300 omschrijvingen, waar andere uit ongeveer 1700 omschrijvingen bestaan. Verder zijn de omschrijvingen in alle bestanden niet overdreven netjes zonder enige vorm van ruis, maar hebben vaak wel woorden die kenmerkend zijn voor de label die bij de omschrijving hoort. Ook komen een aantal veel voorkomende afkortingen en vakspecifieke termen terug die zorgen voor een goed gevarieerde

testset. Deze analyse bestanden zijn daarom een goede doorsnede van de analyses zoals K&R die tegenkomt.

4.3 Eerste experiment

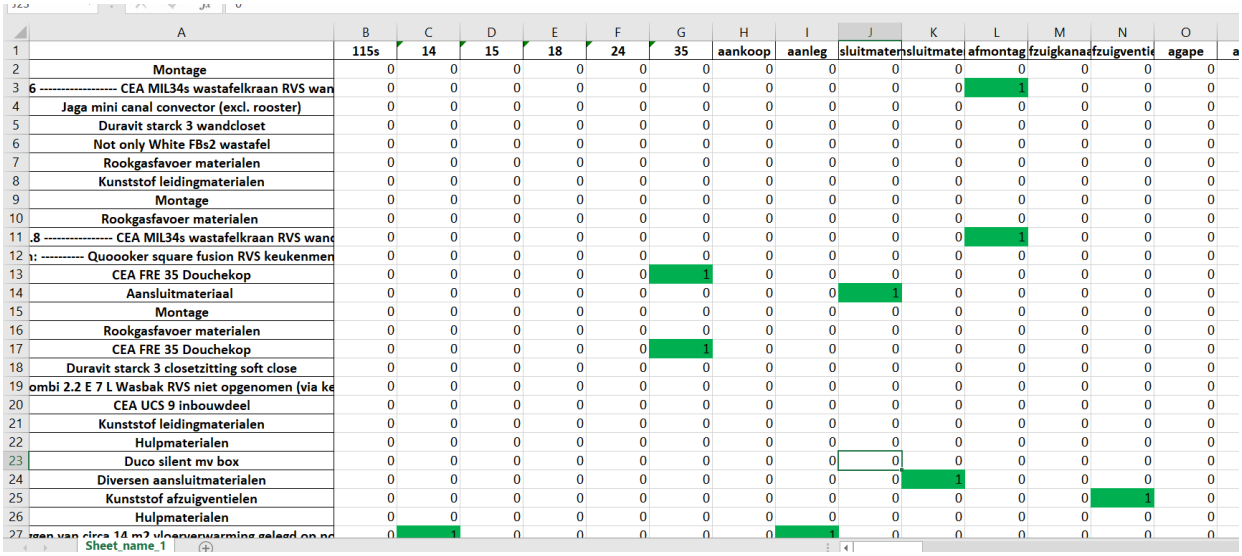
Er is gekozen voor twee verschillende experimenten, waarbij het eerste experiment ook wel het basismodel genoemd kan worden. Het tweede experiment kan het best worden uitgelegd als het uitproberen van verschillende tekst classificatie optimalisatie technieken. Deze technieken worden veel gebruikt in de literatuur en zijn allemaal kleine aanpassingen op het basismodel van het eerste experiment.

Het eerste experiment is uitgevoerd op de dataselectie besproken in secties 4.1 en 4.2. In dit basismodel wordt de meest eenvoudige manier van het Naive Bayes model uitgevoerd, waarbij er geen ruis wordt verwijderd of een moeilijke vectorizer wordt gebruikt. Een vectorizer is de methode waarmee de tekstdata wordt gevectoriseerd. In deze sectie wordt stapsgewijs het hele classificatieproces uitgelegd, waarna in het tweede experiment in sectie 4.4 alleen de verschillen ten opzichte van het basismodel worden besproken.

Het classificatieproces begint met het importeren van de Excel-kolommen met daarin de omschrijvingen en bijbehorende labels. Elk uniek label krijgt zijn eigen labelnummer, waarna dit wordt opgeslagen in een ‘dictionary’. Vervolgens wordt elk woord binnen alle omschrijvingen als aparte ‘feature’ geïdentificeerd. Dit wordt ook wel ‘tokenization’ genoemd. Hierna wordt elke ‘feature’ gevectoriseerd, omdat het Naive Bayes algoritme, net als andere tekst classificatie algoritmes, geen ruwe tekst kan verwerken.

Dit vectoriseren wordt gedaan volgens het ‘Bag of Words’ (BoW) principe, waarbij wordt geteld hoe vaak een feature voorkomt in een omschrijvingscel. Met deze manier van vectoriseren wordt de structuur van de woorden in het document losgelaten. Het draait enkel om het herkennen van de woorden in een omschrijvingscel.

Het trainen van het model gaat op de volgende manier in zijn werk. Alle features uit de trainingsdata worden geleerd en staan daarna in een soort vocabulaire. Met de informatie uit de vocabulaire en de ‘features’ in de omschrijvingscellen in de trainingsdata, kan er een matrix worden gemaakt. Elke omschrijvingscel met haar aparte ‘features’ wordt langsgegaan en voor elke ‘feature’ uit het vocabulaire wordt aangegeven of het in die specifieke omschrijvingscel voorkomt. Dit wordt gedaan door een 1 te plaatsen bij de vocabulaire-‘features’ die in die omschrijvingscel staan en een 0 te plaatsen bij de features die dat niet doen. Zo kunnen alle omschrijvingscellen samen uit één analyse als matrix worden gerepresenteerd, met op elke rij een omschrijvingscel en op elke kolom een ‘feature’ (woord) uit het vocabulaire. Zo’n matrix wordt opgesteld voor de trainingsdata, waarna vervolgens ook een matrix wordt opgesteld voor de testdata. Hierbij staan als rijen de omschrijvingscellen van de testdata en als kolommen de al bekende vocabulaire. Onbekende woorden uit de omschrijvingscellen die niet in het vocabulaire voorkomen, worden genegeerd en komen dus niet terug in de matrix. Deze matrixen worden ook wel ‘document-term matrixen’ genoemd waarbij in deze situatie de documenten de omschrijvingscellen zijn en de termen de ‘features’ uit het vocabulaire. Een voorbeeld van een ‘document-term matrix’ is te vinden in Figuur 9 waarbij een matrix is geëxporteerd naar Excel. In elke kolom is een ‘feature’ te vinden en in elke rij is een omschrijving te vinden. De cellen waarin een 1 voorkomt zijn voor de duidelijkheid groen gemaakt en voor deze cellen geldt dus dat het ‘feature’ voorkomt in de betreffende omschrijving.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		115s	14	15	18	24	35	aankoop	aanleg	sluitmater	nsluitmate	afmontag	fzuigkanaa	fzuigventi	agape
2	Montage	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	6 ---------- CEA MIL34s wastafelkraan RVS wan	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4	Jaga mini canal convector (excl. rooster)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Duravit starck 3 wandcloset	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Not only White FBs2 wastafel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Rookgasfavoer materialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Kunststof leidingmaterialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Montage	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Rookgasfavoer materialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	.8 ---------- CEA MIL34s wastafelkraan RVS wan	0	0	0	0	0	0	0	0	0	0	1	0	0	0
12	: ---------- Quoooker square fusion RVS keukenmen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	CEA FRE 35 Douchekop	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14	Aansluitmateriaal	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	Montage	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Rookgasfavoer materialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	CEA FRE 35 Douchekop	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18	Duravit starck 3 closetzitting soft close	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	ombi 2.2 E 7 L Wasbak RVS niet opgenomen (via ke	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	CEA UCS 9 inbouwdeel	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Kunststof leidingmaterialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Hulpmaterialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Duco silent mv box	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Diversen aansluitmaterialen	0	0	0	0	0	0	0	0	0	1	0	0	0	0
25	Kunststof afzuigventielen	0	0	0	0	0	0	0	0	0	0	0	0	1	0
26	Hulpmaterialen	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	gen van circa 14 m2 vloerverwarming gelegd op pc	0	1	0	0	0	0	0	1	0	0	0	0	0	0

Sheet_name_1

Figuur 9 - Voorbeeld van een deel van een 'document-term matrix', waarin als rijen de omschrijvingen staan en de kolommen de 'features' uit de trainingsdata vocabulaire zijn.

De eerste matrix gebaseerd op de trainingsdata wordt vervolgens gebruikt om samen met de label nummers het Naive Bayes model te trainen. De tekst is nu omgezet in een matrix waarin alleen maar nullen en enen voorkomen, waardoor het algoritme het nu kan lezen. Ook de labels zijn geconverteerd naar label nummers. Aan de hand van deze matrix waarin staat welke 'features' bij welke omschrijvingscellen voorkomen en de label nummers behorende bij die omschrijvingscel, leert het algoritme de kansen dat een bepaalde 'feature' bij een bepaalde label hoort.

Vervolgens wordt na het trainen de tweede matrix van de testdata in het model geplaatst. Op basis van deze matrix, waarin staat welke al bekende 'features' voorkomen in de omschrijvingscellen, voorspelt het model met gebruik van het Naive Bayes algoritme het bijbehorende label nummer. De precieze werking van dit algoritme is al uitgelegd in hoofdstuk 3, maar voor elk label wordt de kans berekent dat de omschrijvingscel bij dat label hoort. Het label met de hoogste kans wordt uiteindelijk toegewezen aan de omschrijvingscel.

Ten slotte worden de voorspelde labels samen met de werkelijke labels vergeleken, waardoor een nauwkeurigheidsscore kan worden berekend. De omschrijvingscellen, de werkelijke labels, de voorspelde labels en de kans van het voorspelde label horende bij die omschrijving worden als output geëxporteerd naar Excel. Dit wordt uitgevoerd voor beide testdata groepen, zodat er vergeleken kan worden.

4.4 Tweede experiment

In deze sectie wordt het tweede experiment uitgelegd, waarbij verschillende tekst classificatie optimalisatie technieken zijn uitgeprobeerd. De volgende technieken worden veel gebruikt in tekst classificatie en zijn allemaal kleine aanpassingen op het basismodel uit sectie 4.3. De verschillende technieken worden in deze sectie kort uitgelegd en zijn in meerdere combinaties uitgeprobeerd tot de beste resultaten zijn gevonden.

4.4.1 TF-IDF

In het eerste experiment is een methode van vectoriseren uitgelegd waarbij er wordt gekeken naar de frequentie van woorden in de omschrijving. Een nadeel van deze methode is dat woorden die veel voorkomen in de volledige analyse als belangrijk kunnen worden gezien, ondanks dat dit vrij onbelangrijke woorden als 'en' of 'de' kunnen zijn. Om deze woorden toch niet na voren te laten

komen als belangrijk, kunnen de frequenties van woorden met een wegingsfactor worden gecorrigeerd naar hoe vaak ze in het totale analyse-bestand voorkomen. Deze benadering wordt TF-IDF genoemd, oftewel 'Term Frequency – Inverse Document Frequency'. Hierbij staat TF voor de frequenties van een woord in een omschrijvingscel, terwijl IDF staat voor het logaritme van het aantal omschrijvingscellen in een analyse gedeeld door het aantal documenten waarin een bepaald woord voorkomt. Deze twee getallen worden met elkaar vermenigvuldigd waardoor een nieuw wegingsfactor voor een woord ontstaat, waarbij er meer belang wordt gehecht aan uniekere woorden.

4.4.2 Verwijderen ruis

Het verwijderen van ruis op verschillende manieren is ook een methode waarmee een tekst classificator betere resultaten kan geven. Met ruis worden tekens, woorden en nummers bedoeld die geen waarde zouden moeten hebben bij de classificatie. Een bekend voorbeeld hiervan zijn stopwoorden, woorden die veel voorkomen maar geen informatie verschaffen over het label. Er zijn standaard lijsten te vinden voor de Nederlandse taal, waarin woorden als 'de', 'het' en 'er' te vinden zijn. Ook is het mogelijk om eigen woorden toe te voegen, zoals het woord 'mm' dat 1806 keer voorkomt in de trainingsdata, zonder dat het iets zegt over het bijbehorende label. Ook kan gebruik gemaakt worden van synoniem- en afkortingslijsten, waarin specifiek voor de situatie van K&R bepaalde woorden in kunnen worden opgenomen. Hierbij kan gedacht worden aan afkortingen als 'st' en 'mtr' die respectievelijk 462 en 230 keer voorkwamen in de trainingsdata. Deze termen zijn afkortingen van de woorden 'stuk' en 'meter' die zelf ook al respectievelijk 215 en 61 keer voorkwamen. De afkorting en het volledig uitgeschreven woord kunnen door die lijst als één 'feature' worden meegenomen in de classificatie.

Ook kan alle interpunctie worden weggehaald, wat de snelheid en soms ook de nauwkeurigheid van het model verbeterd. Andere vergelijkebare methodes zijn het verwijderen van hoofdletters en het verwijderen van onnodige witruimtes in de tekst. Voor bepaalde tekst classificatie voorbeelden is het verwijderen van getallen vaak ook een goede verbetering. De vraag bij K&R is of dit ook echt een verbetering is omdat sommige getallen ook informatie verstrekken over het label. Een voorbeeld hiervan zijn de labels 'Koudtapwaterleiding_Ø_15' en 'Koudtapwaterleiding_Ø_22' waarbij het verschil in de omschrijving vaak ook alleen de cijfers '15' en '22' zijn. Zonder de informatie van die getallen kan de classificator niet het juiste label classificeren. Daarom is ervoor gekozen om de getallen niet te verwijderen in dit experiment.

Ten slotte is er nog een methode genaamd 'stemming'. Hierbij worden alle woorden teruggebracht tot hun stam, waardoor woorden met dezelfde betekenis maar die anders geschreven staan naar dezelfde stam worden teruggebracht. Een voorbeeld hiervan zijn de woorden 'montage' en 'montages', die beiden teruggebracht zullen worden naar hun stam 'montag'. Door de verandering van deze woorden naar hun stam, herkent het model deze woorden als dezelfde woorden. In andere gevallen zou het model deze twee woorden niet aan elkaar kunnen koppelen.

4.4.3 'N-grams' & 'alpha'

In sectie 4.4.2 zijn aanpassingen bij het voorbereiden van de data voordat het algoritme erop los wordt gelaten toegepast. Er kan echter ook in het model met een aantal parameters gewerkt worden. De belangrijkste hiervan zijn de 'N-grams' en de 'alpha'.

Een 'N-gram' is een nieuwe 'feature' bestaande uit N woorden. Het idee is dat wanneer er gekozen wordt voor een 'N-gram' van (1,2) er een aantal extra 'features' worden onttrokken van een omschrijving. Als een omschrijving bestaat uit 'Vervuild filter alarm' bestaan de 'features' uit 'Vervuild', 'filter', 'alarm', 'vervuild filter' en 'filter alarm'. Deze zogenaamde '2-grams' worden

opgeslagen als aparte ‘feature’ en krijgen dus ook een eigen kolom in de ‘document-term matrix’. Hierdoor kan er meer informatie uit een omschrijvingscel gehaald worden, waarbij er ook een ‘N-gram’ range van 3 of 4 woorden gekozen kan worden. Dit vertraagt het model wel en de vraag is of deze techniek goed werkt bij K&R. Deze methode schijnt vaak goed te werken als bepaalde woorden essentieel zijn voor de betekenis van het woord daarna, zoals ‘geen goede test’. Als elk ‘feature’ apart wordt genomen, komen de woorden ‘goede’ en ‘test’ beide voor, waardoor de classificator de betekenis niet op de juiste manier kan interpreteren. Het woord ‘geen’ bij deze tekst is essentieel en zou bij het gebruik van ‘N-grams’ wel beter meegenomen zijn.

Een andere belangrijke parameter is de ‘alpha’ binnen de formule van Naive Bayes. Om uit te leggen wat de ‘alpha’ doet, is het belangrijk om te begrijpen wat het Naive Bayes algoritme doet met woorden in een omschrijving die voor een bepaalde categorie nog nooit zijn voorgekomen. Wanneer het NB-algoritme de kans berekent van een bepaald label bij een bepaalde omschrijving, wordt voor elke ‘feature’ (woord) berekent wat de kans is dat dat ‘feature’ bij het label hoort. Wanneer in de trainingsdata die ‘feature’ nog nooit is voorgekomen bij een bepaalde label, zal het NB-algoritme deze kans 0 maken. Dit is een probleem, omdat de kansen van alle ‘features’ in een omschrijving met elkaar worden vermenigvuldigd en de totale kans van die label bij die omschrijving dan altijd 0 wordt, ongeacht van de kansen van de andere ‘features’. Om dit probleem op te lossen wordt er altijd een vast getal opgeteld bij elke kans. Dit wordt ook wel ‘additive smoothing’ of ‘Laplace smoothing’ genoemd en het getal dat erbij wordt opgeteld is de ‘alpha’. Deze ‘alpha’ staat in Python standaard op 1, maar kan aangepast worden naar een kleiner getal. De ‘alpha’ is dus de kans die gegeven wordt aan een ‘feature’ die nog niet in het trainingsvocabulaire voorkomt, maar de ‘alpha’ wordt ook opgeteld bij de ‘features’ die wel voorkomen in het vocabulaire. Hierdoor krijgen de ‘features’ die wel bekend zijn altijd een hogere kans dan niet bekende ‘features’, maar zijn de kansen van de onbekende ‘features’ niet 0. Vaak reageert een model op een kleiner ‘alpha’ beter, zeker wanneer er een onbalans in de trainingsdata is. Omdat de trainingsdata relatief beperkt is, kan het aanpassen van deze ‘alpha’ een significante verbetering veroorzaken.

Het zoeken van de beste parameters voor de dataset kan gedaan worden door middel van de functie ‘GridSearchCV’. Hierbij wordt er gezocht naar de parameters die het beste presteren op het huidige model.

In dit hoofdstuk is het oplossingsontwerp besproken zoals dat is geïmplementeerd op de situatie van K&R. De resultaten van deze experimenten zijn te vinden in hoofdstuk 5.

5 Resultaten

Hoofdstuk 5 gaat over de resultaten van de experimenten uitgelegd in hoofdstuk 4. De nauwkeurigheid van de voorspellingen worden benoemd met enkele kanttekeningen. Ook worden de parameters van het algoritme uitgelegd en worden een aantal extra inzichten gegeven voor verdere analyse.

5.1 Nauwkeurigheid

De nauwkeurigheid van de voorspellingen van het eerste experiment, waarin gebruik wordt gemaakt van het 'Bag of Words' principe, zijn te vinden in Tabel 1 en Tabel 2. De percentages zijn het resultaat van het delen van de totale aantallen labels in een analyse door de juist voorspelde labels. Voor dit experiment zijn verder geen veranderingen op de data of het model toegepast.

Analyse testbestanden groep 1	**Nauwkeurigheid**
190118 ANA TP6 Loods	11,97%
190515 Analysemodel Electro v2 (database hard)	4,91%
190621 ANA Landstation TenneT W-installaties	1,72%

Tabel 1 - Nauwkeurigheidspercentages eerste experiment eerste groep

Analyse testbestanden groep 2	**Nauwkeurigheid**
190903 ANA E VU Onderzoeksgebouw V2	11,40%
190903 Analysemodel W	3,05%
190905 ANA ASML gebouw 8 W	4,03%

Tabel 2 - Nauwkeurigheidspercentages eerste experiment tweede groep

De nauwkeurigheid valt laag uit, waarbij de nauwkeurigheid bij vier van de gevallen zelfs onder de 5% is. De resultaten van het tweede experiment zijn in Tabel 3 en Tabel 4 te vinden.

Analyse testbestanden groep 1	**Nauwkeurigheid**
190118 ANA TP6 Loods	22,49%
190515 Analysemodel Electro v2 (database hard)	14,73%
190621 ANA Landstation TenneT W-installaties	10,34%

Tabel 3 – Nauwkeurigheidspercentages tweede experiment eerste groep

Analyse testbestanden groep 2	**Nauwkeurigheid**
190903 ANA E VU Onderzoeksgebouw V2	14,21%
190903 Analysemodel W	8,22%
190905 ANA ASML gebouw 8 W	6,45%

Tabel 4 - Nauwkeurigheidspercentages tweede experiment tweede groep

Bij het tweede experiment is gekozen om de data voor te bewerken voordat het op het algoritme wordt toegepast. De TF-IDF methode is toegepast, samen met een 'alpha' van 0.01. Andere voorbewerkingen die zijn gedaan zijn het verwijderen van punctuatie, lege witruimtes en hoofdletters. Ook is ervoor gekozen om een stopwoorden- en afkortingenlijst te maken, waarvan de stopwoorden voornamelijk bestaan uit algemene Nederlandse stopwoorden. De afkortingenlijst bestond alleen uit 'mtr' en 'st', omdat dit verreweg de meest voorkomende afkortingen waren in de 'features' van de trainingsdata. Daarnaast is ervoor gekozen om alle 'features' terug te brengen naar hun stam door middel van de 'Kraaij-Pohlmann Snowball Stemmer'. Ten slotte is ervoor gekozen om gebruik te maken van '4-grams'.

De resultaten van het tweede experiment zijn beter dan het eerste experiment, maar het zijn nog steeds geen nauwkeurigheidsscores waar veel waarde aan te hechten is. Het vervolg van dit hoofdstuk zal ingaan op de kanttekeningen die bij deze resultaten kunnen worden gezet. Daarnaast worden ook de keuzes voor parameters in sectie 5.3 besproken.

5.2 Kanttekeningen nauwkeurigheidsscores

In deze sectie zijn een aantal extra analyses uitgevoerd waardoor er een beter beeld ontstaat van de werking van het model en de inhoud van de trainingsdata. Om te kijken welke invloed de hoeveelheid trainingsdata op de nauwkeurigheidsscores heeft, is er gekeken hoe die scores zich ontwikkelen wanneer het model met minder trainingsdata wordt getraind. Deze resultaten zijn terug te vinden in Tabel 5, waarbij er een voorspelling is gemaakt gebaseerd op 7.500, 16.000 en 20.000 omschrijvingsregels. Dit experiment is alleen uitgevoerd voor groep 1 van de testbestanden. Al deze experimenten zijn uitgevoerd met dezelfde parameters en voorbewerkingen van data als in experiment twee.

Analyse testbestanden	**Nauwkeurigheid 7.500 omschrijvingen**	**Nauwkeurigheid 16.000 omschrijvingen**	**Nauwkeurigheid 20.000 omschrijvingen**
190118 ANA TP6 Loods	4,53%	21,84%	22,49%
190515 Analysemodel Electro v2 (database hard)	5,36%	14,29%	14,73%
190621 ANA Landstation TenneT W-installaties	2,59%	8,62%	10,34%

Tabel 5 - Nauwkeurigheidsscores van de drie testanalsyes bij verschillende hoeveelheid trainingsdata

Logischerwijs neemt de nauwkeurigheid toe naarmate er met meer data wordt getraind. Gelukkig is er meer trainingsdata beschikbaar en deze cijfers wekken wel een nieuwsgierigheid naar de prestaties van het model bij trainingsdata van meer dan 20.000.

De huidige trainingsdata bestaan na het voorbewerken van data uit 8941 ‘features’ en 788 labels. Er zijn in totaal 2500 labels in de database van K&R, dus lang niet alle labels zijn gerepresenteerd in deze trainingsdata. Wel zullen de meest voorkomende labels terug te vinden zijn in deze trainingsset. Na analyse van de score op bestand ‘190118 ANA TP6 Loods’ blijkt dat van de werkelijke labels er maar twee nog niet in het trainingsbestand zaten. Een hiervan (‘keerklep’) is redelijk veel voorkomend, de ander (voeding_zonwering) is unieker.

Ook interessant is dat wanneer de trainingsdata als testdata worden ingevoerd, de labels met de bewerkingen van experiment twee met 73,79% nauwkeurigheid worden voorspeld. Met het basismodel wordt er een score van 54,71% voorspeld. Deze score wordt ook wel de ‘model fit’ genoemd en is altijd beter dan de nauwkeurigheidsscores, maar het geeft wel aan dat er nog duidelijk ruimte voor verbetering is. De ‘model fit ’score is wel redelijk laag, aangezien de trainingsdata hier exact overeenkomt met de testdata. Er wordt dus al ongeveer 27% kwijtgeraakt op omschrijvingen die erg veel op elkaar lijken, maar toch andere labels hebben. Hierdoor raakt het algoritme in de war en wijst het algoritme het verkeerde label toe. Ook de omschrijvingen die bij meerdere onderdelen voorkomen, zoals ‘hulpmaterialen’ of ‘montagekosten’ worden vaak met het verkeerde label voorspeld.

Bij het analyseren van de resultaten die geëxporteerd worden naar Excel kunnen er verschillende opmerkingen worden gemaakt. Er kan bijvoorbeeld gekeken worden hoe ver een fout voorspeld label van het werkelijke label afligt. Het fout voorspelde label heeft toch meer waarde wanneer het

verschil met het werkelijke label alleen op een klein detail fout zit. Er zijn bijvoorbeeld 8 verschillende pvc-buis labels en wanneer het algoritme toch een van deze acht pvc-buizen voorspelt, maar niet exact de juiste, kan worden gezegd dat het algoritme beter werkt dan de nauwkeurigheidsscore doet vermoeden. De nauwkeurigheidsscore kijkt alleen naar de voorspellingen die exact goed en fout zijn, terwijl er niet gekeken wordt naar het verschil tussen de voorspelling en werkelijkheid. Bij de voorspellingen van de analysebestanden uit groep 1 is er daarom gekeken naar de top 3 labels bij elke omschrijving. Vervolgens is er berekent hoe vaak het juist label bij de top 3 zit, waardoor er een iets genuanceerder beeld naar voren komt wat betreft de resultaten. Deze resultaten zijn te vinden in Tabel 6. Hieruit is op te maken dat de nauwkeurigheid van de top 3 labels significant hoger ligt dan wanneer er alleen naar het eerste label gekeken wordt. Dit geeft aan dat het algoritme beter werkt dan de eerste nauwkeurigheidsscores doen geloven en dat de praktische waarde van het algoritme daardoor ook groter is.

Analyse testbestanden groep 1	**Nauwkeurigheid experiment 2**	**Nauwkeurigheid experiment 2 top 3**
190118 ANA TP6 Loods	22,49%	34,14%
190515 Analysemodel Electro v2 (database hard)	14,73%	21,43%
190621 ANA Landstation TenneT W-installaties	10,34%	18,10%

Tabel 6 - Nauwkeurigheidspercentages van groep 1 experiment 2 zonder en met top 3 labels.

Ook op de scores van de top 3 labels zijn weer een extra nuance aan te brengen. Zoals is uitgelegd in de vorige alinea zijn er acht verschillende pvc-buizen en blijkt uit een snelle analyse van het eerste analyse testbestand '190118 ANA TP6 Loods' dat voor deze pvc-buizen ook binnen de top 3 niet altijd het juiste label wordt gekozen. Hetzelfde verhaal geldt voor de PE-buizen en een aantal andere elementen. Een voorbeeld hiervan is te zien in Tabel 7, waarin twee omschrijvingen met het werkelijke label en de top 3 voorspelde labels zijn geprojecteerd. Het laat zien dat er binnen de top 3 wel labels voorkomen die een afvoerleiding van het juiste materiaal teruggeven, maar het juiste label zit er niet bij. Er is een kans dat het juiste label wel in de top 10 voorspelde labels staat, maar voor dit onderzoek is er alleen naar de top 3 gekeken. Het algoritme zit dus naast de percentages te vinden in Tabel 6 ook bij een aantal andere omschrijvingen nog steeds dicht bij het juiste label. Deze omschrijvingen worden niet meegenomen in een van de nauwkeurigheidsscores, maar zijn wel van belang om een goed totaalbeeld te krijgen van de prestaties van het algoritme.

	Om-schrijving	**actual_label**	**prediction_name_1**	**prediction_name_2**	**prediction_name_3**
28	PVC Ultra 3 110 mm	Vuilwater_afvoer_PVC_Ø_75	Afvoerleiding_PE_ontluchtingskappen	Afvoerleiding_PVC	Vuilwater_afvoer_PVC_Ø_110
29	Geberit PE buis 50 mm	Afvoerleiding_PE_50_mm	Afvoerleiding_PE_volvul	Wandclosetcombinatie_Sphinx_of_Grohe_Woningbouw_	Afvoerleiding_PE_160_mm

Tabel 7 - Voorbeeld van twee omschrijvingen waarbij het werkelijke label en de top 3 voorspelde labels voorkomen.

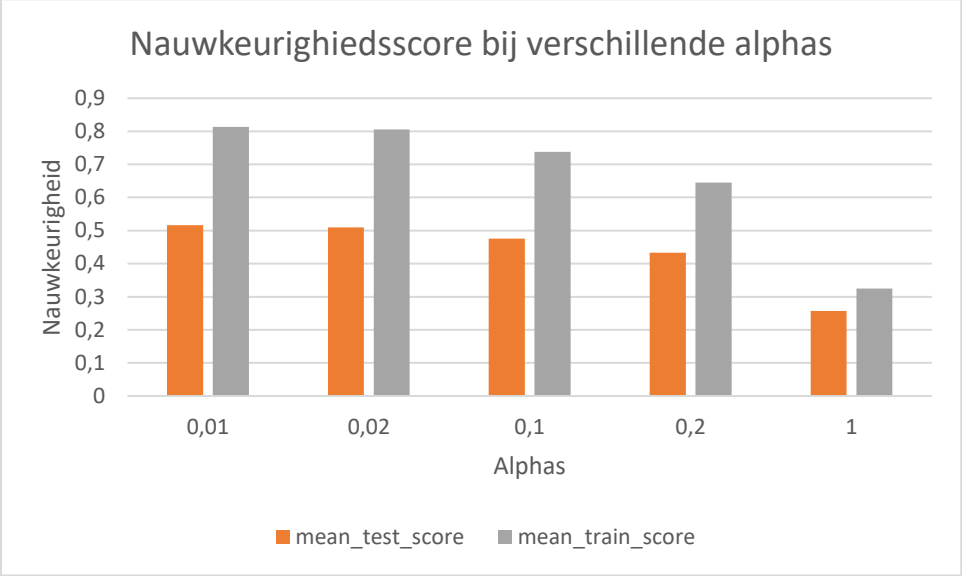
5.3 Parameters

Bij het tweede experiment zijn een deel van de parameters veranderd en voorbewerkingstechnieken toegepast. In deze sectie wordt door middel van een aantal figuren getoond voor welke parameters is gekozen en waarom.

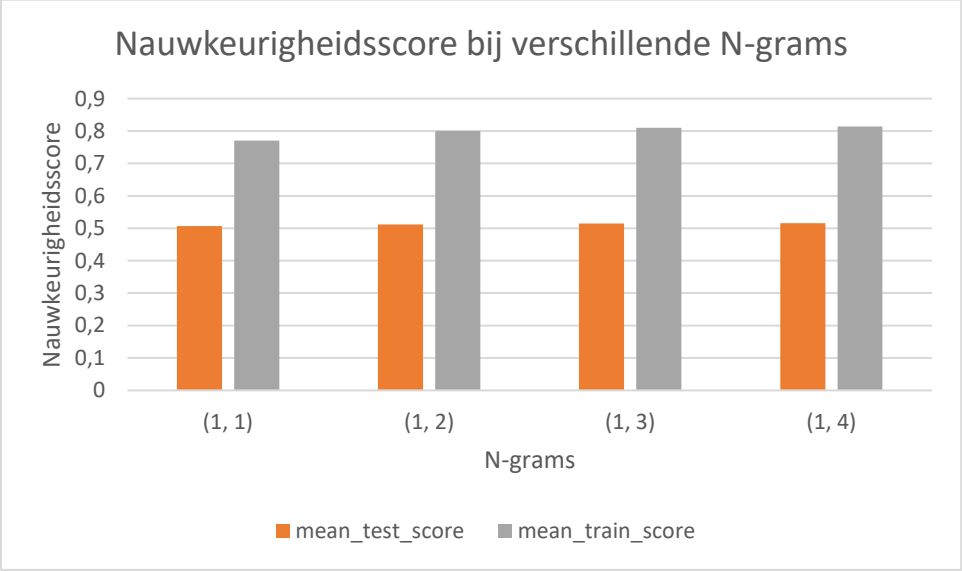
Bij experiment twee is ervoor gekozen om de data te verwijderen van punctuatie en lege witruimtes en het vervangen van hoofdletters door kleine letters. Ook zijn er 100 Nederlandse veel

voorkomende stopwoorden uitgehaald, samen met een kleine zelf opgestelde synoniemen lijst. Daarnaast is gebruik gemaakt van ‘stemming’.

Na het aanpassen van de data is er op zoek gegaan naar de best mogelijk parameters. Hierbij zijn er meerdere combinaties uitgeprobeerd voor de ‘alpha’ van het Naive Bayes model en de ‘N-grams’ tijdens het vectoriseren. In Figuur 10 staan de nauwkeurigheidsscores bij het kiezen van verschillende ‘alphas’ en in Figuur 11 van de verschillende ‘N-grams’. Bij beide is ervoor gekozen om de test- en trainingsscore te bepalen, waarbij de testscore de nauwkeurigheid van voorspellingen is op een nieuwe dataset. Bij de trainingsscore is juist de nauwkeurigheid genomen bij het voorspellen van de al getrainde dataset. De combinatie van ‘alpha’ als 0,01 met ‘N-grams’ (1,4) geven de beste resultaten en zijn daarom gekozen als parameters tijdens het uitvoeren van het tweede experiment.



Figuur 10 - Nauwkeurigheidsscore bij verschillende 'alphas'.

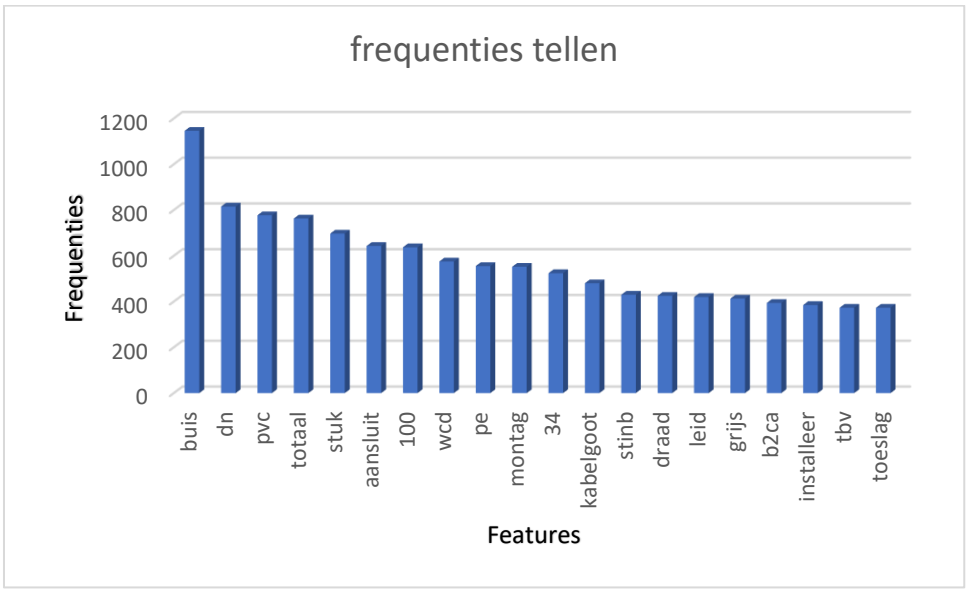


Figuur 11 - Nauwkeurigheidsscore bij verschillende N-grams.

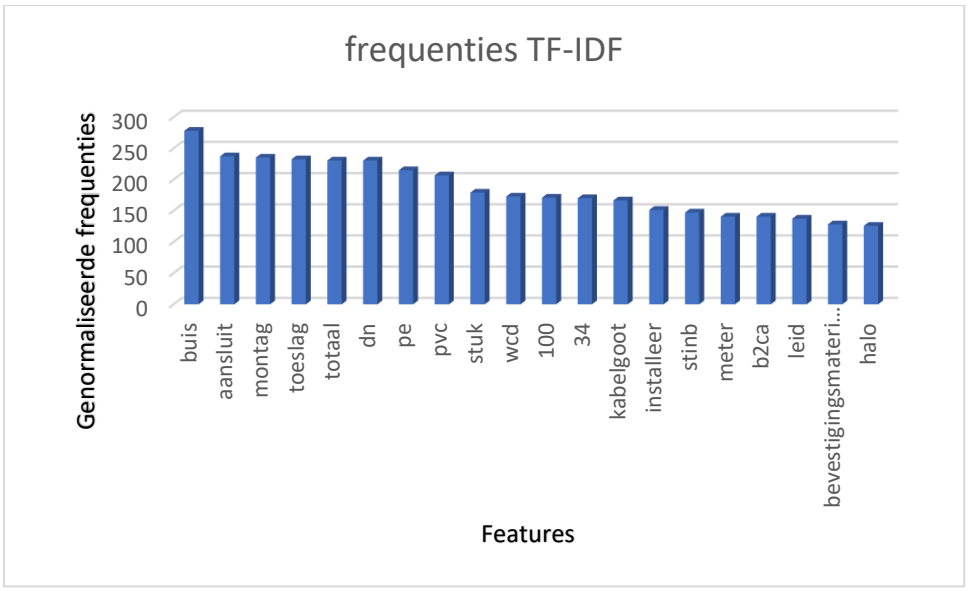
Bij het tweede experiment werd er ook gebruik gemaakt van een ‘TF-IDF’ vectorizer in plaats van een gewone frequentie vectorizer. Om het verschil van aanpak tussen de vectorizers aan te tonen, zijn in

Figuur 12 en Figuur 13 de top 20 meest voorkomende ‘features’ te zien. Op de y-as is ook te vinden hoe vaak elke ‘feature’ in de volledige trainingsset is voorgekomen.

Uit Figuur 12 en Figuur 13 is af te leiden waarom de ‘TF-IDF’ vectorizer beter werkte dan de vectorizer gebruikt in het eerste experiment. De top vijf ‘features’ in Figuur 12 komen significant vaker voor dan de andere vijftien ‘features’, waardoor deze mogelijk een veel te groot belang krijgen. In Figuur 13 is te zien dat de veel voorkomende ‘features’ uit Figuur 12 nog steeds hoog scoren qua frequentie, maar niet meer zo ver boven de rest uit steken. Deze ‘features’ worden dus nog steeds als belangrijk herkent, maar niet meer als buiten proportioneel belangrijk. De ‘feature’ frequenties zijn dus genormaliseerd naar de dataset.



Figuur 12 - Top 20 meest voorkomende 'features', waarbij het tellen van woord frequenties als vectorizer werd gebruikt.



Figuur 13 - Top 20 meest voorkomende 'features', waarbij de TF-IDF vectorizer is gebruikt.

5.4 Databeperking

In sectie 4.1 is uitgelegd waarom er maar met een beperkte trainingsset wordt gewerkt. Dit heeft zijn weerslag op de testresultaten, die over het algemeen teleurstellend laag zijn te noemen. Om het model beter te laten werken, dient het model met meer data getraind te worden. Meer data betekent dat de belangrijkste 'features' voor het model betekenisvollere worden. De trainingsnauwkeurigheid van een Naive Bayes model is bij goed werkende model vaak boven de 90%, terwijl het huidige model tot enkel 70% komt. Ook de andere scores vallen laag uit, maar het is belangrijk om deze kanttekening van een beperkte trainingsset in het oog te houden.

In het volgende hoofdstuk zullen er conclusies getrokken worden gebaseerd op het literatuuronderzoek, de resultaten van het oplossingsontwerp en de ervaringen opgedaan gedurende de scriptie.

6 Conclusie, discussie & aanbevelingen

Op basis van de resultaten uit het vorige hoofdstuk kunnen conclusies getrokken worden. Daarnaast zullen er meerdere aanbevelingen worden gedaan waarmee het model verbeterd of veranderd kan worden. Ook zal er een mogelijk vervolgonderzoeksrichting worden genoemd waarmee K&R dit automatiseringsproject kan doorzetten.

6.1 Conclusie & relevantie

Het onderzoek is begonnen met het doel om een simpel, werkend model te krijgen waarmee kan worden uitgezocht of er voor K&R mogelijkheden zijn in het veld van tekstclassificatie. Deze scriptie moet dus gezien worden als een verkennend onderzoek dat een simpel tekst classificatiemodel heeft geïmplementeerd op beperkte data. Dit gezegd hebbende, zijn er toch hoopvolle resultaten waardoor tekstclassificatie een serieuze verbetering voor het analyseproces van K&R kan zijn in termen van accuraatheid en snelheid. Ook is er een basismodel gebouwd waarop kan worden voortgeborduurd en waaraan verbeteringen en aanpassingen gemaakt kunnen worden. Het beginnen en opzetten van een model is vaak een tijd- en arbeidsintensieve operatie, waarna de stappen naar een geavanceerder model makkelijker genomen kunnen worden. Deze scriptie voorziet K&R met deze basis.

De resultaten zijn bekend, maar het blijft onduidelijk in welke mate meer trainingsdata tot een verbetering van de classificaties leidt. Wel zijn er al een aantal verbeteringen doorgevoerd in het model wat betreft data voorbewerking en het aanpassen van parameters. Hierdoor werden de nauwkeurigheidsscore van de voorspellingen significant beter. Ook blijkt uit het experiment met de top 3 labels een hogere nauwkeurigheidsscore te komen, waarmee een genuanceerder beeld van de nauwkeurigheid van het algoritme naar voren komt.

De wetenschappelijke relevantie van dit onderzoek kan worden betrokken op het feit dat er weinig tot geen literatuur te vinden is over vergelijkbare tekst classificaties. De combinatie van relatieve kort tekstomschrijvingen zonder volledige zinnen met de vele labels in de database van K&R, maakt dit een uniek project. De praktisch relevantie van het onderzoek voor K&R zit in de opzet van een basismodel waarmee onderzocht is of tekst classificatie een manier is om het analyseproces te automatiseren. Er is een begin neergezet waarbij een deel van de omschrijvingen goed wordt voorspeld. Een eerste aanzet tot (gedeeltelijke) automatisering is hiermee gegeven.

6.2 Aanbevelingen

In de punten behandelt in deze sectie worden een aantal aanbevelingen gedaan:

- Een belangrijke, eerste stap om te nemen is het invoeren van meer trainingsdata. Meer trainingsdata levert met de juiste voorbewerking waarschijnlijk een classificator op dat langzaam de belangrijke woorden de juiste waarde gaat geven.
- Een ander idee om de classificator beter te laten presteren, is door de huidige analyses onder te verdelen in analyses die op elkaar lijken. Zo zijn er maar een klein aantal calculatieschema's in Nederland die door veel bedrijven worden gebruikt voor het opstellen van open begrotingen. Door de bedrijven die gebruik maken van hetzelfde calculatieschema te groeperen, kan het model alleen met die data worden getraind. Als er dan een nieuwe open begroting binnenkomt dat gebruik maakt van hetzelfde calculatieschema, kan het model worden aangeroepen dat alleen getraind is met die vergelijkbare data. Doordat de trainings- en testdata meer op elkaar zullen lijken, kunnen de prestaties van de classificator toenemen.

- Een derde aanbeveling is om te kijken naar technieken waarbij er enige afhankelijkheid tussen de woorden wordt geïmplementeerd. Deze methode wordt ook wel 'word embedding' (woord insluitingen) genoemd en bekende voorbeelden zijn Word2vec, fastText en GloVe.
- Ten slotte zijn er nog een aantal andere algoritmes die gedurende het onderzoek langs zijn gekomen als mogelijk goed werkende alternatieven voor het Naive Bayes model. Uit een kleine test kwamen de 'KNeighborsClassifier', 'DecisionTreeClassifier', 'RandomForestClassifier', 'GradientBoostingClassifier', 'LogisticRegression', 'SGDClassifier' en 'ComplementNB' naar voren als classificatoren die gelijkwaardig of iets minder scoorde dan het in deze scriptie gebruikte Multinomial Naive Bayes classificator.

In sectie 7.7 (pagina 44) in de appendix is op een aantal aanbevelingen wat dieper ingegaan. Ook worden daar nog een aantal praktische tips meegegeven waarin opgedane kennis van tijdens het onderzoek is verwerkt.

6.3 Vervolgonderzoek

Zoals aangegeven in sectie 6.1, is de kleine trainingsdataset nu nog een groot probleem. Dit onderzoek komt met een basismodel dat werkt, maar nog niet goed werkt. De grote hoeveelheid labels in combinatie met relatief korte omschrijvingen maakt dit een lastig project. Uit het literatuuronderzoek kwam al naar voren dat er weinig voorbeelden zijn van dit soort projecten met grote hoeveelheden labels. Wel kwam het vaker voor dat er niet hele tekstdocumenten of artikelen, maar ook kleinere berichten op internet werden geclassificeerd.

Het verbeteren van het model naast het opschalen van de trainingsdata kan bereikt worden door of de omschrijvingen langer/uitgebreider te maken, of door het aantal labels te verminderen. Lang niet alle 2500 labels komen op dit moment voor in de trainingsdata, maar naarmate er meer getraind wordt zal de hoeveelheid labels ook toenemen. Op deze beperkte trainingsdata gaat het aantal labels al richting de 1000. Het analyseren van het gebruik van deze labels is makkelijk te combineren met het vergroten van de trainingsdata. Python is in staat om snel en makkelijk veel voorkomende labels te vinden en te onderzoeken. Op basis van die informatie zou er een heroverweging van de database van K&R kunnen plaatsvinden. De zeer specifieke aard van de database maakt het een lastige dataset om een goed werkende classificator te krijgen, misschien wel een té lastige dataset. Uit deze scriptie is de conclusie te trekken dat er met meer trainingsdata er mogelijk toekomstmuziek zit in een tekst classificatie. Wil dat echter werkelijke succesvol gebeuren, zal de structuur van de database met haar specifieke labels moeten veranderen. Hoe dat precies ingevuld zou kunnen worden, is een interessante vraag voor een mogelijk vervolgonderzoek.

In de huidige situatie wordt er bij K&R al gewerkt met hoofdcodes en subcategorieën waarbinnen de labels vallen. Van deze subcategorieën, in K&R's database 'element varianten' genoemd, zijn er maar 260. Het juist voorspellen van deze 'element varianten' zou daarom al een stuk gemakkelijker moeten zijn. Dit is natuurlijk een simpele oplossing, waarbij er niet is nagedacht over hoe er wordt omgegaan met de prijzen die nu aan elk specifiek label zijn gelinkt. Die prijzen in de database zijn een 'unique selling point' van K&R, waardoor het dus geen goed idee is om daar afstand van te nemen. Een makkelijk oplossing voor het probleem van de grote aantallen labels ligt dus niet voor de hand en is daarmee absoluut een interessante hoek voor vervolgonderzoek. Een goede oplossing hiervoor kan ook enorme positieve gevolgen hebben voor de prestaties van een classificator.

Vervolgonderzoek zou dus in de richting kunnen gaan van het verminderen van klassen. K&R moet besluiten of zij hun database anders willen inrichten met minder labels en hoe ze dat willen aanpakken. Als dat niet gebeurd is er op basis van dit onderzoek weinig hoop op een zeer goed

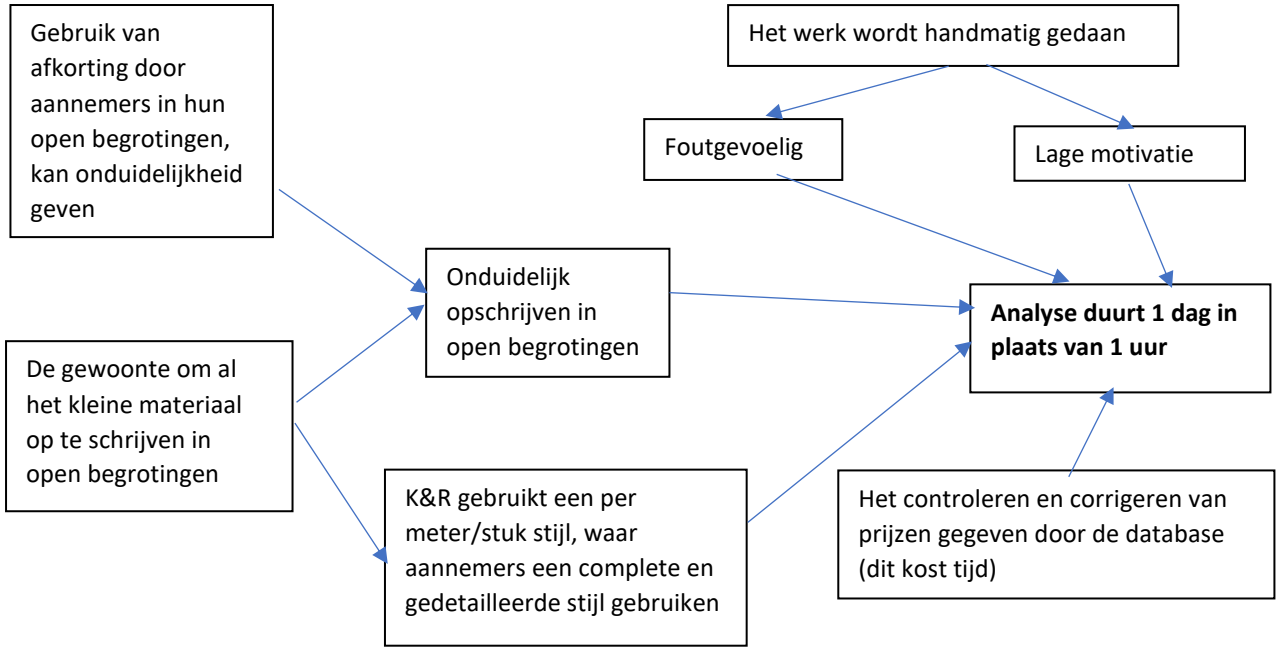
werkende classificator. Deze thesis geeft wel een werkend model waarin er ook bekeken kan worden welke labels veel voorkomen over meerdere analyses, waardoor nieuwe inzichten wat betreft de database verworven kunnen worden.

7 Appendix

7.1 Uitgebreide uitleg probleemkluwen

Om het kernprobleem te identificeren stellen we een lijst samen met alle problemen gerelateerd aan het actieprobleem van K&R. In deze lijst is samen met elk probleem de oorzaken van dat probleem erbij geschreven. De uitgebreide probleemkluwen is in

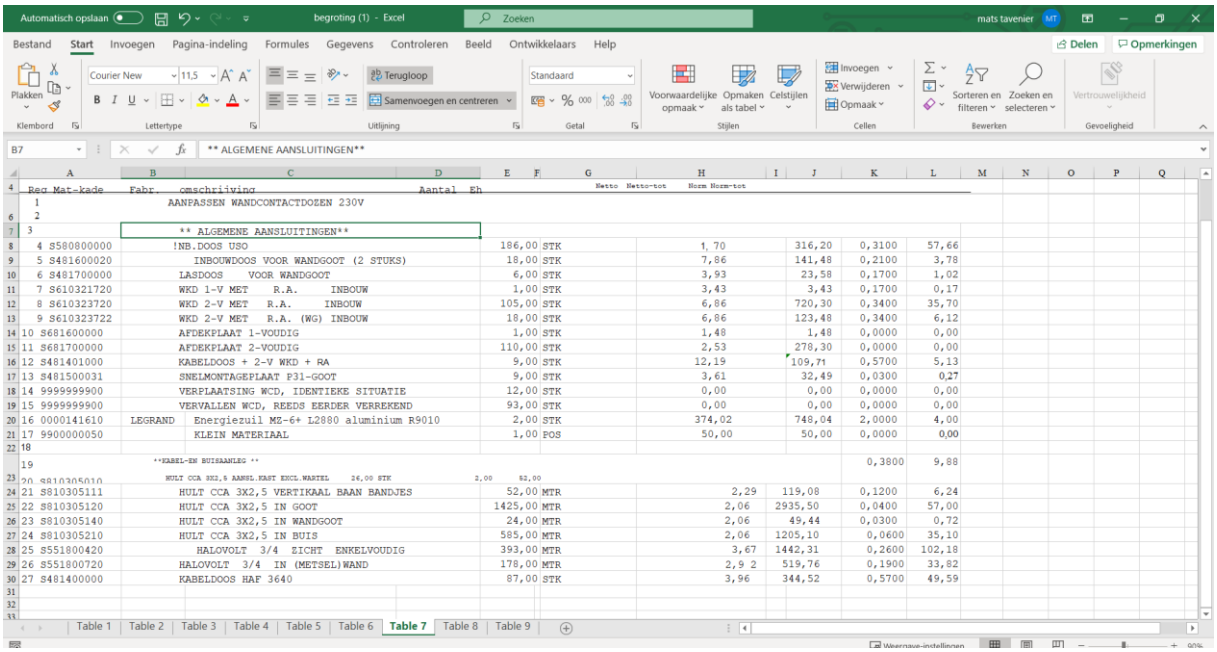
- *Onduidelijk opschrijven in open begrotingen.* Aannemers willen het soms juist onduidelijk maken, zodat klanten meer moeite hebben om de open begrotingen te begrijpen. Deze klanten hebben vaak niet de kennis om de open begrotingen te begrijpen en raken daar dan de weg in kwijt, waardoor aannemers soms kostenposten kunnen toevoegen die ze eigenlijk niet nodig hebben.
- *K&R gebruikt een per meter/stuk stijl, waar aannemers een complete en gedetailleerde stijl gebruiken.* K&R's doel is om de prijzen juist duidelijk te maken naar hun klanten, zodat die kunnen zien waar de kosten vandaan komen en waar het geld heengaat.
- *Gebruik van afkorting door aannemers in hun open begrotingen, kan onduidelijkheid geven.* Dit is makkelijker voor de aannemers en het gebruik van afkortingen is veelvoorkomend in dit veld.
- *De gewoonte om al het kleine materiaal op te schrijven in open begrotingen, ook al kost het bijna niets en voegt het geen nuttige informatie toe.* Onderdeel van het probleem door het verschil in gebruik van opschrijf stijlen. Maar het wordt ook gedaan met het doel om de open begrotingen zo onduidelijk mogelijk te maken.
- *Lage motivatie van de werknemers van K&R.* Het werk is soms zeer repetitief, omdat ze voor bijna elk project de open begrotingen vergelijkbaar moeten maken aan hun database. Veel materialen zitten standaard in een bouwproject en komen dus vaak terug.
- *Foutgevoelig.* Het werk is handmatig, dus is er kans op fouten.
- *Het werk wordt handmatig gedaan.* K&R werkt nog niet heel lang met deze database en hebben nog niet geprobeerd het te automatiseren.
- *Het controleren en corrigeren van prijzen gegeven door de database (dit kost tijd).* K&R doet te weinig analyses per jaar om een database te hebben die altijd de juiste markconforme prijzen geeft.
- *Het hele analyseproces van het vergelijken van een open begrotingen met de database duurt gemiddeld 1 dag in plaats van 1 uur.* De verschillende opschrijf stijlen van aannemers en K&R en het feit dat ze het vergelijkbaar maken handmatig moeten doen, vergt veel tijd.



Figuur 14 - Uitgebreide probleemkluwen

7.2 Geconverteerde open begroting naar Excel

Deze screenshot (Figuur 15 - Screenshot open begroting ExcelFiguur 15) heeft betrekking tot het stuk in sectie 2.1. Om goed te zien wat er soms veranderd in de opmaak, is er ook een screenshot van het pdf-bestand bij gezet (Figuur 16). In de screenshots is te zien hoe bepaalde kolommen zijn samengevoegd bij het converteren, zoals de 'Reg' en 'Mat-kode' kolommen en de 'Aantal' en 'Eh' kolommen. Dit veroorzaakt problemen wanneer het wordt overgezet in de tool van K&R, omdat je deze nu samengevoegde kolommen wel los wil hebben. Dit moet dus eerst handmatig aangepast worden, voordat het gekopieerd kan worden naar de tool.



Reg	Mat-kode	Fabr.	omschrijving	Aantal	Eh	Netto	Netto-tot	Norm	Norm-tot
1			AANPASSEN WANDCONTACTDOZEN 230V						
2									
3			** ALGEMENE AANSLUITINGEN**						
4	S580800000		!NB.DOOS U50	186,00	STK	1, 70	316,20	0,3100	57,66
5	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18,00	STK	7,86	141,48	0,2100	3,78
6	S481700000		LASDOOS VOOR WANDGOOT	6,00	STK	3,93	23,58	0,1700	1,02
7	S610321720		WKD 1-V MET R.A. INBOUW	1,00	STK	3,43	3,43	0,1700	0,17
8	S610323720		WKD 2-V MET R.A. INBOUW	105,00	STK	6,86	720,30	0,3400	35,70
9	S610323722		WKD 2-V MET R.A. (WG) INBOUW	18,00	STK	6,86	123,48	0,3400	6,12
10	S681600000		AFDEKPLAAT 1-VOUDIG	1,00	STK	1,48	1,48	0,0000	0,00
11	S681700000		AFDEKPLAAT 2-VOUDIG	110,00	STK	2,53	278,30	0,0000	0,00
12	S481401000		KABELDOOS + 2-V WKD + RA	9,00	STK	12,19	109,71	0,5700	5,13
13	S481500031		SNELMONTAGEPLAAT P31-GOOT	9,00	STK	3,61	32,49	0,0300	0,27
14	9999999900		VERPLAATSING WCD, IDENTIEKE SITUATIE	12,00	STK	0,00	0,00	0,0000	0,00
15	9999999900		VERVALLEN WCD, REEDS EERDER VERREKEND	93,00	STK	0,00	0,00	0,0000	0,00
16	0000141610	LEGRAND	Energiezuil MZ-6+ L2880 aluminium R9010	2,00	STK	374,02	748,04	2,0000	4,00
17	9900000050		KLEIN MATERIAAL	1,00	POS	50,00	50,00	0,0000	0,00
18									
19			**KABEL- EN BUISAANLEG **					0,3800	9,88
20	S810305010		HULT CCA 3X2,5 AANSL.KAST EXCL.WARTEL	26,00	STK				
21	S810305111		HULT CCA 3X2,5 VERTIKAAL BAAN BANDJES	52,00	MTR	2,29	119,08	0,1200	6,24
22	S810305120		HULT CCA 3X2,5 IN GOOT	1425,00	MTR	2,06	2935,50	0,0400	57,00
23	S810305140		HULT CCA 3X2,5 IN WANDGOOT	24,00	MTR	2,06	49,44	0,0300	0,72
24	S810305210		HULT CCA 3X2,5 IN BUIS	585,00	MTR	2,06	1205,10	0,0600	35,10
25	S551800420		HALOVOLT 3/4 ZICHT ENKELVOUDIG	393,00	MTR	3,67	1442,31	0,2600	102,18
26	S551800720		HALOVOLT 3/4 IN (METSEL)WAND	178,00	MTR	2,9 2	519,76	0,1900	33,82
27	S481400000		KABELDOOS HAF 3640	87,00	STK	3,96	344,52	0,5700	49,59

Figuur 15 - Screenshot open begroting Excel

Calculatie: 10470201/06112 Offertepost: 010 Heijmans Utiliteit 14.11.2019 Blad : 1

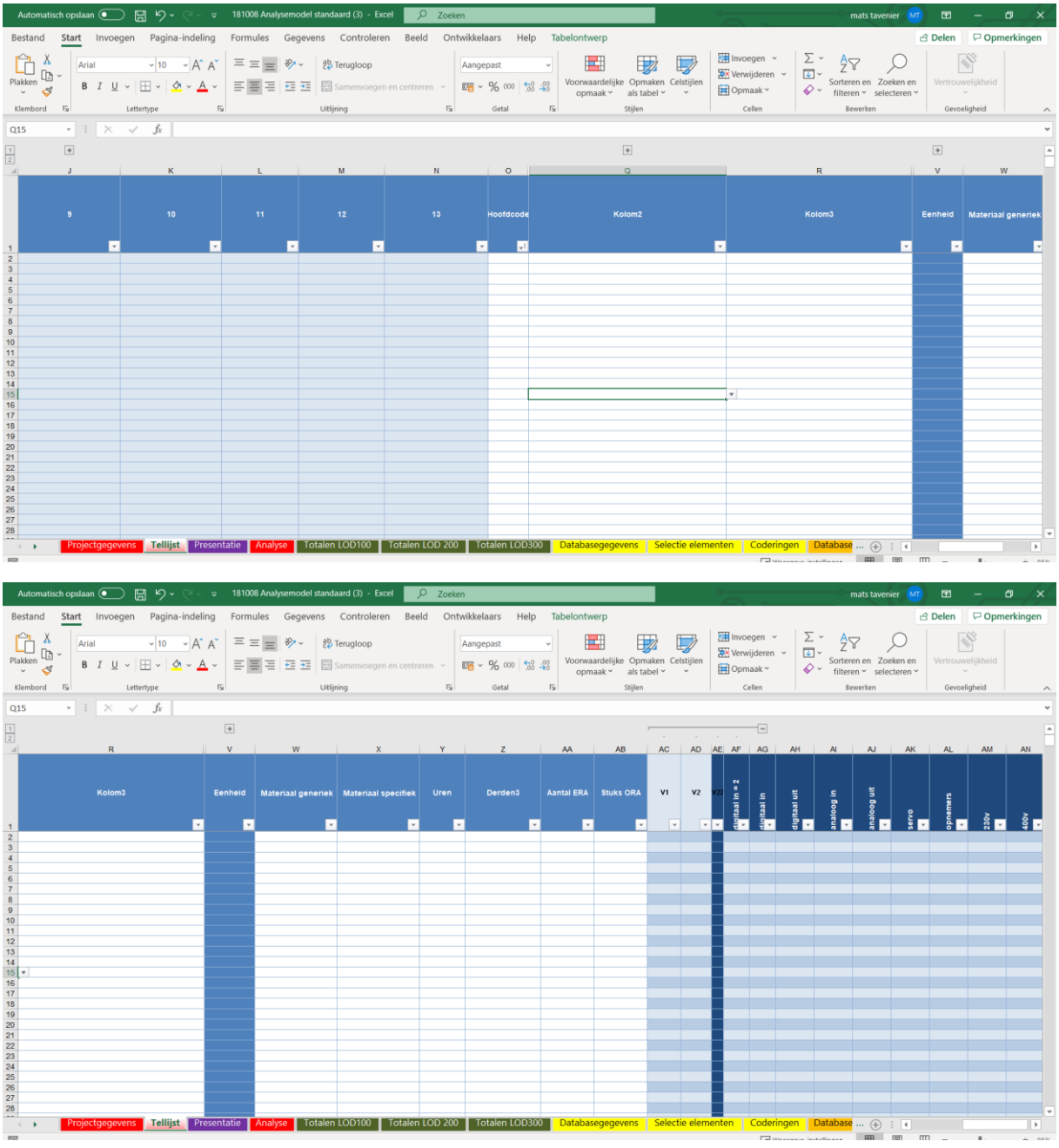
Offertepost : MW 112; WANDCONTACTDOZEN 230V
Valuta : EUR
Werk: HUIZE WESTERLICHT, ALKMAAR
Aanvrager:
Bestek: HUIZE WESTERLICHT ALKMAAR

Reg	Mat-kode	Fabr.	Omschrijving	Aantal	Eh	Netto	Netto-tot	Norm	Norm-tot
1			AANPASSEN WANDCONTACTDOZEN 230V						
2									
3			** ALGEMENE AANSLUITINGEN **						
4	S580800000		INB.DOOS U50	186,00	STK	1,70	316,20	0,3100	57,66
5	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18,00	STK	7,86	141,48	0,2100	3,78
6	S481700000		LASDOOS VOOR WANDGOOT	6,00	STK	3,93	23,58	0,1700	1,02
7	S610321720		WKD 1-V MET R.A. INBOUW	1,00	STK	3,43	3,43	0,1700	0,17
8	S610323720		WKD 2-V MET R.A. INBOUW	105,00	STK	6,86	720,30	0,3400	35,70
9	S610323722		WKD 2-V MET R.A. (WG) INBOUW	18,00	STK	6,86	123,48	0,3400	6,12
10	S681600000		AFDEKPLAAT 1-VOUDIG	1,00	STK	1,48	1,48	0,0000	0,00
11	S681700000		AFDEKPLAAT 2-VOUDIG	110,00	STK	2,53	278,30	0,0000	0,00
12	S481401000		KABELDOOS + 2-V WKD + RA	9,00	STK	12,19	109,71	0,5700	5,13
13	S481500031		SNELMONTAGEPLAAT P31-GOOT	9,00	STK	3,61	32,49	0,0300	0,27
14	9999999900		VERPLAATSING WCD, IDENTIEKE SITUATIE	12,00	STK	0,00	0,00	0,0000	0,00
15	9999999900		VERVALLEN WCD, REEDS EERDER VERREKEND	93,00	STK	0,00	0,00	0,0000	0,00
16	0000141610	LEGRAND	Energiezuil MZ-6+ L2880 aluminium R9010	2,00	STK	374,02	748,04	2,0000	4,00
17	9900000050		KLEIN MATERIAAL	1,00	POS	50,00	50,00	0,0000	0,00
18									
19			** KABEL- EN BUISAANLEG **						
20	S810305010		HULT CCA 3X2,5 AANSL.KAST EXCL.WARTEL	26,00	STK	2,00	52,00	0,3800	9,88
21	S810305111		HULT CCA 3X2,5 VERTIKAAL BAAN BANDJES	52,00	MTR	2,29	119,08	0,1200	6,24
22	S810305120		HULT CCA 3X2,5 IN GOOT	1425,00	MTR	2,06	2935,50	0,0400	57,00
23	S810305140		HULT CCA 3X2,5 IN WANDGOOT	24,00	MTR	2,06	49,44	0,0300	0,72
24	S810305210		HULT CCA 3X2,5 IN BUIS	585,00	MTR	2,06	1205,10	0,0600	35,10
25	S551800420		HALOVOLT 3/4 ZICHT ENKELVOUDIG	393,00	MTR	3,67	1442,31	0,2600	102,18
26	S551800720		HALOVOLT 3/4 IN (METSEL)WAND	178,00	MTR	2,92	519,76	0,1900	33,82
27	S481400000		KABELDOOS HAF 3640	87,00	STK	3,96	344,52	0,5700	49,59

Figuur 16 - Open begrotingen in PDF

7.3 Niet ingevulde tool uit sectie 2.2

De figuren hieronder zijn screenshots van een nog niet ingevulde tool. In sectie 2.2 wordt naar dit bestand gerefereerd.



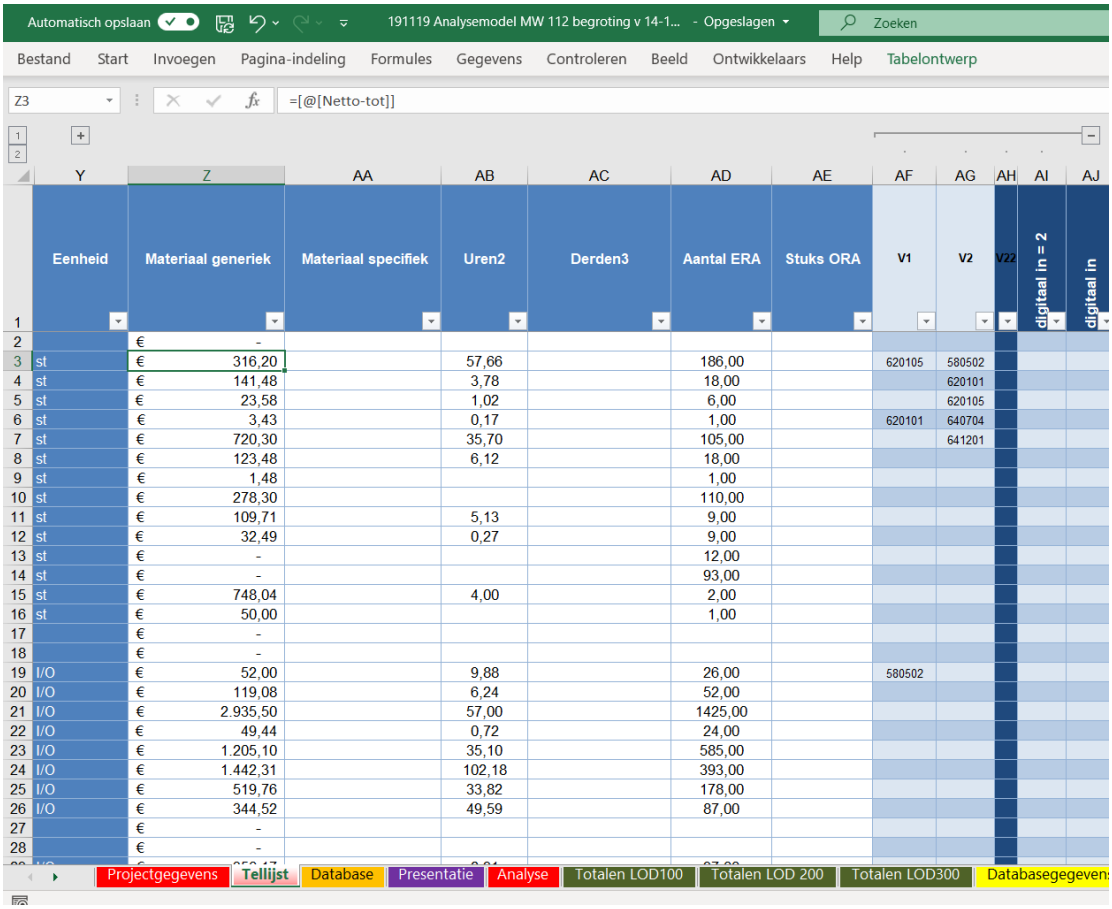
Figuur 17 - Voorbeeld van niet ingevulde Excel-tool

7.4 Aanpassen tool

Alle gegevens worden in de tool geplaats, waarbij er wordt gecheckt of de gegevens nog steeds overeenkomen met de informatie zoals weergeven in toegestuurde open begrotingen. Belangrijk is dat alle aantallen nog steeds onder het juiste kopje vallen, zodat er geen fouten worden gemaakt bij het optellen van prijzen en aantallen. Wanneer dit gecontroleerd is, wordt de rest van de tool ingevuld.

Om het overzichtelijk te houden, blijven we met hetzelfde voorbeeld werken als de vorige twee secties. Kolommen ‘Z’, ‘AB’ en ‘AD’ worden gelinkt aan een van de eerder ingevulde kolommen die te zien zijn in Figuur 3 in sectie 2.2. De namen van deze kolommen ‘Z’, ‘AB’ en ‘AD’ zijn respectievelijk ‘Materiaal generiek’, ‘Uren2’ en ‘Aantal ERA’. Deze kolommen zijn er om overzichtelijk te krijgen wat de prijzen, uren en aantallen zijn per kostenpost en in totaal. In dit geval zijn die kolommen gelinkt aan de kolommen ‘Netto-tot’, ‘Norm-tot’ en ‘Aantal’ (respectievelijk kolom ‘H’, ‘J’ en ‘E’). De kolommen ‘AA’, ‘AC’ en ‘AE’ zijn voor deze analyse niet aan de orde. Allen kolom ‘AC’ wordt in andere gevallen wel gebruikt. Deze kolom wordt ingevuld wanneer aannemers gebruiken maken van onderaannemers. Vaak gebruiken ze dan een extra kolom waarin ze de totaalprijs van die onderaannemer vermelden. Bij dit voorbeeld wordt er geen gebruik gemaakt van onderaannemers, dus wordt deze kolom ook niet ingevuld.

In Figuur 18 is te zien hoe het na het linken van de verschillende kolommen de tool eruitziet.

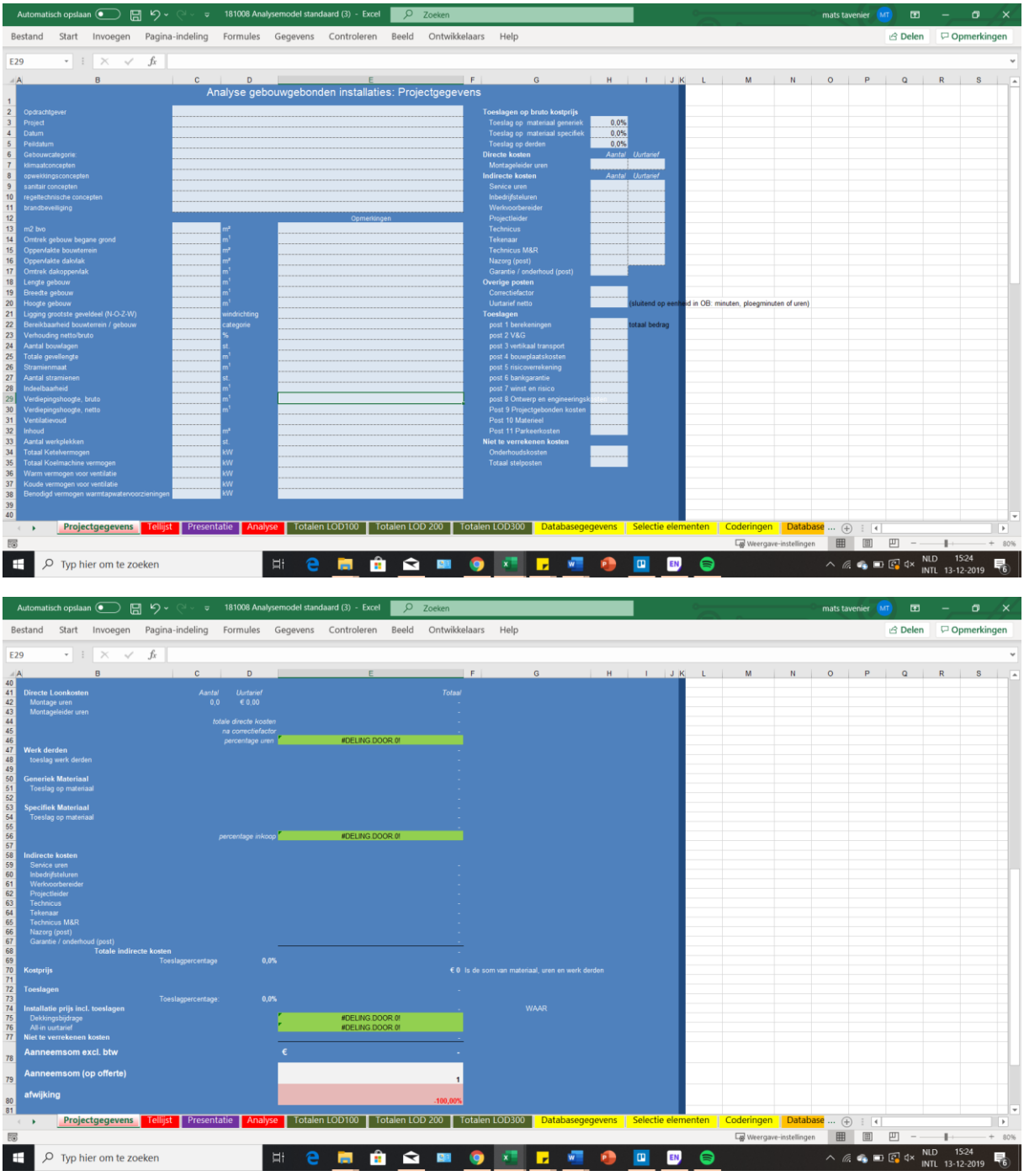


	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	Eenheid	Materiaal generiek	Materiaal specifiek	Uren2	Derden3	Aantal ERA	Stuks ORA	V1	V2	V22	digitaal in = 2	digitaal in
2		€ -										
3	st	€ 316,20		57,66		186,00		620105	580502			
4	st	€ 141,48		3,78		18,00			620101			
5	st	€ 23,58		1,02		6,00			620105			
6	st	€ 3,43		0,17		1,00		620101	640704			
7	st	€ 720,30		35,70		105,00			641201			
8	st	€ 123,48		6,12		18,00						
9	st	€ 1,48				1,00						
10	st	€ 278,30				110,00						
11	st	€ 109,71		5,13		9,00						
12	st	€ 32,49		0,27		9,00						
13	st	€ -				12,00						
14	st	€ -				93,00						
15	st	€ 748,04		4,00		2,00						
16	st	€ 50,00				1,00						
17		€ -										
18		€ -										
19	I/O	€ 52,00		9,88		26,00		580502				
20	I/O	€ 119,08		6,24		52,00						
21	I/O	€ 2.935,50		57,00		1425,00						
22	I/O	€ 49,44		0,72		24,00						
23	I/O	€ 1.205,10		35,10		585,00						
24	I/O	€ 1.442,31		102,18		393,00						
25	I/O	€ 519,76		33,82		178,00						
26	I/O	€ 344,52		49,59		87,00						
27		€ -										
28		€ -										

Figuur 18 - Kolommen 'Z' t/m 'AE' uit tabblad 'Tellijst' uit de Excel-tool

7.5 Tabblad ‘Projectgegevens’

De figuren hieronder zijn screenshots van een nog niet ingevuld tabblad ‘Projectgegevens’. In sectie 2.3 wordt hiernaar verwezen.

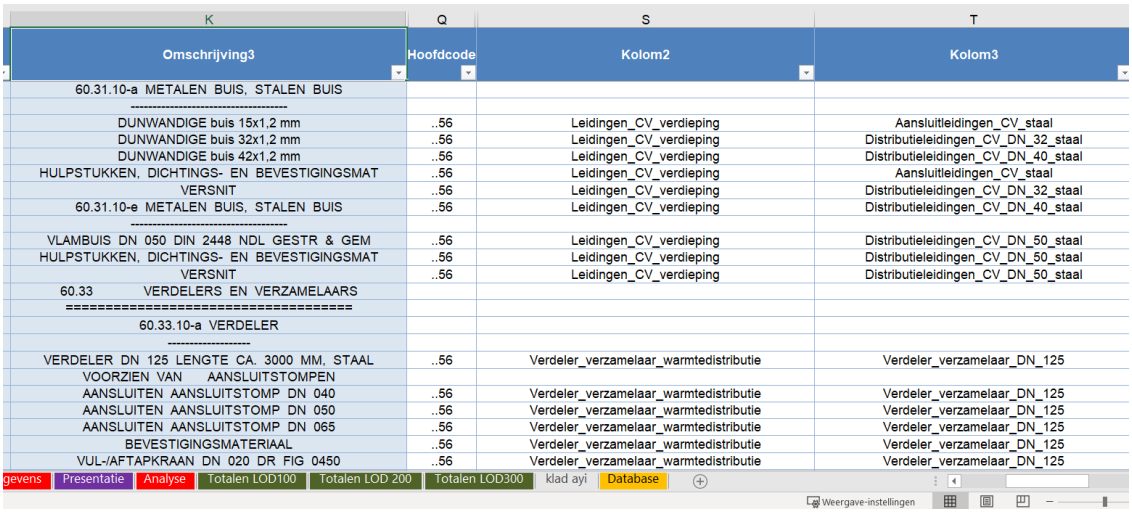


Figuur 19 - Nog niet ingevulde tabblad 'Projectgegevens'

7.6 Moeilijkheden preparen trainingsdata

In sectie 4.1 (pagina 23) wordt de trainingsdata voor het oplossingsontwerp besproken, waarbij een aantal moeilijkheden is voorgekomen. Deze moeilijkheden worden in deze appendix besproken.

Aan het prepareren van de trainingsdata hebben enkele haken en ogen gezeten. Zo is het niet mogelijk gebleken om snel meerdere Excel-analyses in een keer te importeren naar het Python-programma. Dit komt door een onregelmatigheid in het opslaan van de omschrijvingen en labels binnen de Excel-bestanden. In Figuur 20 en Figuur 21 is hier een voorbeeld van gegeven, waarbij er twee verschillende analyse-bestanden te zien zijn. In Figuur 20 zijn de omschrijvingen in kolom 'K' neergezet, terwijl in Figuur 21 de omschrijvingen in kolom 'A' te vinden zijn. Dit probleem doet zich ook voor bij de label-kolommen, waarbij de labels in respectievelijk kolom 'T' en R te vinden zijn.



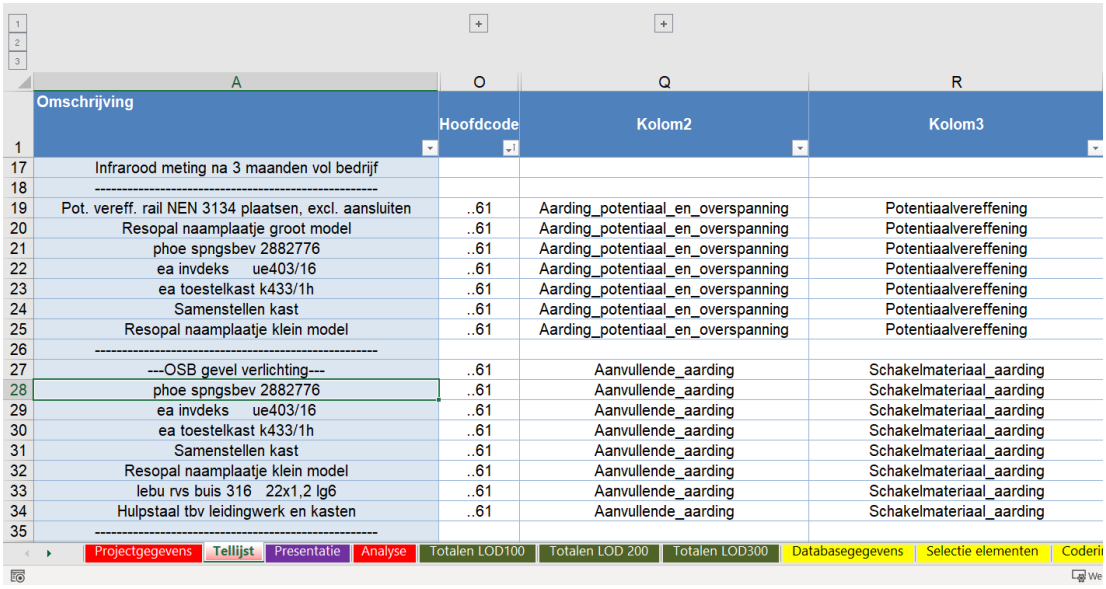
	K	Q	S	T
	Omschrijving3	Hoofdcode	Kolom2	Kolom3
	60.31.10-a METALEN BUIS, STALEN BUIS			

	DUNWANDIGE buis 15x1,2 mm	..56	Leidingen_CV_verdieping	Aansluitleidingen_CV_staal
	DUNWANDIGE buis 32x1,2 mm	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_32_staal
	DUNWANDIGE buis 42x1,2 mm	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_40_staal
	HULPSTUKKEN, DICHTINGS- EN BEVESTIGINGSMAT	..56	Leidingen_CV_verdieping	Aansluitleidingen_CV_staal
	VERSNIT	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_32_staal
	60.31.10-e METALEN BUIS, STALEN BUIS	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_40_staal
	----------			
	VLAMBUIS DN 050 DIN 2448 NDL GESTR & GEM	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_50_staal
	HULPSTUKKEN, DICHTINGS- EN BEVESTIGINGSMAT	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_50_staal
	VERSNIT	..56	Leidingen_CV_verdieping	Distributieleidingen_CV_DN_50_staal
	60.33 VERDELERS EN VERZAMELAARS			
	==========			
	60.33.10-a VERDELER			

	VERDELER DN 125 LENGTE CA. 3000 MM, STAAL	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125
	VOORZIEN VAN AANSLUITSTOMPEN			
	AANSLUITEN AANSLUITSTOMP DN 040	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125
	AANSLUITEN AANSLUITSTOMP DN 050	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125
	AANSLUITEN AANSLUITSTOMP DN 065	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125
	BEVESTIGINGSMATERIAAL	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125
	VUL-/AFTAPKRAAN DN 020 DR FIG 0450	..56	Verdeler_verzamelaar_warmtedistributie	Verdeler_verzamelaar_DN_125

gevens | Presentatie | Analyse | Totalen LOD100 | Totalen LOD 200 | Totalen LOD300 | klad ayi | Database

Figuur 20 - Voorbeeld 1 van omschrijvings- en labelkolom



	A	O	Q	R
1	Omschrijving	Hoofdcode	Kolom2	Kolom3
17	Infrarood meting na 3 maanden vol bedrijf			
18	----------			
19	Pot. vereff. rail NEN 3134 plaatsen, excl. aansluiten	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
20	Resopal naamplaatje groot model	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
21	phoe spngsbev 2882776	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
22	ea invdeks ue403/16	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
23	ea toestelkast k433/1h	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
24	Samenstellen kast	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
25	Resopal naamplaatje klein model	..61	Aarding_potentiaal_en_overspanning	Potentiaalvereffening
26	----------			
27	---OSB gevel verlichting---	..61	Aanvullende_aarding	Schakelmateriaal_aarding
28	phoe spngsbev 2882776	..61	Aanvullende_aarding	Schakelmateriaal_aarding
29	ea invdeks ue403/16	..61	Aanvullende_aarding	Schakelmateriaal_aarding
30	ea toestelkast k433/1h	..61	Aanvullende_aarding	Schakelmateriaal_aarding
31	Samenstellen kast	..61	Aanvullende_aarding	Schakelmateriaal_aarding
32	Resopal naamplaatje klein model	..61	Aanvullende_aarding	Schakelmateriaal_aarding
33	lebu rvs buis 316 22x1,2 lg6	..61	Aanvullende_aarding	Schakelmateriaal_aarding
34	Hulpstaal tbv leidingwerk en kasten	..61	Aanvullende_aarding	Schakelmateriaal_aarding
35	----------			

Projectgegevens | Tellijst | Presentatie | Analyse | Totalen LOD100 | Totalen LOD 200 | Totalen LOD300 | Databasegegevens | Selectie elementen | Coderi

Figuur 21 - Voorbeeld 2 van omschrijvings- en labelkolom

Er is in deze Excel-bestanden geen regelmaat te vinden, wat het Python-programma in moeilijkheden brengt. Vanuit Python kunnen er specifieke kolommen uit specifieke Excel-tabbladen geïmporteerd worden. Daarnaast is het ook mogelijk om meerdere Excel-bestanden in een keer te importeren, maar dit werkt alleen goed als de omschrijvingen en labels in elk Excel-bestand in dezelfde kolom

geschreven staan. Bij K&R is dit niet het geval, waardoor elk Excel-bestand apart geïmporteerd moet worden. Dit brengt een aantal problemen met zich mee.

Ten eerste kost dit veel tijd, want voor het importeren van elk bestand moet het Python-programma worden aangepast op de juiste kolommen van een specifiek Excel-bestand. Ook levert dit apart importeren van elk Excel-bestanden een tweede probleem op. In Python is het opslaan van het model relatief makkelijk te coderen, maar het voortdurend updaten van het model is een volgende stap waarin veel extra tijd gestoken zou moeten worden. Deze twee redenen hebben geleid naar de volgende relatief eenvoudige en snelle oplossing. Er zijn 30 analyse-bestanden gekozen, waarbij van elke analyse de omschrijvings- en labelkolom zijn gekopieerd naar een ander Excel-bestand. In dit trainingsbestand zijn dus alle omschrijvingen en labels in dezelfde kolom geplakt, waardoor dit bestand in een keer geïmporteerd kan worden. Ook hoeft het Python-programma maar een keer het model op te slaan in plaats van het voortdurend updaten.

De reden dat er maar voor 30 analyse-bestanden gekozen is, komt door het tijdsintensieve proces van het preparen van de data. Door de beperkingen van de Corona-maatregelen moesten alle werkzaamheden thuis worden uitgevoerd, waarbij alle Excel-bestanden door een enkele laptop moesten worden behandeld. Geschikte Excel-bestanden moesten via OneDrive worden uitgezocht, waarna de Excel-bestanden met in elk bestand de volledige database van K&R, gedownload moest worden. De OneDrive crashte regelmatig, wat samen met de lange download- en opentijd van de Excel-bestanden leidde tot een proces waarbij alleen het prepareren van deze 30 bestanden al meerdere dagen kostte.

7.7 Praktische tips

7.7.1 'Word embedding'

In het model dat gepresenteerd is in deze scriptie, wordt er bij beide experimenten uitgegaan van onafhankelijkheid tussen alle 'features' of woorden. In plaats van dat er alleen gekeken wordt naar woordfrequenties, is het ook mogelijk om enige afhankelijkheid in de woorden te implementeren. Deze methode wordt 'word embedding' (woord insluitingen) genoemd en bekende voorbeelden zijn Word2vec, fastText en GloVe. Het idee is dat bij het vectoriseren van de data de woordvectoren worden opgeslagen waarin patronen kunnen worden gevonden. Hierdoor kan er meer relevante informatie uit de tekst worden gehaald.

De oudste methode is Word2Vec, waarbij een 'deep neural network' wordt gebruikt om patronen te ontdekken. De basis van deze methode draait om het idee dat elk woord zijn eigen omliggende woorden kan voorspellen. Het model is al wat ouder, maar lijkt nog steeds goed te werken. De tweede methode is fastText, waarbij er informatie wordt gehaald uit de sub-woorden binnen de al bekende woorden. Deze methode reageert goed op woorden in de testdata die nog niet bekend waren in de trainingsdata. De derde methode GloVe haalt informatie uit het samen voorkomen van verschillende woorden binnen de trainingsdata. Hoe meer bepaalde woorden samen voorkomen, hoe hoger de kans op een relatie tussen de twee woorden wordt bepaald.

Van deze drie methodes zijn voorgetrainde technieken, maar het is ook mogelijk om de 'word embedding' zelf te trainen op de trainingsdata. Het verschilt per tekst classificatie project wat het beste werkt. Het gebruik van 'word embedding' voorkomt het gebruik van een 'document-term matrix', wat veel geheugen kost. Een ander groot voordeel bij deze methode van vectoriseren is dat de woorden die op elkaar lijken vergelijkbare vectoren krijgen, waardoor patronen makkelijker herkent kunnen worden.

7.7.2 Alternatieven modellen

Deze scriptie is gefocust op het Naive Bayes model vanwege de snelle en relatief eenvoudige implementatie. Er is veel data beschikbaar en andere tekst classificator algoritmes kunnen daar erg traag van worden. Wel kan er verder gekeken worden naar andere modellen en waar nodig geïnvesteerd worden door K&R. Uit een aantal kleine testen die wel konden worden uitgevoerd door de beschikbare computers kwamen KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, SGDClassifier en ComplementNB naar voren als classificatoren die gelijkwaardig of iets minder scoorde dan de Multinomial Naive Bayes classificator. Dit waren tests zonder aanpassingen of opschonen van data, waardoor er op dat vlak mogelijk ook nog winst te behalen is.

Wanneer er met meer data getraind kan worden, blijft het interessant om andere classificatoren uit te proberen. Door het aanpassen van parameters kunnen deze classificatoren misschien wel beter presteren dan het huidige model. Een ander classificatie algoritme dat vaak qua scores redelijk dicht bij Naive Bayes schijnt te liggen, is het 'lineare discriminant analysis' (LDA). Of dat voor de dataset van K&R ook geldt, is nog niet onderzocht.

Bij tekst classificatie wordt vaak toch geprobeerd om modellen zo simpel mogelijk te houden, omdat die modellen het beste werken. De voorgestelde classificatoren zijn net als Naive Bayes nog steeds relatief simpel en zijn daarom goede alternatieven.

7.7.3 Extra aanbevelingen

Naast de conclusies en voorstellen van vervolgonderzoek, zijn er ook nog een aantal kleinere aanbevelingen die gemaakt kunnen worden. Dit zijn de praktische tips waar binnen de kaders van deze scriptie niet aan toegekomen is, maar wel mogelijk toekomst in zit.

Het importeren van data naar Python is een van de knelpunten binnen deze scriptie en er zijn een aantal mogelijkheden om dit te verbeteren. Zo lijkt het mogelijk om meerdere Excel-bestanden tegelijk te importeren naar een Python-programma, mits de data binnen de Excel-bestanden in dezelfde kolommen staan. Ook is het mogelijk om rechtstreeks PDF-bestanden te importeren naar Excel. PDF is vaak de vorm waarin de open begrotingen worden aangeleverd aan K&R. Als dit goed zou werken, hoeven de open begrotingen niet meer geconverteerd naar Excel te worden. Onduidelijk is wel hoe accuraat de data transitie van PDF naar Python gaat.

Een ander probleem in Python waartegen aan is gelopen, is het updaten van het classificatie model met extra trainingsdata. Hoe het updaten van het model met nieuwe trainingsdata precies in zijn werk gaat, is helaas niet gevonden. Wel lijkt de functie 'partial_fit' hierbij te kunnen helpen. Om door te bouwen op het basismodel zoals dat er nu staat, is het verstandig om te verdiepen in de werking van deze functie en de mogelijke toepassing ervan.

Een andere methode op het gebied van tekst classificatie is 'lemmatization'. Deze methode lijkt op 'stemming' waarbij woorden worden teruggebracht naar hun stam. 'Lemmatization' lijkt misschien zelfs beter te werken dan 'stemming', maar er zijn helaas geen officiële bibliotheken beschikbaar voor de Nederlandse taal. Wel blijft het een methode dat erg interessant kan zijn wanneer dit wel ooit beschikbaar wordt.

Naast het tellen van woordfrequenties en zoeken van patronen in tekst, zijn er nog andere manieren om 'features' te selecteren. Deze technieken zijn meer experimenteel, maar zouden in ieder geval uitgeprobeerd kunnen worden. Voorbeelden hiervan zijn karakter, punctuatie, hoofdletter of zelfstandig naamwoord frequenties tellen. Deze technieken zijn waarschijnlijk een wilde gok, maar het kan voorkomen dat een van deze technieken verbazingwekkend goed werkt.

8 Referenties

1. Kao A, Poteet SR. Natural language processing and text mining2007. 1-265 p.

2. Jo T. Text mining : concepts, implementation, and big data challenge. Cham, Switzerland: Springer; 2019. Available from: https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1827895

3. Sharda R, Delen D, Turban E. Business intelligence : a managerial perspective on analytics. Global edition. Third edition. ed. Harlow: Pearson Education Limited; 2014.

4. Feldman R, Ronen, Sanger, James. The text mining handbook: Advanced approaches in analyzing unstructured data2007.

5. Aggarwal CC. Machine learning for text. Cham, Switzerland: Springer; 2018. Available from: https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5589130

6. Fanny F, Yohan M, Fidelson T. A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification. Jurnal Informatika: Jurnal Pengembangan IT [Internet]. 2018; 3(2):[157-60 pp.].

7. Sharma N, Singh M, editors. Modifying Naive Bayes classifier for multinomial text classification. 2016 International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2016; 2016.

8. Zhang Z. Naïve bayes classification in R. Annals of Translational Medicine. 2016;4(12):1-5.

9. Kim SB, Rim HC, Yook DS, Lim HS. Effective methods for improving naive Bayes text classifiers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)2002. p. 414-23.

10. Jiang L, Cai Z, Zhang H, Wang D. Naive Bayes text classifiers: A locally weighted learning approach. Journal of Experimental and Theoretical Artificial Intelligence. 2013;25(2):273-86.

11. Zhang L, Jiang L, Li C, Kong G. Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems. 2016;100:137-44.

12. Kumar Y, Sahoo G. Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers. International Journal of Information Technology and Computer Science. 2013;5:57-64.

13. McKinney W. Python for data analysis. Sebastopol, CA: O'Reilly Media; 2013. Available from: http://swbplus.bsz-bw.de/bsz378290444cov.htm