Titel TBD.

Optionele ondertitel.

Thomas Vanderveken.

Scriptie voorgedragen tot het bekomen van de graad van
Professionele bachelor in de toegepaste informatica

Promotor: TBD
Co-promotor: TBD
Academiejaar: 2023–2024
Derde examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Samenvatting

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Lijst van codefragmenten

1

Inleiding

Regelgevende filings die door institutionele beleggers, zoals pensioenfondsen en vermogensbeheerders worden ingediend, bieden belangrijke inzichten in marktpatronen en beleggingsstrategie. De 13F-meldingen die worden ingediend bij de Amerikaanse Securities and Exchange Commission (SEC) zijn een van de belangrijkste bronnen van deze informatie. Deze registraties geven informatie weer over de bezittingen van institutionele beleggers, waardoor ze van cruciaal belang zijn voor het uitvoeren van financieel onderzoek en het analyseren van beleggingen. 13F-meldingen van voor 2013 leveren echter aanzienlijke problemen op vanwege hun variabele vormen en structuren, die de menselijke verwerking en analyse complexer maken. De opkomst van geavanceerde AI-technologie biedt een potentiële kans om deze problemen aan te pakken. Natural Language Processing (NLP) en Machine Learning (ML) bieden geavanceerde technieken aan voor het extraheren en organiseren van gegevens uit tekst zonder vooraf gedefinieerde structuur. Door gebruik te maken van deze technologieën is het mogelijk om het proces van het standaardiseren en combineren van eerdere 13F meldingen in een goed georganiseerde relationele databank te automatiseren, waardoor de toegang en het gebruik wordt verbeterd.

Het doel van dit proefschrift is het creëren van een proof-of-concept toepassing die Natural Language Processing (NLP) en Machine Learning (ML) technieken gebruikt om 13F aanvragen van voor 2013 te standaardiseren en te integreren in een relationele databank. Het voorgestelde systeem is gericht op het stroomlijnen van het gegevens extractieproces door de meldingen automatisch om te zetten in een gestandaardiseerd formaat met een hoge efficiëntie en nauwkeurigheid. Dit zou niet alleen de analyse van financiële gegevens uit het verleden optimaliseren, maar ook het werk en de kosten verminderen die gepaard gaan met handmatige gegevensverwerking.

Bovendien zouden de gestandaardiseerde gegevens het begrip van investerings-

patronen uit het verleden verbeteren en het creëren van voorspellingsmodellen ondersteunen. Het onderzoek zal beginnen met een uitgebreide literatuurstudie om de meest efficiënte Natural Language Processing (NLP) en Machine Learning (ML) strategieën voor dit specifiek doel te bepalen. Daarna zal een proof-of-concept toepassing worden gecreëerd en beoordeeld worden op nauwkeurigheid, efficiëntie en bruikbaarheid.
De inleiding geeft een beknopt overzicht van de redenen, doelen en het belang van het onderzoek weer. Dit werk wil een nuttige bijdrage leveren aan de analyse van financiële gegevens en onderzoekers en analisten een nuttig hulpmiddel bieden door de moeilijkheden aan te pakken die gepaard gaan met het verwerken van oudere 13F-meldingen.

1.1. Probleemstelling

13F meldingen van de SEC voor 2013, zijn belangrijke bestanden voor financieel onderzoek, ze bevatten namelijk data over de stocks dat investment managers beheren. Maar deze zijn vaak inconsistent in opmaak en moeilijker toegankelijk, wat manuele analyse bemoeilijkt. Er ontbreekt namelijk een geautomatiseerd systeem om deze gegevens te standaardiseren en in een databank te integreren. Dit bemoeilijkt de opportuniteiten voor diepgaande analyses en het verkrijgen van inzichten in beleggingstrends.

1.2. Onderzoeksvraag

Hoe kunnen AI-technologieën zoals Natural Language Processing (NLP) en Machine Learning (ML) effectief worden toegepast om 13F-meldingen van de SEC van vóór 2013 te standaardiseren en te integreren in een gestructureerde databank, zodat de historische gegevens efficiënter kunnen worden geanalyseerd en vergeleken?

1.3. Onderzoeksdoelstelling

Het hoofddoel van dit onderzoek is het ontwikkelen van een geautomatiseerde methode die gebruikmaakt van AI-technologieën, zoals NLP en ML, om de data uit de 13F meldingen van voor 2013 te extraheren, standaardiseren en te integreren in een relationele databank. Dit moet leiden tot een efficiëntere en meer accurate extractie van gegevens uit deze documenten, waardoor de toegankelijkheid en bruikbaarheid van de data voor financieel onderzoek en investeringsanalyse aanzienlijk worden verbeterd.

1.4. Opzet van deze bachelorproef

Het verdere verloop van deze bachelorproef is opgebouwd als volgt:

In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.
In Hoofdstuk 4 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.
In Hoofdstuk 4 wordt de proof-of-concept besproken. De inhoud omvat de ingewikkelde technische specificaties, structuur en tools, samen met de functionele elementen zoals de modellen en de databank.
In Hoofdstuk 5, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2

Stand van zaken

De Securities and Exchange Commission (SEC) vereist dat institutionele vermogensbeheerders een kwartaalrapport indienen dat bekend staat als Form 13F als ze zeggenschap hebben over $100 miljoen of meer in sectie 13(f) effecten. Sectie 13(f) van de Securities Exchange Act van 1934 verplicht de openbaarmaking van effectenbezit door grote institutionele beleggers om de transparantie te vergroten. In 1975 implementeerde het Congres deze bepaling om de toegankelijkheid van informatie over de investeringsactiviteiten van deze bedrijven te verbeteren. De bedoeling was om het vertrouwen van beleggers in de integriteit van de effectenmarkten in de Verenigde Staten te vergroten door middel van een openbaarmakingsprogramma Securities en Commission (2023). Melding 13F biedt een uitgebreid overzicht van de aandelenbeleggingen van S&P 500 bedrijven en is een zeer belangrijk hulpmiddel voor analisten, onderzoekers en beleggers die inzicht willen verkrijgen in markttrends en de beleggingsbenaderingen van belangrijke marktspelers. Het onverwerkte tekstformaat waarin deze inzendingen worden aangeleverd, vormt echter een aanzienlijke belemmering voor effectieve gegevensextractie en -analyse, vooral voor inzendingen van voor 2013. Voor 2013 ontbrak het bij 13F-meldingen vaak aan standaardisatie en systematische opmaak, wat nu wel gebruikelijk is bij recentere aanmeldingen. Kunstmatige intelligentie (AI) en Machine Learning (ML) technologieën hebben de extractie en organisatie van gegevens uit ongestructureerde tekst de afgelopen jaren aanzienlijk veranderd. Geavanceerde methodologieën zoals Natural Language Processing (NLP) en Deep Learning (DL) modellen vergemakkelijken de omzetting van tekstuele 13F meldingen in gestructureerde datasets die geschikt zijn voor grondige analyse en studie. Standaardisatie is cruciaal voor historische gegevens, omdat het ontbreken van uniformiteit geautomatiseerde verwerking kan bemoeilijken. Door gebruik te maken van deze technologieën kunnen we zowel huidige als oudere 13F aanvragen omzetten in georganiseerde gegevens, die vervolgens kunnen worden opgeslagen in databanken,

waardoor patronen eenvoudiger kunnen worden opgehaald, gevisualiseerd en geanalyseerd.

Het doel van deze literatuurstudie is het onderzoeken en beoordelen van de verschillende Artificial Intelligence (AI) en Machine Learning (ML) technieken die kunnen worden gebruikt om gegevens uit 13F-meldingen van voor 2013 te extraheren, te organiseren en op te slaan. Het doel van het onderzoek is het bepalen van de meest efficiënte methoden om de ongeorganiseerde inhoud van deze documenten om te zetten in een gestructureerd formaat dat geschikt is voor analyse en opslag in een database. Dit houdt in dat er een onderzoek wordt gedaan naar verschillende kunstmatige intelligentie methodologieën, zoals Natural Language Processing (NLP) en Text mining, en dat bepaalde tools zoals NLTK en SpaCy worden geëvalueerd. De literatuurstudie zal ook de integratie van gestructureerde gegevens in een Database Management System (DBMS) onderzoeken, om te garanderen dat de geëxtraheerde gegevens gemakkelijk beschikbaar zijn voor later onderzoek en analyse. Het doel van deze evaluatie is om een uitgebreide kennis te krijgen van de meest effectieve procedures en technologie voor het verwerken van 13F-meldingen.

2.1. Wat zijn 13F meldingen

Het doel van deze sectie is om een beknopte inleiding te geven aan 13F-meldingen, met bijzondere aandacht voor de structuur van deze meldingen, de informatie die ze bevatten en de reden van hun bestaan.

2.1.1. Definitie en doel

Volgens (Securities & Commission, 2023) zijn 13F-meldingen verplichte wettelijke documenten die de Amerikaanse Securities and Exchange Commission (SEC) vereist onder Sectie 13(f) van de Securities Exchange Act van 1934. Deze deponeringen worden gebruikt om de portefeuilles van institutionele beleggingsbeheerders te rapporteren.
Het belangrijkste doel van 13F meldingen is om duidelijkheid en openheid te bieden over de beleggingsactiviteiten van belangrijke institutionele beleggers. Deze vereiste vergemakkelijkt het toezicht op beleggingsposities van verschillende instellingen, zoals beleggingsfondsen, pensioenfondsen en andere belangrijke beleggingsbeheerders, door het publiek en regelgevende instanties.

2.1.2. Belangrijke kenmerken

Het doel van dit deel is het bespreken van enkele kenmerken van de 13F-meldigen waaronder wie het moet indienen en wat ze moeten inhouden.

Vereisten voor rapportage:

Institutionele beleggingsbeheerders die minimaal $100 miljoen aan beheerd vermogen beheren, zijn verplicht om elk kwartaal een 13F-melding in te dienen. Deze rapporten moeten gedetailleerde informatie bevatten over de aandelenportefeuille van de instelling. Dit omvat onder andere de naam van het aandeel, het CUSIP-nummer, het aantal aandelen dat wordt gehouden, en de marktwaarde ervan.

Omvang van de informatie:

De rapportages over het aandelenbezit van grote beleggingsinstellingen richten zich voornamelijk op aandelen, terwijl andere soorten activa, zoals obligaties, derivaten en private equity, buiten beschouwing worden gelaten. Elk kwartaalrapport biedt een beknopt overzicht van de aandelenportefeuille van de instelling aan het einde van de rapportageperiode. Dit overzicht geeft waardevolle inzichten in de investeringsstrategieën en methoden die de instelling hanteert, wat bijdraagt aan de transparantie en begrip van hun beleggingsbenadering.

Opmaak en toegankelijkheid:

De 13F-meldingen van voor 2013 zijn varierend in opmaak, dit is wat data extractie moeilijk maakt Dit zijn enkele afbeeldingen van enkele van de tienduizenden 13F-meldingen

Voor 2013

Hier ziet men een aantal voorbeelden van informatie tabel van enkele 13F meldigen. Zoals men ziet zijn er veel verschillende manieren waarop deze meldingen zijn ingediend.

FORM 13F INFORMATION TABLE

COLUMN 1	COLUMN 2	COLUMN 3	COLUMN 4	COLUMN 5			COLUMN 6	COLUMN 7	COLUMN 8		
NAME OF ISSUER	TITLE OF CLASS	CUSIP	VALUE (x$1000)	SHRS OR PRN AMT	SH/ PRN	PUT/ CALL	INVESTMENT DISCRETION	OTHER MANAGERS	VOTING AUTHORITY SOLE	SHARED	NONE
<S>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>
Alcatel-Lucent	SPONSORED ADR	013904305	475	291,367	SH		SOLE	0	0	291,367	0
Alcoa Inc	Common	013817101	309	35,283	SH		SOLE	0	0	35,283	0
Anadarko Pete Corp	Common	032511107	2,458	37,132	SH		SOLE	0	0	37,132	0
Apache Corp	Common	037411105	1,172	13,330	SH		SOLE	0	0	13,330	0
Apple, Inc	Common	037833100	339	580	SH		SOLE	0	0	580	0
ARM HLDGS PLC	SPONSORED ADR	042068106	1,400	58,855	SH		SOLE	0	0	58,855	0

Figuur 2.1: Een correcte 13F melding duidelijk gestructureerd en geen missende waarden

									Voting Authority		
Name of Issuer	Title of class	CUSIP	Value (x$1000)	Shares/ Prn Amt	Sh/ Prn	Put/ Call	Invstmt Dscretn	Other Managers	Sole	Shared	None
<S>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>
1ST SOURCE CORP	Common Stock	336901103	306	13,531	SH		Sole		13,531		
ACE LTD	Common Stock	H0023R105	247	3,332	SH		Sole		3,332		
ACTIVISION BLIZZARD INC	Common Stock	00507V109	268	22,349	SH		Sole		22,349		
AECOM TECHNOLOGY CORP	Common Stock	00766T100	229	13,945	SH		Sole		13,945		

Figuur 2.2: Een 13F melding met missende waarden

NAME OF ISSUER	TITLE OF CLASS	CUSIP	VALUE (x$1000)	SHRS/ PRN AMT	SH/ PRN	PUT/INVSTMNT CALL DSCRETN	VOTING AUTHORITY SOLE	SHARED	NONE
<S>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>
ABN AMRO HOLDING NV ADR	Common	937102	27486	1497900	SH	SOLE	1424355		73545
ABN AMRO HOLDING NV ADR	Common	937102	322	17530	SH	UNKNOWN	17530		
ACMAT CORP CLASS A	Common	4616207	490	51890	SH	SOLE	51890		
AKZO NOBEL NV SPONSORED ADR	ADR	10199305	31141	752646	SH	SOLE	721408		31238
AKZO NOBEL NV SPONSORED ADR	ADR	10199305	84	2030	SH	UNKNOWN	2030		

Figuur 2.3: Een 13F melding met missende waarden en één entry overspant minstens één rij

SECURITY DESCRIPTION <S>	CLASS <C>	CUSIP <C>	SHARES <C>	MARKET VALUE <C>	SOLE (A) <C>	SHARED (B) <C>	SHARED OTHER (C) <C>	MGR <C>	VOTING AUTHORITY SOLE (A) <C>	SHARED (B) <C>	NONE (C) <C>
Ade Corporation	COM	00089C107	212,950	2,438	X				201,900	0	11,050
AVX, Corp.	COM	002444107	1,498,286	24,467	X				1,402,846	0	95,440
Aegis Realty Inc.	COM	00760P104	7,200	81	X				7,200	0	0

Figuur 2.4: Een 13F melding gebruik makend van 'X' in plaats van Bv. Sole

FORM 13F INFORMATION TABLE Name of Issuer	Title of Class	Cusip	Value (X1000)	Shs of SH/PRN Prn amt	Discrtn	Put/call	Voting Authority	
AFLAC Common	COM	001055102	1,427	33,500	SH		SOLE	SOLE
Abbott Labs Common	COM	002824100	1,441	22,350	SH		SOLE	SOLE
Apache Corp Common	COM	037411105	670	7,620	SH		SOLE	SOLE
Apple Inc Common	COM	037833100	1,170	2,003	SH		SOLE	SOLE
Becton Dickinson Common	COM	075887109	206	2,750	SH		SOLE	SOLE

Figuur 2.5: Een 13F melding zonder cijfer datas in de laatste kolommen

	<C>						<C>			
NAME OF ISSUER	TITLE OF CLASS	CUSIP	VALUE(K)	SH/P	AMT	S/P	P/C	INV DSC	MANAGERS	SOLE	SHARED	NONE
ABBOTT LABS	COMMON	002824100	51694	982400	SH		SOLE	0	953000	29400		
AUTOMATIC DATA PROCESSIN	COMMON	053015103	185140	3514433	SH		SOLE	0	3391733	122700		
AVON PRODS INC	COMMON	054303102	80208	2864583	SH		SOLE	0	2766433	98150		
BERKSHIRE HATHAWAY INC DELCL B	COMMON	084670702	40021	517134	SH		SOLE	0	499084	18050		
COCA COLA CO	COMMON	191216100	144373	2145533	SH		SOLE	0	2071133	74400		

Figuur 2.6: Een 13F melding zonder gestructureerde data maar werkend met tabs, geen tabel structuur

```
<ns1:informationTable xmlns:ns1="http://www.sec.gov/edgar/document/thirteenf/informationtable">
  <ns1:infoTable>
    <ns1:nameOfIssuer>AB ACTIVE ETFS INC</ns1:nameOfIssuer>
    <ns1:titleOfClass>SHORT DURATION H</ns1:titleOfClass>
    <ns1:cusip>00039J830</ns1:cusip>
    <ns1:figi>BBG01N1MX948</ns1:figi>
    <ns1:value>2003805</ns1:value>
    <ns1:shrsOrPrnAmt>
      <ns1:sshPrnamt>57067</ns1:sshPrnamt>
      <ns1:sshPrnamtType>SH</ns1:sshPrnamtType>
    </ns1:shrsOrPrnAmt>
    <ns1:investmentDiscretion>SOLE</ns1:investmentDiscretion>
    <ns1:votingAuthority>
      <ns1:Sole>31139</ns1:Sole>
      <ns1:Shared>0</ns1:Shared>
      <ns1:None>25928</ns1:None>
    </ns1:votingAuthority>
  </ns1:infoTable>
```

Figuur 2.7: Een recente (2024) 13F melding gestructureerd in XML

Na 2013

2.2. Text mining en gerelateerde technieken

Text mining, of ook bekend tekstdatamining, is de procedure om ongestructureerde tekst om te zetten in een gestructureerd formaat om significante patronen te ont-

dekken en nieuwe inzichten te verwerven. Text mining maakt de analyse van uitgebreide tekstdatasets mogelijk om significante thema's, patronen en verborgen verbanden bloot te leggen. Deze techniek is essentieel voor het omzetten van ongestructureerde gegevens in gestructureerde gegevens, die vervolgens kunnen worden gebruikt voor analyse en besluitvorming(IBM, 2024).

2.2.1. Document datatypes

Text mining kan verschillende soorten gegevens omvatten, waaronder:

- **Gestructureerde Gegevens**: Deze gegevens zijn gestandaardiseerd in een tabelvorm, wat ze makkelijker maakt om op te slaan en te verwerken voor analyse en machine learning-algoritmen. Voorbeelden includeren databanken met kolommen en rijen.
- **Ongestructureerde Gegevens**: Deze gegevens hebben geen vooraf gedefinieerd formaat en kunnen tekst uit bronnen zoals sociale media of productrecensies bevatten, evenals rijke media zoals video- en audiobestanden. Aangezien financiële documenten vaak in ongestructureerd formaat bestaan, is text mining essentieel om deze gegevens om te zetten in een bruikbaar formaat.
- **Semi-gestructureerde Gegevens**: Deze gegevens vormen een mix tussen gestructureerde en ongestructureerde formaten. Ze hebben enige organisatie, maar voldoen niet volledig aan de vereisten van een relationele database. Voorbeelden hiervan zijn XML, JSON en Html-bestanden.

Dit onderscheid zijn van groot belang voor het begrijpen van hoe text mining toegepast kan worden over de verschillende datastructuren, dit opent de mogelijkheid om de data de extraheren en belangrijke inzichten te verwerven(AWS, 2024).

2.2.2. Text mining vs. Text analytics

Hoewel text mining en text analytics vaak door elkaar worden gebruikt, kan er een genuanceerd onderscheid tussen de twee gemaakt worden. Bij text mining gaat het meestal om het identificeren van patronen en trends in ongestructureerde gegevens, terwijl text analytics gericht is op het afleiden van kwantitatieve inzichten door gegevens op een gestructureerde manier te analyseren. Deze observaties kunnen vervolgens grafisch worden weergegeven om de ontdekkingen effectief over te brengen aan een breder publiek.(IBM, 2024)

Text mining: vinden van verstopte patronen

Text mining omvat het extraheren van waardevolle informatie en het identificeren van verborgen patronen uit uitgebreide verzamelingen ongeorganiseerde of gedeeltelijk georganiseerde tekstuele gegevens. Text mining is een gespecialiseerde

vorm van datamining die is ontworpen om vooral tekstuele informatie te verwerken. Het belangrijkste doel van text mining is om tekst om te zetten in analyseerbare gegevens om inzichten, trends en patronen te ontdekken die niet direct voor de hand liggen. Deze aanpak omvat een reeks methodologieën, waaronder het ophalen van informatie, natuurlijke taalverwerking (NLP) en machinaal leren. Het primaire doel is het begrijpen en analyseren van uitgebreide tekstdatabases (Gaikwad e.a., 2014).

Voor- en nadelen text mining

Volgens (Kinter, 2024) en (Gaikwad e.a., 2014) zijn er veel voor- en nadelen aan text mining:

- Voordelen:
  - Het corpus van teksten kan worden geanalyseerd met technieken zoals informatie-extractie om de namen van verschillende entiteiten en hun relaties te identificeren.
  - De complexe taak om effectief om te gaan met grote hoeveelheden ongestructureerde gegevens om patronen bloot te leggen, wordt aangepakt door het gebruik van text mining.
  - Bedrijven kunnen een uitgebreid inzicht krijgen in huidige trends en patronen door inzichten te analyseren die zijn verkregen uit vele gegevensbronnen. Deze inzichten helpen bedrijven bij het nemen van weloverwogen zakelijke beslissingen.
- Nadelen
  - Text mining gebruikt vaak een grote hoeveelheid gegevens. Het efficiënt opslaan, beheren en verwerken van deze gegevens vereist daarom een grote hoeveelheid opslagruimte en rekenkracht, wat duur kan zijn.
  - Text mining, gegevensanalyse en patroonherkenning zijn sterk afhankelijk van de kwaliteit van de gegevens. De nauwkeurigheid van de resultaten kan worden beïnvloed door variaties in de gegevenskwaliteit, die worden beïnvloed door de structuur en voorbewerking van de gegevens.

Text analyse: het afleiden van semantische betekenis

Tekstanalyse daarentegen houdt zich meer bezig met het begrijpen en interpreteren van de inhoud van tekst om er informatie van hoge kwaliteit uit af te leiden. In tegenstelling tot text mining, dat zich richt op het ontdekken van nieuwe patronen, is tekstanalyse gericht op het extraheren en interpreteren van bestaande informatie uit tekstgegevens. Dit proces omvat de toepassing van semantische analysetechnieken om de betekenis, context en bedoeling achter de woorden in de tekst te begrijpen (Gaikwad e.a., 2014).

Tekstanalyse maakt vaak gebruik van Natural Language Processing (NLP) om de structuur van zinnen te ontleden, entiteiten te identificeren en sentiment te analyseren. Deze technieken zijn cruciaal voor taken zoals sentimentanalyse, waarbij het doel is om de emotionele toon van een tekst te bepalen, of onderwerpmodellering, waarbij het doel is om de belangrijkste thema's te identificeren die in een set documenten worden besproken. Tekstanalyse kan ook meer geavanceerde methoden omvatten, zoals entiteitherkenning, waarbij belangrijke stukken informatie (zoals namen, data en locaties) in een tekst worden geïdentificeerd en geclassificeerd.

conclusie

Terwijl text mining vaak verkennend is, waarbij gezocht wordt naar onbekende patronen, is tekstanalyse gerichter, waarbij de nadruk ligt op het extraheren van specifieke informatie van hoge kwaliteit uit de tekst. In een juridische context kan tekstanalyse bijvoorbeeld worden gebruikt om relevante clausules uit een contract te halen, terwijl text mining kan worden gebruikt om trends in juridische beslissingen in de loop van de tijd te identificeren (Gaikwad e.a., 2014).

2.2.3. Text mining technieken

(Talib e.a., 2016) spreekt over enkele technieken zoals Information Extraction (IE), Information retrieval (IR), en Meerdere NLP technieken die gebruikt worden in data mining, deze zullen hier besproken worden

Information retrieval vs Information extraction

Information retrieval

Informatie retrieval (Information Retrieval, IR) verwijst naar de interactie tussen mens en computer wanneer een gebruiker informatie zoekt die overeenkomt met zijn of haar zoekopdracht in een database of computersysteem. Dit proces omvat het ophalen van relevante inhoud op basis van de behoeften van de gebruiker. Het systeem vergelijkt de zoekopdracht van de gebruiker met een reeks documenten om de meest relevante te identificeren en presenteert deze uiteindelijk in een geprioriteerde lijst. Dit gespecialiseerde vakgebied, zoals beschreven door Krallinger (2024), is essentieel om gebruikers in staat te stellen snel en efficiënt informatie te lokaliseren en extraheren uit uitgebreide en vaak ongestructureerde gegevensbronnen zoals tekstdocumenten, databases of het internet.

De effectiviteit van een IR-systeem wordt gemeten aan de hand van metrieken zoals precisie en recall. Precisie is de verhouding tussen het aantal relevante documenten dat wordt opgehaald en het totale aantal opgehaalde documenten, terwijl recall de verhouding is tussen het aantal relevante documenten dat wordt opgehaald en het totale aantal relevante documenten in de dataset (Javija, 2024). Deze metrics helpen om informatie-overload te verminderen door ervoor te zorgen dat alleen de meest relevante informatie aan de gebruiker wordt gepresenteerd. De methoden en technieken die worden gebruikt in IR-systemen zijn fundamenteel

voor het aandrijven van technologieën zoals zoekmachines, die een snelle en efficiënte informatie ophaling mogelijk maken (Krallinger e.a., 2017).
Enkele IR technieken zijn maar niet gelimiteerd tot (IBM, 2024):

- Tokenizatie: Dit is het proces van het opbreken van text in zinnen en woorden genoemd tokens. Deze zijn dan gebruikt in de modellen voor clustering en documentmatching taken(IBM, 2024).
- Stemming is een tekstvoorbewerkingsmethode die wordt gebruikt in natuurlijke taalverwerking (NLP) om woorden te vereenvoudigen door ze om te zetten naar hun basisvorm. Het doel van stemming is om woorden te stroomlijnen en te normaliseren en zo de efficiëntie van het ophalen van informatie, het categoriseren van teksten en andere natuurlijke taalverwerkingsactiviteiten (NLP) te verbeteren(SaturnCLoud, 2024).

Information Extraction
Informatie-extractie (IE) is gericht op het extraheren van gestructureerde informatie uit ongestructureerde documenten met behulp van technieken zoals Natural Language Processing (NLP). In tegenstelling tot Information Retrieval (IR), waarbij relevante documenten worden opgehaald, richt IE zich op het identificeren van specifieke gegevens binnen deze teksten, waardoor informatie toegankelijker en analyseerbaarder wordt (Javija, 2024). IE-systemen moeten kosteneffectief en aanpasbaar zijn en in staat zijn om zich over verschillende domeinen uit te breiden. Op gebieden zoals financiën wordt Named-Entity Recognition (NER) gebruikt om vooraf gedefinieerde gegevenstypen, zoals namen en data, uit documenten te extraheren, wat efficiënt gegevensbeheer vergemakkelijkt (Gupta e.a., 2020). Geautomatiseerd leren in IE vermindert fouten en afhankelijkheid van handmatig toezicht, waardoor het proces efficiënter en contextueel waardevoller wordt. De toenemende hoeveelheid ongestructureerde gegevens, vooral online, benadrukt het belang van effectieve IE-systemen (Javija, 2024).
Enkele IE technieken zijn maar niet gelimiteerd tot (IBM, 2024):

- Feature selection en Feature extraction
  - Feature selection: is een essentiële stap bij het verwerken van gegevens met een groot aantal dimensies. Het gaat om het kiezen van een kleinere set belangrijke kenmerken uit de originele set om de efficiëntie en nauwkeurigheid van het leren te verbeteren. Door overbodige en inconsequente kenmerken te elimineren, wordt de omvang van de gegevensverwerking verkleind, wordt de tijd die nodig is voor het leren verminderd en worden de resultaten gestroomlijnd. Eigenschapsselectie is een proces dat de belangrijkste oorspronkelijke kenmerken behoudt, in tegenstelling tot kenmerkextractie waarbij gegevens worden veranderd in ken-

merken die goed zijn in het herkennen van patronen. Eigenschapsselectie is cruciaal voor het verminderen van de dimensionaliteit van gegevens. Technieken voor kenmerkselectie omvatten een reeks benaderingen, zoals supervised, unsupervised en semi-supervised modellen. Deze methoden worden geclassificeerd op basis van hun associatie met leermethoden (filter, wrapper, inbeddingsmodellen) en andere criteria. Eigenschapsselectie is een veelgebruikte techniek in gebieden zoals beeldherkenning en tekst mining. Het verbetert de prestaties van modellen voor machinaal leren door een evenwicht te bereiken tussen hoge nauwkeurigheid en lage rekenvereisten (Cai e.a., 2018).

- Feature extraction: is een essentiële stap in machinaal leren, omdat het uitgebreide invoergegevens omzet in een beter hanteerbare en lager-dimensionale kenmerkenset. Deze procedure vereenvoudigt de gegevens door de complexiteit ervan te verminderen, terwijl belangrijke informatie toch behouden blijft. Het is vooral nuttig bij taken zoals categorisatie. Kenmerkextractietechnieken transformeren de initiële kenmerkruimte in een gecondenseerde, alternatieve ruimte door een gereduceerde, representatieve verzameling kenmerken te behouden in plaats van ze weg te gooien. Principale Componenten Analyse (PCA) en Bag of Words zijn vaak gebruikte technieken. PCA vermindert bijvoorbeeld de dimensionaliteit van gegevens door de oorspronkelijke variabelen om te zetten in ongecorrigeerde componenten. Dit proces verbetert de rekenefficiëntie en verhoogt de nauwkeurigheid van modellen voor machinaal leren (Suhaidi e.a., 2021).
- Verschil? FS behoud de originele features terwijl FE nieuwe maakt.

- Named Entity Recognition (NER) is een kerntaak in Natural Language Processing (NLP) die tot doel heeft entiteiten, zoals personen, organisaties en plaatsen, binnen een gegeven tekst te herkennen en te categoriseren. NER, of Named Entity Recognition, wordt op grote schaal gebruikt in een verscheidenheid aan toepassingen, variërend van het ophalen van informatie tot geautomatiseerde klantenservice.

  Recent onderzoek benadrukt dat, hoewel NER-modellen opmerkelijke prestaties hebben behaald op typische datasets en vaak hoge F-scores laten zien, deze maat alleen geen goed inzicht geeft in hun effectiviteit. Geavanceerde NER-modellen vertonen bijvoorbeeld F-scores van meer dan 90% op datasets zoals OntoNotes. Desalniettemin kan deze eenzame metriek variaties in prestaties verdoezelen tussen verschillende categorieën entiteiten, soorten taal en onbekende data (Vajjala & Balasubramaniam, 2022).

Aspect	Information Retrieval	Information Extraction
Focus	Document Retrieval	Feature Retrieval
Uitvoer	Geeft een set van documenten terug	Geeft feiten van een document terug
Doel	Het doel is om documenten te vinden die relevant zijn voor de informatiebehoefte van de gebruiker.	Het doel is om vooraf gespecificeerde kenmerken uit documenten te halen of informatie weer te geven.
Aard van informatie	Echte informatie ligt verborgen in documenten	Extraheer informatie uit de documenten
Application	Used in many search engines – Google is the best IR system for the web.	Used in database systems to enter extracted features automatically.
Methodology	Typically uses a bag of words model of the source text.	Typically based on some form of semantic analysis of the source text.
Theoretical Basis	Mostly use the theory of information, probability, and statistics.	Emerged from research into rule-based systems.

Tabel 2.1: Comparison of Information Retrieval and Information Extraction

Conclusie

Information Retrieval (IR) en Information Extraction (IE) zijn twee technologieën die informatie toegankelijk maken via verschillende methodologieën.IR richt zich op het ophalen van relevante documenten, terwijl IE specifieke, gestructureerde informatie extraheert voor nauwkeurige gegevensanalyse.IR is essentieel voor grote datasets en zoekmachines, terwijl IE cruciaal is voor het extraheren van bruikbare inzichten. De kracht van IR ligt in het beheren en ophalen van informatie uit ongestructureerde bronnen, waardoor het onmisbaar is voor grote databases. IE is van vitaal belang voor datamining, kennisbeheer en geautomatiseerde processen. Naarmate het datavolume toeneemt, zal de wisselwerking tussen IR en IE steeds belangrijker worden. Het begrijpen en benutten van beide technologieën zal cruciaal zijn voor het optimaliseren van informatieverwerkingssystemen en om ervoor te zorgen dat gebruikers snel en accuraat de benodigde informatie kunnen verkrijgen. Voor dit onderzoek zal men informatie extractie gebruiken.

NLP

Summarization

Een andere kritische NLP techniek is tekstsamenvatting, waarbij een beknopte weergave van originele tekstdocumenten wordt gegenereerd. Dit proces omvat voorbewerkingsstappen zoals tokeniseren, stopwoorden verwijderen en stemmen, gevolgd door het creëren van lexiconlijsten tijdens de verwerkingsfase. Historisch gezien was het samenvatten van tekst gebaseerd op woordfrequentie, maar moderne methoden maken gebruik van geavanceerde text mining technieken om de relevantie en nauwkeurigheid van de resultaten te verbeteren. Kenmerken zoals zinslengte, thematische woorden en vaste zinnen worden gebruikt om belangrijke informatie te extraheren en deze technieken kunnen op meerdere documenten tegelijk worden toegepast(Talib e.a., 2016).

Part of speech tagging

Part-of-speech (POS) tagging is een essentiële activiteit in natuurlijke taalverwerking (NLP) waarbij een grammaticale classificatie, zoals zelfstandig naamwoord, werkwoord of bijvoeglijk naamwoord, wordt toegewezen aan elk woord in een zin. Tagging vergemakkelijkt computationeel begrip van de syntactische organisatie van tekst, een kritisch onderdeel voor veel toepassingen van natuurlijke taalverwerking (NLP) (Martinez,2012).Ondanks de uitdagingen zoals tweeslachtige woorden bereiken moderne POS taggers hoge nauwkeurigheidspercentages (rond 96-97%) en worden ze veel gebruik bij het ophalen van informatie, tekstanalyse en andere NLP-taken(Martinez, 2012).

Text categorization

TODO - REview use Tekstclassificatie is een methodische procedure die bestaat uit vier essentiële stappen: kenmerken extraheren, de dimensionaliteit verminderen, een classificator selecteren en de resultaten evalueren. Eerst wordt tekst omgezet in een numeriek formaat door middel van kenmerkextractie, zoals woord frequentie of Word2Vec. Dimensionaliteitsreductiewordt gebruikt om de gegevens te vereenvoudigen en cruciale informatie te behouden. Dit wordt bereikt door technieken zoals principale componentenanalyse(PCA) of lineaire discriminantanalyse (LDA) toe te passen. De selectie van een classificeer is essentieel, omdat deep learning-methoden vaak conventionele machinelearning-algoritmen overtreffen in termen van nauwkeurigheid. Uiteindelijk wordt de doeltreffendheid van het model beoordeeld door de prestaties te meten met behulp van metrieken zoals de Matthews correlatiecoëfficiënt (MCC), oppervlakte onder de ROC-curve (AUC) en nauwkeurigheid. Van deze maatstaven wordt nauwkeurigheid beschouwd als de meest directe en eenvoudige manier om de prestatie van het model te evalueren. Gupta et al. (2020) ontdekten dat supportvectormachines (SVM) beter presteerden dan andere benaderingen zoals Naive Bayes (NB),k-nearest neighbour (KNN),

beslisbomen en regressie in termen van nauwkeurigheid, recall en F1-maatstaven.

Sentiment analysis
TODO - DEL (unrelated) Natural Language Processing (NLP) omvat verschillende technieken om onnauwkeurig en dubbelzinnig taalgebruik om te zetten in nauwkeurige en ondubbelzinnige berichten, met toepassingen in sectoren als financiën, e-commerce en sociale media. Een belangrijke techniek binnen NLP is Sentimentanalyse (SA), ook wel bekend als opinion mining, waarbij onderliggende meningen uit tekstgegevens worden gehaald. SA is vooral nuttig voor taken als emotieherkenning en polariteitsdetectie, met toepassingen variërend van voorspelling van de aandelenmarkt tot analyse van feedback van klanten (Gupta e.a., 2020).
SA kan worden benaderd met lexicon gebaseerde methoden, die vertrouwen op tools zoals SentiWordNet voor woord-naar-sentiment mappen, of machine learning (ML) technieken die tekst classificeren met behulp van algoritmes zoals Naïve Bayes (NB) en support vectormachines (SVM's). Hoewel ML-benaderingen geen kostbare woordenboeken vereisen, vereisen ze wel domein specifieke datasets, wat een beperking kan zijn. Bovendien zijn deep learning-methoden onlangs gecombineerd met traditionele ML-technieken om de nauwkeurigheid en betrouwbaarheid van SA te verbeteren, met name in financiële voorspellingen (Gupta e.a., 2020).
TODO - Add REGEX? TODO - Add statistical table extraction TODO - Train llama(3, 8B, instruct, unsloth.ai model (pruned), finetune)? (NO GPT -> payed) TODO - Restructure texts

2.3. Technieken en Tools

2.3.1. SpaCy vs NLTK

Tabel geeft een vergelijkende analyse van SpaCy en NLTK op basis van belangrijke functies die relevant zijn voor tekstsamenvatting. TODO - Prune tabel (Amade e.a., 2024)
Prestatiewaarden (Precisie, Recall, F-Score): SpaCy toont hogere precisie (0.72 vs. 0.51) en F-Score (0.69 vs. 0.58) in vergelijking met NLTK, wat wijst op superieure nauwkeurigheid bij het genereren van samenvattingen.
Snelheid van Tokenizatie en Tagging: SpaCy is aanzienlijk sneller dan NLTK in zowel tokeniseren als taggen. SpaCy kan bijvoorbeeld tekst tokenen in 0,2 milliseconden vergeleken met de 4 milliseconden van NLTK, waardoor het geschikter is voor toepassingen die real-time verwerking vereisen. **Ondersteuning voor Geavanceerde NLP-functies:** SpaCy ondersteunt geavanceerde functies zoals topic modellering, vectorisatie en TF-IDF (Term Frequency-Inverse Document Frequency), die niet standaard beschikbaar zijn in NLTK. Dit maakt SpaCy een meer uitgebreide tool voor taken die diep semantisch begrip en machine learning integratie vereisen.

Functie	NLTK	SpaCy
Precisie	0.51	0.72
Recall	0.65	0.65
F-Score	0.58	0.69
Tokenisatiesnelheid	4 ms	0.2 ms
Taggingsnelheid	443 ms	1 ms
Ondersteuning voor Classificatie	Ja	Ja
Onderwerpmodellering	Nee	Ja
Vectorisatie	Nee	Ja
Parsing	Ja	Ja
TF-IDF Implementatie	Nee	Ja
Programmeerparadigma	Procedureel	Objectgeoriënteerd
Gebruiksvriendelijkheid	Vereist meer aanpassing en tijd	Meer geautomatiseerd en gebruiksvriendelijk
Ondersteunde Taalmodellen	Basis Tokenizatie en parsing	Geavanceerde modellen met voorgetrainde vectors
Grootte en Afhankelijkheden van de Bibliotheek	Lichtgewicht, minimale afhankelijkheden	Zwaarder, meer afhankelijkheden door geavanceerde functies

Tabel 2.2: Vergelijkende Analyse van SpaCy en NLTK

Gebruiksvriendelijkheid: SpaCy is gebruiksvriendelijker met ingebouwde mogelijkheden, wat de behoefte aan maatwerkprogrammering vermindert, in tegenstelling tot NLTK, dat meer tijd en inspanning vergt.

subsubsection

Conclusie Samenvattend, SpaCy is een krachtigere en efficiëntere tool voor tekstsamenvatting vanwege zijn hogere precisie, snelheid en ondersteuning voor geavanceerde NLP-functies. NLTK, hoewel veelzijdig, is beter geschikt voor eenvoudigere taken of projecten die meer aanpassing vereisen. De keuze tussen deze tools hangt af van de specifieke eisen van het project, waaronder de complexiteit van de taak, de benodigde functies en de beschikbare middelen.

2.3.2. Database Management Systemen (DBMS)

In deze sectie gaan wij bekijken welke databank gebruikt zal worden na het structuren en standaardiseren van de 13f meldingen. Hier zal besproken worden of er SQL of nosql gebruikt zal worden vervolgens zal er een specifieke databank gekozen worden

SQL vs. NOSQL

Bij het kiezen tussen SQL- en Nosql-databases is het belangrijk om de onderliggende architectuur en toepassingsmogelijkheden te begrijpen. SQL-databases zijn ontworpen voor het organiseren van gestructureerde data, waardoor ze ideaal zijn voor online transaction processing (OLTP). Ze presteren uitstekend in situaties waarin complexe query's, consistentie en relationeel databeheer vereist zijn. Nosql-databases daarentegen ondersteunen horizontale schaalbaarheid en zijn geoptimaliseerd voor het verwerken van grote hoeveelheden ongestructureerde data, wat hen geschikt maakt voor big data-analyse. De keuze tussen beide hangt grotendeels af van de specifieke behoeften van de organisatie, zoals de focus op datastructuur of schaalbaarheid (Khan e.a., 2023).
In dit onderzoek is gekozen voor een SQL-database. Deze keuze is gebaseerd op de noodzaak om gestructureerde data uit de 13F-meldingen te beheren, waarbij consistente gegevensintegriteit en de mogelijkheid om complexe query's uit te voeren cruciaal zijn. SQL-databases bieden de benodigde functionaliteiten voor het beheer van relationele gegevens en het uitvoeren van geavanceerde analyses, wat essentieel is voor het succes van dit project (Khan e.a., 2023).

SQL-databank

Op basis van de gedetailleerde analyse die door (Javija, 2024) werd PostgreSQL gekozen voor ons proefschrift vanwege de geavanceerde functies, robuuste gegevensintegriteit en uitbreidbare architectuur. In tegenstelling tot andere SQL-databases, blinkt PostgreSQL uit in het verwerken van complexe datamanipulatie, het bieden van sterke ACID compliance en het ondersteunen van aangepaste datatypes en functies. Dit maakt PostgreSQL bijzonder geschikt voor bedrijfstoepassingen en datawarehousing waar schaalbaarheid en geavanceerd databeheer cruciaal zijn. Hoewel PostgreSQL een steilere leercurve heeft dan sommige alternatieven, maken de uitgebreide functie set en betrouwbaarheid het een optimale keuze om aan de complexe eisen van ons project te voldoen.

2.4. Uitdagingen en beperkingen

In dit hoofdstuk zullen enkele uitdagingen en beperkingen benoemt worden

2.4.1. Variable structuur

Vóór 2013 bevatten 13F-meldingen enigszins verschillende variabele structuren, maar deze variaties zijn belangrijk omdat ze het moeilijk maken om de gegevens te lezen en te analyseren. Vóór 2013 waren de 13F-meldingen variabeler in de manier waarop ze gegevens verstrekten, waardoor het moeilijk was om beleggingsportefeuilles door de tijd heen te vergelijken. Door dit gebrek aan standaardisatie moesten onderzoekers en analisten zorgvuldig omgaan met deze gegevens om consistente en nauwkeurige bevindingen te krijgen. Diepgaand onderzoek is nodig om interpretatiefouten te minimaliseren en investeringen en investeringspatronen als gevolg van verschillende rapportagestandaarden volledig te begrijpen.

2.4.2. Gegevenskwaliteit en Validatie

TODO - Als Llama lukt

2.4.3. Databaseprestaties

Uitdaging: Hoewel PostgreSQL goed presteert bij grote hoeveelheden gestructureerde data, kan het moeilijk zijn om de prestaties te optimaliseren naarmate de hoeveelheid data en het aantal gelijktijdige gebruikers toeneemt.
Beperking: Bij zeer grote datasets of een hoge mate van gelijktijdige toegang kunnen er prestatieproblemen optreden. Het kan nodig zijn om uitgebreide optimalisaties en schaalstrategieën te implementeren, zoals partitionering of het gebruik van read replicas.

2.5. Leemtes in huidig onderzoek

Dit hoofdstuk geeft enkele gaten weer in het huidig onderzoek.

2.5.1. Onbehandelde kwesties

Training van Eigen Large Language Models (LLMs)

Het trainen van een eigen LLM voor financiële toepassingen vereist veel tijd en middelen. Het model moet worden getraind op uitgebreide financiële datasets voor nauwkeurige resultaten. Dit kan uw infrastructuur belasten en vereist expertise in machine learning en datawetenschap, met mogelijke problemen op het gebied van data-integriteit en privacy.

Beveiliging en Privacy

Het beschermen van gevoelige financiële gegevens tegen ongeautoriseerde toegang en datalekken is complex en vereist naleving van privacywetgeving. Onvoldoende beveiliging kan leiden tot datalekken, verlies van vertrouwen en juridische problemen. Implementeer encryptie, toegangscontrole en regelmatige beveiligingsaudits om gegevens te beschermen. Zorg ervoor dat uw systemen voldoen aan

relevante regelgeving en best practices voor gegevensbeveiliging.

Schaalbaarheid en Prestaties
Groeiende hoeveelheden gegevens kunnen leiden tot prestatieproblemen bij opslag en analyse, wat complexe oplossingen vereist voor snelle toegang. Slechte prestaties kunnen vertragingen veroorzaken in rapportage en analyse, wat de besluitvorming en efficiëntie beïnvloedt. Gebruik schaalbare databases en technieken zoals gegevenspartitionering en caching. Monitor en optimaliseer regelmatig de prestaties om problemen te voorkomen.

3

Methodologie

Dit hoofdstuk geeft een overzicht van de methodologie die is gebruikt om dit onderzoek uit te voeren en de Proof of Concept (POC) te creëren. De tekst biedt een uitgebreide analyse van het belang van elke fase van het onderzoek en licht de redenering achter de gekozen methodologieën en benaderingen toe. Dit hoofdstuk maakt duidelijk hoe de gekozen benaderingen helpen om de onderzoeksdoelen te bereiken door een goed georganiseerd overzicht te bieden. Het belang van elke fase wordt benadrukt, waardoor inzicht wordt verkregen in de achterliggende gedachte van de beslissingen die tijdens het onderzoeksproces zijn genomen.

3.1. Literatuur studie

De eerste fase van dit onderzoek bestond uit een uitgebreid onderzoek van bestaande literatuur. Het doel van deze fase was om een grondig begrip te krijgen van de concepten en technologieën die gebruikt zouden worden bij de implementatie van de Proof of Concept (POC). De bovengenoemde stap omvatte een uitgebreide analyse van verschillende publicaties, papers, blogs en handleidingen om relevante toepassingen en benaderingen te ontdekken. De belangrijkste onderwerpen die in deze fase werden onderzocht waren Text Mining, Natural Language Processing (NLP) en Database Management Systemen (DBMS).

3.2. Requirements analyse

Het doel van deze requirements analyse is om de eisen en voorwaarden vast te stellen voor een methodologie waarin alle benodigde middelen en processen kosteloos beschikbaar moeten zijn. Deze methodologie richt zich op projecten, tools, en diensten die gratis toegankelijk zijn, zodat er geen financiële belemmeringen zijn voor de uitvoering.

Algemene Eisen

Toegankelijkheid:
De gekozen tools en resources moeten wereldwijd gratis toegankelijk zijn, zonder geografische of economische beperkingen.

Licentievoorwaarden:
Er moet gebruik worden gemaakt van open-source software of diensten met een volledig gratis licentie, zoals GNU General Public License (GPL), MIT-licentie, of soortgelijke. Er mogen geen verborgen kosten of verplichtingen zijn verbonden aan het gebruik van deze software of diensten.

3.3. Apparaten

Hier een korte lijst van de gebruikte apparaten tijdens dit onderoek.
TODO - Tabel?

- Laptop Dell XPS 15 9500
  - OS: Windows 11 Pro
  - CPU: Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz
  - RAM: 32 GB
  - GPU: NVIDIA GeForce GTX 1650 Ti 4GB VRAM
- Google Collab CPU
  - CPU: Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz
- Google Collab GPU (free credits)
- RAM: 13GB
  - RAM: 13GB
  - GPU: Nvidia T4 15GB VRAM

3.4. Dataset creation

Tijdens deze fase hebben we een dataset gegenereerd met de 13F-dossiers als basis, die de basis vormde voor de constructie van het Proof of Concept (POC). De informatie werd zorgvuldig samengesteld door pertinente financiële gegevens uit de 13F-papieren te halen, waarbij gegarandeerd werd dat de informatie geordend en geformatteerd werd op een manier die geschikt is voor latere analyse en verwerking binnen het POC-kader.

3.5. Comparitive study (NLP vs ML vs DM)

Llama

3.6. POC

In het volgende deel van ons onderzoek willen we een Proof of Concept (POC) uitvoeren met SpaCy, een zeer gewaardeerd Python framework voor natuurlijke taalverwerking (NLP). SpaCy werd geselecteerd vanwege zijn sterke vaardigheid in het beheren van uitgebreide tekstdatasets, wat cruciaal is voor het analyseren van ingewikkelde financiële informatie, zoals die in 13F filings.
Het doel van de proof of concept (POC) is het onderzoeken en verifiëren van de doeltreffendheid van SpaCy bij het uitvoeren van essentiële NLP (natural language processing) activiteiten, zoals het extraheren van tekst, het herkennen van named entities (NER) en het ophalen van informatie. Het uitvoeren van deze taken is cruciaal voor de nauwkeurige identificatie en extractie van belangrijke entiteiten, zoals bedrijfsnamen en financiële statistieken, uit het ongeorganiseerde materiaal in 13F-papers.
Een cruciaal onderdeel van deze proof of concept (POC) is het aanpassen van SpaCy's natuurlijke taalverwerking (NLP) aan de specifieke eisen van het project. Een van deze benaderingen is het trainen van gespecialiseerde NER-modellen met behulp van domeinspecifieke gegevens. Dit kan helpen om de nauwkeurigheid van het herkennen van financiële woorden en entiteiten te verhogen, waardoor de algehele betrouwbaarheid van het systeem toeneemt.
Door de implementatie van deze proof of concept (POC) willen we de haalbaarheid van het gebruik van SpaCy voor dit project aantonen en een basis leggen voor de uitgebreide implementatie van de natuurlijke taalverwerkingsoplossingen (NLP) die nodig zijn voor het verwerken en onderzoeken van financiële gegevens uit 13F filings.

3.7. Database

PostgreSQL is gekozen als databasebeheeroplossing voor dit project om te voldoen aan de vereisten voor gegevensbeheer. PostgreSQL is een vrij beschikbaar databasesysteem dat de eigenschappen van objectgeoriënteerde en relationele databases combineert. Het is zeer betrouwbaar, kan grote hoeveelheden gegevens aan en heeft uitstekende mogelijkheden voor het uitvoeren van gecompliceerde queries. Deze kwaliteiten maken het een uitstekende optie voor het beheren van de gedetailleerde financiële informatie uit 13F deponeringen.
De keuze voor PostgreSQL werd beïnvloed door drie belangrijke factoren:
PostgreSQL garandeert de integriteit van gegevens door de ACID (Atomicity, Consistency, Isolation, Durability) principes volledig te ondersteunen. Precisie en consistentie zijn van het grootste belang bij het verwerken van gevoelige financiële gegevens.
PostgreSQL heeft geavanceerde mogelijkheden, waaronder ondersteuning voor JSON-gegevenstypen, full-text zoeken en aangepaste indexering. Deze functies

zijn met name voordelig voor het effectief verwerken van semigestructureerde gegevens die kunnen voortkomen uit natuurlijke taalverwerkingstaken (NLP).
Het flexibele ontwerp van PostgreSQL maakt het mogelijk om de database aan te passen aan de unieke eisen van ons project, met name wat betreft het opslaan en opvragen van financiële informatie. Dit wordt bereikt door de mogelijkheid om nieuwe functies, operatoren en datatypes te bouwen.
Schaalbaarheid is een belangrijke factor bij het overwegen van PostgreSQL's vermogen om enorme datasets goed te beheren, vooral gezien de omvang en complexiteit van de gegevens die geproduceerd worden tijdens het verwerken van 13F aanvragen. PostgreSQL heeft de mogelijkheid om zowel horizontaal als verticaal uit te breiden, zodat het effectief kan voldoen aan de toenemende eisen van het project naarmate er meer gegevens worden toegevoegd.

3.8. Analyse van de resultaten

Hier zal men een kort overzicht van de verworven resultaten weergeven en wat we er mogelijks mee kunnen doen

4

Methodologie

4.0.1. Voorbereiding

Data

TODO - Explain the source (raw) data and processed data for training

Omgeving

TODO - explain env (libs, acces to model,...)

4.0.2. Praktische Vergelijking Technieken

TODO- Expand intro Llama, Statisticly table extraction, Spacy (IR, IE, NER), REGEX
TODO - show outputs to serve as example and to make it visualy more intresting

Manuele extractie
Handmatige extractie houdt in dat documenten of bestanden met de hand worden doorgenomen en dat de benodigde informatie wordt overgezet in een gestructureerd formaat, zoals een spreadsheet of een database. Dit proces wordt vaak gebruikt bij kleine datasets of wanneer de gegevens niet beschikbaar zijn in een digitaal formaat.

- **Voordelen**:
  - Nauwkeurig voor kleine datasets
  - Simpel
- **Nadelen**:
  - Niet praktisch voor grote datasets
  - Tijdrovend
  - Menselijke fouten

  - Niet schaalbaar

Vanwege de aard van handmatige extractie is deze niet geschikt voor deze POC, omdat de volledige 13F dataset honderdduizenden bestanden bevat.

REGEX

Reguliere expressies (regex) zijn patronen die gebruikt worden om opeenvolgingen van tekens in tekst te matchen. Ze kunnen worden gebruikt om specifieke patronen te identificeren en te extraheren uit tekstbestanden, wat bijzonder nuttig kan zijn voor het parsen van gestructureerde of semigestructureerde gegevens.

- **Voordelen**:
  - Flexibel: Regex laat toe om op maat gemaakte patronen maken die passen bij een grote verscheidenheid aan gegevens formaten.
  - Integratie: Regex is gemakkelijk te integreren in bestaande software
- **Nadelen**:
  - Complex: naarmate patronen ingewikkelder worden, kan regex moeilijk te lezen en onderhouden beginnen worden.
  - Gelimiteerd: Regex kan moeite hebben met ongestructureerde of zeer variabele gegevens. Als de gegevens niet voldoen aan een voorspelbaar patroon of als er significante afwijkingen zijn, kan regex niet goed presteren en belangrijke informatie missen of fouten genereren.

Vanwege de aard van de informatie in de 13F-rapportagetabellen bleek het schrijven van een regex aanvankelijk een noodzakelijke stap voor het ontwikkelen van een werkend proof of concept (POC). Dit werd echter al snel stopgezet door de grote variëteit in opmaak en structuur, een missende/extra waarden, indentatie verschillen en extra spaties. Deze werd hierdoor niet gebruikt voor het extraheren van de tabel informatie maar wel voor de algeme informatie van het indiendende bedrijf.

Statistic table extraction

- **Voordelen**:
  - Reliable
- **Nadelen**:
  -

IR en IE met Spacy

Het verschil tussen Information Retrieval en Information Extraction blijkt klein, maar beide methoden hebben moeite om missende data op te vangen. Dit was deels verwacht omdat het model vooral getraind is op gestructureerde data en minder goed kan omgaan met ontbrekende of ongestructureerde informatie. Information Retrieval richt zich op het vinden van relevante informatie op basis van zoektermen, terwijl Information Extraction specifieke entiteiten of relaties uit een tekst haalt. Het ontbreken van gegevens kan tot onnauwkeurigheden leiden, wat aangeeft dat aanvullende technieken nodig zijn om nauwkeuriger resultaten te verkrijgen. Het model presteert minder goed wanneer de gegevens inconsistent of onvolledig zijn, wat het belang benadrukt van kwalitatief goede, goed gestructureerde data bij deze methoden.

Named entity recognition met Spacy

Hoewel deze benadering beter presteert dan IR (Information Retrieval) en IE (Information Extraction), is het nog steeds niet voldoende. Named Entity Recognition (NER) ondervindt ook problemen, vooral met ontbrekende gegevens. De tool heeft moeite om correcte entiteiten te identificeren en te extraheren wanneer er data ontbreekt of incompleet is, wat de nauwkeurigheid en effectiviteit van het systeem beïnvloedt. Hierdoor blijft de algehele prestatiesubstantie onder de verwachtingen, en zijn er aanvullende aanpassingen en verbeteringen nodig om de resultaten te optimaliseren.

Llama

Llama presteert het beste in vergelijking met alle andere technieken, maar ondervindt nog steeds moeilijkheden wanneer het wordt geconfronteerd met structuren die aanzienlijk afwijken van wat het eerder heeft gezien. Ondanks deze uitdaging, heeft Llama echter wel de capaciteit om effectief om te gaan met ontbrekende waarden. Het systeem kan robuust omgaan met gegevenshiaten, maar de prestaties kunnen worden beperkt wanneer het wordt geconfronteerd met ongebruikelijke of radicaal verschillende structuren die het nog niet eerder heeft aangetroffen.

Analyse resulaten LLama?

TODO review neccessity

4.0.3. Databank

Als databank zal men gebruik maken van postrgesql

4.0.4. Conclusie

TODO - summrisation It is possible but not now by because, dont want to spend money (not literally), do not have the time to expand the trainings data set which is recommended when you want to implement this als it will be a good choice to choose a stronger variant of llama3.1 (ipv 8B params 70B or maybe 405B (rivals gpt4o - best llm)params) or maybe a GPT model but ofcourse better models require better hardware which requires more money -> Expad this + transl

5

Conclusie

A

Onderzoeksvoorstel

A.1. Inleiding

Waarover zal je bachelorproef gaan? Introduceer het thema en zorg dat volgende zaken zeker duidelijk aanwezig zijn:

- kaderen thema
- de doelgroep
- de probleemstelling en (centrale) onderzoeksvraag
- de onderzoeksdoelstelling

Denk er aan: een typische bachelorproef is *toegepast onderzoek*, wat betekent dat je start vanuit een concrete probleemsituatie in bedrijfscontext, een **casus**. Het is belangrijk om je onderwerp goed af te bakenen: je gaat voor die *ene specifieke probleemsituatie* op zoek naar een goede oplossing, op basis van de huidige kennis in het vakgebied.
De doelgroep moet ook concreet en duidelijk zijn, dus geen algemene of vaag gedefinieerde groepen zoals *bedrijven*, *developers*, *Vlamingen*, enz. Je richt je in elk geval op it-professionals, een bachelorproef is geen populariserende tekst. Eén specifiek bedrijf (die te maken hebben met een concrete probleemsituatie) is dus beter dan *bedrijven* in het algemeen.
Formuleer duidelijk de onderzoeksvraag! De begeleiders lezen nog steeds te veel voorstellen waarin we geen onderzoeksvraag terugvinden.
Schrijf ook iets over de doelstelling. Wat zie je als het concrete eindresultaat van je onderzoek, naast de uitgeschreven scriptie? Is het een proof-of-concept, een rapport met aanbevelingen, ...Met welk eindresultaat kan je je bachelorproef als een succes beschouwen?

A.2. Literatuurstudie

Hier beschrijf je de *state-of-the-art* rondom je gekozen onderzoeksdomein, d.w.z. een inleidende, doorlopende tekst over het onderzoeksdomein van je bachelorproef. Je steunt daarbij heel sterk op de professionele *vakliteratuur*, en niet zozeer op populariserende teksten voor een breed publiek. Wat is de huidige stand van zaken in dit domein, en wat zijn nog eventuele open vragen (die misschien de aanleiding waren tot je onderzoeksvraag!)?
Je mag de titel van deze sectie ook aanpassen (literatuurstudie, stand van zaken, enz.). Zijn er al gelijkaardige onderzoeken gevoerd? Wat concluderen ze? Wat is het verschil met jouw onderzoek?
Verwijs bij elke introductie van een term of bewering over het domein naar de vakliteratuur, bijvoorbeeld (Hykes, 2013)! Denk zeker goed na welke werken je refereert en waarom.
Draag zorg voor correcte literatuurverwijzingen! Een bronvermelding hoort thuis *binnen* de zin waar je je op die bron baseert, dus niet er buiten! Maak meteen een verwijzing als je gebruik maakt van een bron. Doe dit dus *niet* aan het einde van een lange paragraaf. Baseer nooit teveel aansluitende tekst op eenzelfde bron.
Als je informatie over bronnen verzamelt in JabRef, zorg er dan voor dat alle nodige info aanwezig is om de bron terug te vinden (zoals uitvoerig besproken in de lessen Research Methods).
Je mag deze sectie nog verder onderverdelen in subsecties als dit de structuur van de tekst kan verduidelijken.

A.3. Methodologie

Hier beschrijf je hoe je van plan bent het onderzoek te voeren. Welke onderzoekstechniek ga je toepassen om elk van je onderzoeksvragen te beantwoorden? Gebruik je hiervoor literatuurstudie, interviews met belanghebbenden (bv. voor requirementsanalyse), experimenten, simulaties, vergelijkende studie, risico-analyse, PoC, ...?
Valt je onderwerp onder één van de typische soorten bachelorproeven die besproken zijn in de lessen Research Methods (bv. vergelijkende studie of risico-analyse)? Zorg er dan ook voor dat we duidelijk de verschillende stappen terug vinden die we verwachten in dit soort onderzoek!
Vermijd onderzoekstechnieken die geen objectieve, meetbare resultaten kunnen opleveren. Enquêtes, bijvoorbeeld, zijn voor een bachelorproef informatica meestal **niet geschikt**. De antwoorden zijn eerder meningen dan feiten en in de praktijk blijkt het ook bijzonder moeilijk om voldoende respondenten te vinden. Studenten die een enquête willen voeren, hebben meestal ook geen goede definitie van de populatie, waardoor ook niet kan aangetoond worden dat eventuele resultaten representatief zijn.
Uit dit onderdeel moet duidelijk naar voor komen dat je bachelorproef ook tech-

nisch voldoende diepgang zal bevatten. Het zou niet kloppen als een bachelorproef informatica ook door bv. een student marketing zou kunnen uitgevoerd worden.
Je beschrijft ook al welke tools (hardware, software, diensten, ...) je denkt hiervoor te gebruiken of te ontwikkelen.
Probeer ook een tijdschatting te maken. Hoe lang zal je met elke fase van je onderzoek bezig zijn en wat zijn de concrete *deliverables* in elke fase?

A.4. Verwacht resultaat, conclusie

Hier beschrijf je welke resultaten je verwacht. Als je metingen en simulaties uitvoert, kan je hier al mock-ups maken van de grafieken samen met de verwachte conclusies. Benoem zeker al je assen en de onderdelen van de grafiek die je gaat gebruiken. Dit zorgt ervoor dat je concreet weet welk soort data je moet verzamelen en hoe je die moet meten.
Wat heeft de doelgroep van je onderzoek aan het resultaat? Op welke manier zorgt jouw bachelorproef voor een meerwaarde?
Hier beschrijf je wat je verwacht uit je onderzoek, met de motivatie waarom. Het is **niet** erg indien uit je onderzoek andere resultaten en conclusies vloeien dan dat je hier beschrijft: het is dan juist interessant om te onderzoeken waarom jouw hypothesen niet overeenkomen met de resultaten.

B

Bijlagen

Bibliografie

Amade, D., Chandra, R., Sinha, V. K., & Anand, D. (2024). Automatic Text Summarization Using NLTK & Spacy [Available at SSRN: https://ssrn.com/abstract=4742012 or http://dx.doi.org/10.2139/ssrn.4742012].

AWS. (2024). What is the difference between structured and unstructured data? [accessed: 2024-08-09].

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/https://doi.org/10.1016/j.neucom.2017.11.077

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17). https://doi.org/10.5120/14937-3507

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovations*, *6*(1), 39. https://doi.org/10.1186/s40854-020-00205-1

Hykes, S. (2013, maart 21). *The future of Linux Containers (PyCon 2013)*. Verkregen september 1, 2016, van https://www.youtube.com/watch?v=wW9CAH9nSLs

IBM. (2024). What is text-mining [accessed: 2024-08-08].

Javija, R. (2024). Difference between Information Retrieval and Information Extraction (GeeksForGeeks, Red.) [16-07-2024].

Khan, W., Kumar, T., Zhang, C., Raj, K., Roy, A. M., & Luo, B. (2023). SQL and NoSQL Database Software Architecture Performance Analysis and Assessments—A Systematic Literature Review [Submission received: 13 January 2023 / Revised: 6 May 2023 / Accepted: 8 May 2023 / Published: 12 May 2023]. *Big Data Cogn. Comput.*, *7*(2), 97. https://doi.org/10.3390/bdcc7020097

Kinter, P. (2024). Text mining: applications and techniques [accessed: 2024-08-09].

Krallinger, M., Rabal, O., Lourenço, A., Oyarzábal, J., & Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, *117*, 7673–7761. https://doi.org/10.1021/acs.chemrev.6b00851

Martinez, A. R. (2012). Part-of-speech tagging [WIREs Comp Stat 2012, 4:107–113. doi: 10.1002/wics.195]. https://doi.org/https://doi.org/10.1002/wics.195

SaturnCLoud. (2024). Stemming in Natural Language Processing [accessed: 2024-08-10].

Securities, U., & Commission, E. (2023). Frequently Asked Questions About Form 13F [Accessed: 2024-08-08].

Suhaidi, M., Kadir, R. A., & Tiun, S. (2021). A REVIEW OF FEATURE EXTRACTION METHODS ON MACHINE LEARNING. *JOURNAL INFORMATION AND TECHNOLOGY MANAGEMENT JISTM*, *6*(22), 51–59. https://gaexcellence.com/index.php/jistm/article/view/1125

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(11). https://doi.org/10.14569/IJACSA.2016.071107

Vajjala, S., & Balasubramaniam, R. (2022). What do we Really Know about State of the Art NER? https://doi.org/https://doi.org/10.48550/arXiv.2205.00034