Geavanceerde Extractie van 13F-gegevens met Tekstmining AI en NLP.

Optionele ondertitel.

Thomas Vanderveken.

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

Promotor: Lena De Mol
Co-promotor: Lieven Smits
Academiejaar: 2023–2024
Derde examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Allereerst wil ik mijn oprechte dank uitspreken aan dhr. Smits voor het aanbieden van dit uitdagende en fascinerende onderwerp voor mijn bachelorproef. Het thema van natuurlijke taalverwerking (NLP) en grote taalmodellen (LLM's) sprak mij bijzonder aan, vooral omdat het in de lessen slechts kort aan bod kwam. De complexiteit van het onderwerp motiveerde mij om dieper in de materie te duiken en mijn kennis uit te breiden, voortbouwend op wat ik tijdens mijn stage heb geleerd.

Daarnaast wil ik mijn ouders hartelijk bedanken voor hun voortdurende steun. Van het transporteren van mij en mijn spullen naar en van mijn kot tot het verzorgen van verse kleren en maaltijden, hun hulp was cruciaal. Hun onvoorwaardelijke steun heeft niet alleen mijn bachelorproef, maar mijn gehele academische traject mogelijk gemaakt. Voor alles wat ze hebben gedaan, ben ik hen enorm dankbaar.

Ik wil ook mijn vrienden bedanken voor hun begrip en geduld tijdens mijn afwezigheid. Jullie hebben me de ruimte gegeven die ik nodig had om me volledig op mijn proefschrift te concentreren, en jullie vriendschap en steun waren een grote troost in drukke tijden.

Mijn mede-kotgenoten verdienen ook een speciale vermelding. Jullie constante aanmoediging en de motiverende gesprekken hielpen me om door te zetten, zelfs op de momenten dat ik het moeilijk vond. Jullie enthousiasme en steun hebben mijn werkproces aanzienlijk verbeterd.

Ten slotte wil ik mijn medestudenten bedanken voor hun waardevolle hulp en advies. Jullie bereidheid om te helpen, zelfs als ik op bepaalde gebieden tekortschiet, en jullie raad hebben bijgedragen aan de verbetering van mijn werk.

Aan iedereen die op welke manier dan ook heeft bijgedragen aan dit project en mijn academische reis: jullie steun is onmisbaar geweest en ik waardeer dit enorm.

Samenvatting

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Lijst van codefragmenten

1

Inleiding

Amerikaanse investeringsmaatschappijen met een jaaromzet van ten minste USD 100.000.000, zoals Soros Fund Management, Appaloosa Management en Berkshire Hathaway, zijn periodiek verplicht hun financiële posities te rapporteren aan de Securities and Exchange Commission (SEC). Dit doen ze onder andere via de zogenaamde 13F-filings. Deze filings worden opgeslagen in de Edgar-database en zijn openbaar toegankelijk. SEC EDGAR Search for 13F-HR Forms (2001-2012)
Regelgevende filings van institutionele beleggers, zoals pensioenfondsen en vermogensbeheerders, bieden belangrijke inzichten in marktpatronen en beleggingsstrategieën. Ze vormen vaak de basis voor het genereren van aan- en verkoopsignalen op beurzen zoals de NYSE en Nasdaq. De informatie die in deze registraties wordt weergegeven, is van cruciaal belang voor financieel onderzoek en beleggingsanalyses.
13F-meldingen van vóór 2013 brengen echter aanzienlijke uitdagingen met zich mee vanwege hun variabele vormen en structuren, wat de verwerking en analyse ervan bemoeilijkt. Tot die tijd werden de meldingen ingediend als ”flat files”, oftewel .txt-bestanden, zonder strikte regels over de locatie van specifieke informatie. Dit maakt het systematisch extraheren van nuttige gegevens lastig.
De opkomst van geavanceerde AI-technologie biedt mogelijkheden om deze problemen aan te pakken. Natural Language Processing (NLP) en Machine Learning (ML) bieden geavanceerde methoden om gegevens uit ongestructureerde tekst te extraheren en te organiseren. Deze technologieën kunnen het proces van het standaardiseren en integreren van vroegere 13F-meldingen in een goed georganiseerde relationele databank automatiseren, wat de toegang en het gebruik van deze gegevens aanzienlijk zou verbeteren.
Sinds 2013 verplicht de SEC bedrijven om hun rapporten in .xml-formaat in te dienen, wat de structuur van de gegevens aanzienlijk heeft verbeterd. Toch blijven de gegevens van vóór 2013 belangrijk, omdat ze kunnen dienen als testdata voor het

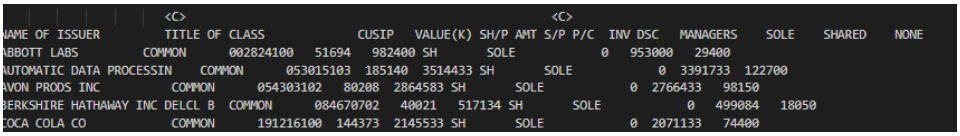
ontwikkelen van beleggingsstrategieën.
Dit proefschrift richt zich op het ontwikkelen van een proof-of-concept toepassing die NLP en ML-technieken gebruikt om 13F-meldingen van vóór 2013 te standaardiseren en te integreren in een relationele databank. Het doel is om het proces van gegevensextractie te stroomlijnen door de meldingen automatisch om te zetten in een gestandaardiseerd formaat met hoge efficiëntie en nauwkeurigheid. Dit zou niet alleen de analyse van historische financiële gegevens optimaliseren, maar ook de werklast en kosten van handmatige gegevensverwerking verminderen.
De gestandaardiseerde gegevens zouden bovendien het begrip van investeringspatronen uit het verleden verbeteren en het ontwikkelen van voorspellingsmodellen ondersteunen. Het onderzoek begint met een literatuurstudie om de meest efficiënte NLP- en ML-strategieën te identificeren, waarna een proof-of-concept toepassing wordt ontwikkeld en geëvalueerd op nauwkeurigheid, efficiëntie en bruikbaarheid.
De inleiding biedt een beknopt overzicht van de redenen, doelen en het belang van dit onderzoek. Het werk beoogt een waardevolle bijdrage te leveren aan de analyse van financiële gegevens en biedt onderzoekers en analisten een nuttig hulpmiddel door de uitdagingen aan te pakken die gepaard gaan met de verwerking 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.

Enkele voorbeelden

Een voorbeeld van voor 2013. Een voorbeeld van na 2013.



```
                    <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 1.1: 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 1.2: Een recente (2024) 13F melding gestructureerd in XML

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 en is dit überhaupt mogelijk?

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. Het onderzoek stoelt op een POC te implementeren. Dit wil zeggen dat indien het omzettingsproces slaagt dit een aanleiding kan zijn om een volwaardige software tool te ontwerpen (door derden).

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 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen. In Hoofdstuk 4 worden de vereisten voor het ontwikkelen van de proof of concept besproken, inclusief een prioritering op basis van hun belang.
In Hoofdstuk **??** 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 6, 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 usd 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 Amerikaanse 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 zowel huidige als oudere 13F-aanvragen worden omgezet 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. Achtergrond informatie

2.1.1. Reden voor 13F meldingen

TODO

2.1.2. Andere filings

Naast de 13F filings zijn er verschillende andere belangrijke SEC-filings die investeringsbedrijven systematisch moeten indienen en die nuttig kunnen zijn voor het opbouwen van een eigen investeringsportfolio. Hier zijn enkele van de meest relevante:

Form 10-K

De 10-K is een jaarlijkse rapportage die een uitgebreid overzicht biedt van de prestaties van een bedrijf gedurende het afgelopen jaar. Deze filing bevat gedetailleerde financiële gegevens, informatie over bedrijfsactiviteiten, risicofactoren en managementanalyses. Het is een cruciaal document voor investeerders die inzicht willen krijgen in de financiële gezondheid en strategie van een bedrijf (Ganesh, 2024) .

Form 10-Q

De 10-Q is een kwartaalrapportage die bedrijven verplicht zijn in te dienen. Het biedt een update over de financiële prestaties en operationele activiteiten van het bedrijf in de afgelopen drie maanden. Dit document is minder uitgebreid dan de 10-K, maar biedt toch waardevolle informatie over de recente ontwikkelingen en

trends (Ganesh, 2024) (Baker, 2022).

Form 8-K
De 8-K is een actuele rapportage die bedrijven moeten indienen wanneer er belangrijke gebeurtenissen plaatsvinden die van invloed kunnen zijn op de aandelenprijs of de operationele status van het bedrijf. Dit kan bijvoorbeeld gaan om fusies, overnames, wijzigingen in het management of andere significante gebeurtenissen. Het is belangrijk voor investeerders om deze filings te volgen, omdat ze snel inzicht geven in belangrijke veranderingen binnen een bedrijf (Ganesh, 2024)(Team, g.d.) (Baker, 2022).

Roxy Statement (DEF 14A)
De Proxy Statement bevat informatie over de jaarlijkse aandeelhoudersvergadering, inclusief details over stemprocedures, bestuursleden, en compensatie van het management. Dit document is waardevol voor investeerders die geïnteresseerd zijn in corporate governance en de belangen van aandeelhouders (Ganesh, 2024) (Baker, 2022).

Schedule 13D
Schedule 13D moet worden ingediend door elke persoon of entiteit die meer dan 5% van de aandelen van een publiek bedrijf aanschaft. Dit document bevat de identiteit van de aandeelhouder en de redenen voor de aankoop, wat nuttig kan zijn voor investeerders die de bewegingen van grote aandeelhouders willen volgen (Team, g.d.) (Baker, 2022).

Form D
Form D wordt gebruikt voor het melden van een privéplaatsing van effecten. Dit document kan nuttig zijn voor investeerders die geïnteresseerd zijn in alternatieve investeringsmogelijkheden of startups die kapitaal aantrekken zonder een volledige openbare aanbieding te doen (Ganesh, 2024).
Door deze verschillende SEC-filings te analyseren, kunnen investeerders een beter begrip krijgen van de bedrijven waarin ze geïnteresseerd zijn en weloverwogen beslissingen nemen bij het opbouwen van hun investeringsportefeuille.

2.2. 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.

2.2.1. 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.

2.2.2. 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 usd 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.

CUSIP
Volgens (Hayes, 2024) is CUSIP een afkorting voor ”Committee on Uniform Security Identification Procedures”. Het is een comité binnen de American Bankers Association (ABA) dat een systeem heeft ontwikkeld voor het identificeren van Amerikaanse en Canadese effecten.
Een CUSIP-nummer is een uniek nummer bestaande uit 9 cijfers en letters dat aan elk effect wordt toegekend. Ze bieden een gestandaardiseerde methode voor het identificeren van effecten om de clearing en afwikkeling van transacties op de handelsmarkt te vergemakkelijken. Deze nummers werden beheerd door Standard & Poor’s maar FactSet Research Systems heeft dit overgekocht in naam van American Bankers Association(ABA) in 2022 zij beheren deze nummer nu ook (Hayes, 2024).
De CUSIP-code is een voorbeeld van een National Securities Identifying Number, een standaard identificatiemethode voor effecten.

Omvang van de informatie:
De rapportages over het aandelenbezit van grote beleggingsinstellingen richten zich voornamelijk op aandelen, terwijl andere soorten activa, zoals obligaties, deri-

vaten en private equity, buiten beschouwing worden gelaten (Securities & Commission, 2023). 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

Hieronder zijn een aantal voorbeelden van de informatie tabel van enkele 13F meldigen zichtbaar. Zoals te zien is, 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

```
                                                                                SHARED
                                                MARKET SOLE  SHARED  OTHER            SOLE    SHARED  NONE
ECURITY DESCRIPTION                  CLASS  CUSIP    SHARES  VALUE  (A)   (B)     (C)    MGR  (A)     (B)     (C)
     <S>                             <C>    <C>      <C>     <C>    <C>   <C>     <C>    <C>  <C>     <C>     <C>
-------------------------------------------------------------------------------------------------------------------
de Corporation                       COM    00089C107  212,950   2,438  X                        201,900    0    11,050
-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------
VX, Corp.                            COM    002444107  1,498,286 24,467 X                      1,402,846    0    95,440
-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------
egis 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    Cusip      Value    Shs of SH/PRN  Discrtn Put/call   Voting
          Class                (X1000)  Prn amt                Authority
AFLAC Common            COM          001055102        1,427      33,500SH          SOLE            SOLE
Abbott Labs Common      COM          002824100        1,441      22,350SH          SOLE            SOLE
Apache Corp Common      COM          037411105          670       7,620SH          SOLE            SOLE
Apple Inc Common        COM          037833100        1,170       2,003SH          SOLE            SOLE
Becton Dickinson Common COM          075887109          206       2,750SH          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

2.3. 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 ontdekken en nieuwe inzichten te verwerven (IBM, 2024). 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.

2.3.1. Document datatypes

Volgens (AWS, 2024) kan text mining kan verschillende soorten gestructureerde gegevens omvatten, waaronder:

1. **Gestructureerde Gegevens**: Deze gegevens worden in tabelvorm gebracht, wat het opslaan en verwerken voor analyse en machine learning-algoritmen vergemakkelijkt. Voorbeelden includeren databanken met kolommen en rijen.
2. **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.
3. **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.

2.3.2. Text mining versus 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:

1. Het corpus van teksten kan worden geanalyseerd met technieken zoals informatie-extractie om de namen van verschillende entiteiten en hun relaties te identificeren.
2. 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.
3. 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

1. 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.
2. 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 entiteitsherkenning, 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.
Voorbeeld 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).

2.3.3. 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 versus Information extraction

Information retrieval
Informatie retrieval (Information Retrieval, IR) verwijst , zoals beschreven doorKrallinger e.a. (2017) naar de interactie tussens 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, 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.
Enkele IR technieken zijn maar niet gelimiteerd tot (IBM, 2024):

1. 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).

2. 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):

1. Feature selection en Feature extraction

   (a) **Feature selection**: Volgens Cai e.a. (2018) is Feature Selection 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 kenmerken 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.

   (b) **Feature extraction**: is een essentiële stap in machinaal leren, omdat het uitgebreide invoergegevens omzet in een beter hanteerbare en lager-dimensionale kenmerkenset (Suhaidi e.a., 2021). Deze procedure vereenvoudigt de gegevens door de complexiteit ervan te verminderen, terwijl belangrijke in-

formatie 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.

(c) Het verschil is dat FS de originele features behoud terwijl FE nieuwe maakt.

2. Volgens Vajjala en Balasubramaniam (2022) is **Named Entity Recognition (NER)** 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. Deze kan echter eenzame metriek variaties in prestaties verdoezelen tussen verschillende categorieën entiteiten, soorten taal en onbekende data.

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. IE extraheert specifieke, gestructureerde informatie 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 er gebruikt gemaakt worden van informatie extractie.

1. Aspect	Information Retrieval	Information Extraction
2. Focus	Document Retrieval	Feature Retrieval
3. Uitvoer	Geeft een set van documenten terug	Geeft feiten van een document terug
4. 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.
5. Aard van informatie	Echte informatie ligt verborgen in documenten	Extraheer informatie uit de documenten
6. Toepassing	Gebruikt in veel zoekmachines – Google is het beste IR-systeem voor het web.	Gebruikt in databasesystemen om automatisch geëxtraheerde kenmerken in te voeren.
7. Methodologie	Maakt doorgaans gebruik van een bag-of-words-model van de brontekst.	Gebaseerd op een vorm van semantische analyse van de brontekst.
8. Theoretische Basis	Maakt voornamelijk gebruik van de theorie van informatie, waarschijnlijkheid en statistiek.	Voortgekomen uit onderzoek naar regelgebaseerde systemen.

Tabel 2.1: Vergelijking tussen Information retrieval en Information Extraction

Natural Language Processing

Samenvatten van tekst

Een andere kritische NLP techniek is tekstsamenvatting, waarbij een beknopte weergave van originele tekstdocumenten wordt gegenereerd (Talib e.a., 2016). 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.

Part Of Speech (POS)

Volgens Martinez (2012) is Part-of-speech (POS) tagging 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.

2.3.4. REGEX in gegevens extractie

Reguliere expressies (regex) zijn een krachtig hulpmiddel voor patroonherkenning in tekst. Ze stellen gebruikers in staat om zoekpatronen te definiëren die specifieke reeksen van tekens kunnen identificeren en extraheren uit een grotere tekst, wat ze onmisbaar maakt bij taken zoals gegevensreiniging, parsing en informatie-extractie.

Gebruik van Regex

Regex kan gebruikt worden voor:

1. **Tekstvoorverwerking**: Regex kan tekstgegevens opschonen en standaardiseren door ongewenste tekens, witruimtes of inconsistenties in de opmaak te verwijderen.
2. **Patroonherkenning**: Het is effectief voor het vinden van specifieke patronen in tekst, zoals datums, telefoonnummers of e-mailadressen.
3. **Tokenisatie**: Regex kan helpen om tekst op te splitsen in tokens (woorden, zinnen) door patronen voor delimiters zoals spaties of interpunctie te definiëren.

Voordelen en Beperkingen

Volgens Nagarjoon (2022) zijn er enkele voordelen en beperkingen die gepaard gaan met REGEX.:

Voordelen

1. **Efficiëntie**: Regex is zeer efficiënt voor eenvoudige patroonherkenningstaken.
2. **Flexibiliteit**: Het kan worden aangepast om complexe tekstpatronen met nauwkeurige controle te herkennen.

Beperkingen

1. **Complexiteit**: Het opstellen van complexe regex-patronen kan uitdagend en foutgevoelig zijn.
2. **Schaalbaarheid**: Regex kan moeite hebben met zeer grote tekstcorpora of wanneer de logica van het patroon te complex wordt, waardoor het minder geschikt is voor sommige NLP-taken die een diepgaande semantische begrip vereisen.

2.4. LLM's: GPT versus Llama

Large Language Models (LLM's) zoals GPT (Generative Pre-trained Transformer) en Llama (Large Language Model Meta AI) hebben de verwerking van natuurlijke taal (NLP) gerevolutioneerd door transformer architecturen te gebruiken om tekst te begrijpen en te produceren die sterk lijkt op menselijke taal. Deze modellen worden getraind met behulp van uitgebreide datasets en hebben een breed spectrum aan toepassingen, inclusief maar niet beperkt tot tekstproductie en het beantwoorden van vragen.

2.4.1. Generatieve Pre-trained Transformer (GPT)

GPT is ontwikkeld door OpenAI en heeft zich ontwikkeld tot een van de meest prominente LLM. Het staat bekend om zijn vermogen om logische en contextueel geschikte taal te produceren als reactie op invoerprompts. Het ontwerp en de trainingsgegevens van GPT zorgen ervoor dat het model in veel taken uitblinkt, met name in taken waarbij ingewikkelde redeneringen en begrip van subtiele taalkundige aanwijzingen nodig zijn.

- **Verbeterde Creativiteit en Samenwerking**: GPT-4 is aanzienlijk creatiever en beter in staat om samen te werken aan complexe taken, zoals het genereren en bewerken van creatieve en technische teksten. Het kan zich aanpassen aan de schrijfstijl van de gebruiker en zelfs complexe opdrachten uitvoeren (openai, 2024).
- **Uitgebreide Algemene Kennis en Probleemoplossend Vermogen**: Dankzij een bredere algemene kennis en geavanceerde probleemoplossende capaciteiten kan GPT-4 moeilijke problemen met grotere nauwkeurigheid oplossen dan zijn voorganger, GPT-3.5 (openai, 2024).
- **Toepassingen in Diverse Sectoren**: GPT-4 wordt al gebruikt in verschillende innovatieve toepassingen, zoals Duolingo voor taalonderwijs, Be My Eyes voor visuele toegankelijkheid, en Stripe voor fraudebestrijding. Dit toont aan hoe veelzijdig en nuttig het model is in de praktijk (openai, 2024).

Nadelen

- **Bekende Beperkingen**: Ondanks de verbeteringen heeft GPT-4 nog steeds beperkingen, zoals sociale vooroordelen, hallucinaties, en kwetsbaarheid voor vijandige prompts. Dit kan leiden tot onnauwkeurigheden of ongewenste resultaten in bepaalde situaties (openai, 2024).
- **Toegang en Beschikbaarheid**: GPT-4 is momenteel alleen beschikbaar voor gebruikers van ChatGPT Plus, dit is een betalende service. Dit beperkt de directe toegang tot het model voor een bredere gebruikersgroep (openai, 2024).

2.4.2. Het grote taalmodel Meta AI (LLaMA3.1)

Meta AI heeft LLaMA gemaakt, een open-source alternatief voor modellen zoals GPT. Het presenteert een vergelijkbaar op transformers gebaseerd ontwerp, maar geeft prioriteit aan toegankelijkheid, use the flexibiliteit en kosteneffectiviteit, vooral voor academische en onderzoekstoepassingen. LLaMA modellen worden aangeboden in verschillende groottes, zoals LLaMA-3 en LLaMA-3.1, die elk verschillende prestaties en verwerkingscapaciteit bieden.

Voordelen

1. **Open Source Toegang**: Llama 3.1 is openlijk toegankelijk, waardoor ontwikkelaars modellen volledig kunnen aanpassen, trainen op nieuwe datasets, en verder kunnen afstemmen zonder beperkingen van gesloten bronmodellen (Meta, 2024).
2. **State-of-the-Art Capaciteiten**: Het 405B-model wordt beschreven als het grootste en meest capabele openlijk beschikbare funderingsmodel, dat kan concurreren met toonaangevende gesloten bronmodellen op het gebied van algemene kennis, meertalige vertaling en andere taken (Meta, 2024).
3. **Uitgebreide Context Lengte**: Met een contextlengte van 128K ondersteunt Llama 3.1 geavanceerde toepassingen zoals lange tekstsamenvattingen en complexe conversaties (Meta, 2024).
4. **Ecosysteem Ondersteuning**: Een breed ecosysteem met meer dan 25 partners, waaronder AWS, NVIDIA, en Google Cloud, ondersteunt Llama 3.1 vanaf dag één, wat integratie en ontwikkeling vergemakkelijkt (Meta, 2024).

Nadelen

1. **Vereiste Compute Resources**: Het 405B-model vereist aanzienlijke computermiddelen, wat het uitdagend maakt voor de gemiddelde ontwikkelaar om mee te werken zonder toegang tot hoogwaardige infrastructuur (Meta, 2024).

Benchmarks	GPT-4o	Meta Llama-3.1-405B	Meta Llama-3.1-70B	Meta Llama 3-70B	Meta Llama-3.1-8B	Meta Llama-3-8B
boolq	0.905	0.921	0.909	0.892	0.871	0.82
gsm8k	0.942	0.968	0.948	0.833	0.844	0.572
hellaswag	0.891	0.92	0.908	0.874	0.768	0.462
human_eval	0.921	0.854	0.793	0.39	0.683	0.341
mmlu_humanities	0.802	0.818	0.795	0.706	0.619	0.56
mmlu_other	0.872	0.875	0.852	0.825	0.74	0.709
mmlu_social_sciences	0.913	0.898	0.878	0.872	0.761	0.741
mmlu_stem	0.696	0.831	0.771	0.696	0.595	0.561
openbookqa	0.882	0.908	0.936	0.928	0.852	0.802
piqa	0.844	0.874	0.862	0.894	0.801	0.764
social_iqa	0.79	0.797	0.813	0.789	0.734	0.667
truthfulqa_mc1	0.825	0.8	0.769	0.52	0.606	0.327
winogrande	0.822	0.867	0.845	0.776	0.65	0.56

Figuur 2.8: Afbeelding toont een statistische vergelijking van GPT- en LLaMA-modellen op veelgebruikte NLP-benchmarks. (Graph, 2024)

2.4.3. Validatie van Large Language Models (LLMs) voor Gegevensextractie uit 13F-bestanden

Volgens (Huang, 2024) zijn er bij de evaluatie van Large Language Models (LLMs) voor het extraheren van gegevens uit 13F-bestanden specifieke uitdagingen en metrieken die in overweging moeten worden genomen In de bredere context van modelvalidatie, zoals besproken door Ray (2024), zijn er verschillende technieken die kunnen worden aangepast om de prestaties van LLMs bij gegevensextractie uit 13F-bestanden te evalueren. Deze technieken variëren afhankelijk van de specifieke aspecten van het model en de aard van de data.

Validatiemetrics en Uitdagingen

1. Ray (2024) benadrukt dat de nauwkeurigheid van een LLM bij het extraheren van specifieke gegevens zoals aandelenposities, waarden en data uit de complexe structuur van 13F-bestanden moet worden gemeten. Dit kan worden gedaan door de geëxtraheerde gegevens te vergelijken met een gecontroleerde dataset van handmatig gevalideerde informatie. Holdout Validation kan hier worden toegepast, waarbij het model wordt getraind op een deel van de data en getest op een ander deel om de nauwkeurigheid te evalueren.

2. Ray (2024) geeft aan dat 13F-bestanden een strikte structuur en format hebben. Het is cruciaal om te controleren of de LLM in staat is om gegevens correct te interpreteren en om te zetten naar het gewenste formaat. Metrics zoals precisie en recall kunnen helpen bij het beoordelen van hoe goed de LLM

voldoet aan de specificaties van het document. Hier kunnen de F1 Score en Cross-Validation nuttig zijn om een gebalanceerd beeld te krijgen van het model's prestaties over verschillende datasplitsingen.

3. Ray (2024) merkt op dat de validatie ook de robuustheid van het model onder verschillende omstandigheden moet testen. Dit houdt in dat de LLM moet presteren bij variaties in de indeling van 13F-bestanden, of bij fouten en inconsistenties in de gegevens. Adversarial Testing kan hierbij worden ingezet om specifieke zwakheden in het model bloot te leggen.

4. Ray (2024) wijst erop dat naast nauwkeurigheid, de snelheid waarmee een LLM gegevens kan extraheren ook belangrijk is. Evalueren hoe snel en efficiënt een model grote hoeveelheden 13F-bestanden kan verwerken, is cruciaal voor praktische toepassingen. Human Evaluation kan nuttig zijn om niet alleen de snelheid, maar ook de praktische bruikbaarheid en de outputkwaliteit van het model te beoordelen.

Best Practices voor Evaluatie

1. Huang (2024) beveelt het gebruik van een gouden dataset aan, bestaande uit 13F-bestanden met handmatig gevalideerde gegevens, om de prestaties van de LLM te meten. Deze dataset moet divers en representatief zijn voor de variaties in 13F-bestanden. Het gebruik van Zero-shot Evaluation kan helpen om het vermogen van het model te beoordelen om goed te presteren op onverwachte varianten binnen deze dataset.

2. Ray (2024) stelt voor om een iteratief evaluatieproces te implementeren waarbij feedback van eerdere evaluaties wordt gebruikt om het model verder te verfijnen. Dit helpt bij het verbeteren van de nauwkeurigheid en robuustheid van het model. Cross-Validation kan hier opnieuw nuttig zijn om het model te blijven testen tijdens het verfijningsproces.

3. Ray (2024) suggereert dat offline evaluaties, zoals het testen van het model op een vaste dataset, gecombineerd moeten worden met online evaluaties door het model te testen op echte gegevensstromen. Dit biedt een uitgebreide beoordeling van zowel de theoretische als praktische prestaties van de LLM. Perplexity kan hierbij worden gebruikt om de voorspellende kwaliteit van het model in online settings te meten.

Door deze benaderingen en metrics te integreren in het validatieproces, kan er een effectieve en betrouwbare LLM ontwikkeld worden voor het extraheren van gegevens uit 13F-bestanden, waardoor de nauwkeurigheid en efficiëntie van de gegevensverwerking verbeterd.

Conclusie
Samengevat is GPT een geavanceerde, commercieel verkrijgbare optie voor sommige NLP-taken, terwijl LLaMA een praktisch open-source alternatief is, vooral wanneer financiële beperkingen of de behoefte aan personalisatie van het grootste belang zijn. Door zorgvuldige selectie, training en finetuning van LLaMA-modellen kunnen betrouwbare, op maat gemaakte resultaten worden verkregen voor gerichte toepassingen, zonder de kosten van betalende modellen zoals GPT4o.
Dus wordt voor dit onderzoek Llama geselecteerd om de POC te ontwikkelen.

2.4.4. Data vereisten voor LLMs te trainen

Volgens Scispace (2024) is er minstens 0.5% van de originele dataset nodig om een LLM effectief te trainen. Dit minimale percentage kan leiden tot verbeterde prestaties, waarbij modellen, ondanks het gebruik van een fractie van de data, een aanzienlijke nauwkeurigheid behalen in specifieke taken.

2.5. Technieken en Tools

2.5.1. SpaCy versus NLTK

De volgende tabel geeft een vergelijkende analyse van SpaCy en NLTK op basis van belangrijke functies die relevant zijn voor tekstsamenvatting.

Functie	NLTK	SpaCy
01. Precisie	0.51	0.72
02. Recall	0.65	0.65
03. F-Score	0.58	0.69
04. Tokenisatie-snelheid	4 ms	0.2 ms
05. Taggingsnelheid	443 ms	1 ms
06. Ondersteuning voor Classificatie	Ja	Ja
07. Onderwerpmodellering	Nee	Ja
08. Vectorisatie	Nee	Ja
09. Ondersteunde Taalmodellen	Basis Tokenizatie en parsing	Geavanceerde modellen met voorgetrainde vectors
10. Grootte en Afhankelijkheden van de Bibliotheek	Lichtgewicht, minimale afhankelijkheden	Zwaarder, meer afhankelijkheden door geavanceerde functies

Tabel 2.2: Vergelijkende Analyse van SpaCy en NLTK (Amade e.a., 2024)

Here are some examples demonstrating the strengths of SpaCy over NLTK:

Prestatiewaarden (Precisie, Recall, F-Score):
SpaCy behaalt een precisie van 0.72 en een F-Score van 0.69 bij het genereren van samenvattingen, terwijl NLTK respectievelijk 0.51 en 0.58 scoort. Dit wijst op een betere nauwkeurigheid en effectiviteit van SpaCy in vergelijking met NLTK.

Snelheid van Tokenizatie en Tagging:
Volgens Saadani (2024) tokeniseert SpaCy tekst in 0,2 milliseconden, terwijl NLTK 4 milliseconden nodig heeft. Dit maakt SpaCy veel sneller en geschikter voor real-time toepassingen zoals chatbots en live datastreams.
Tokenisatie is het proces van het splitsen van tekst in kleinere eenheden, zoals woorden of zinnen, die ”tokens”worden genoemd. Bijvoorbeeld, de zin ”Het is een mooie dag”wordt getokeniseerd in de tokens: [”Het”, ïs”, ëen”, ”mooie”, ”dag”].
Tagging verwijst naar het toewijzen van labels of tags aan de getokeniseerde eenheden om hun grammaticale rol of betekenis te identificeren. Bijvoorbeeld, in de zin ”Het is een mooie dag,”zou tagging de woorden als volgt kunnen labelen:

[(”Het”, B-ARTIKEL), (ïs”, B-WERKWOORD), (ëen”, B-ARTIKEL), (”mooie”, B-BIJVOEGLIJK NAAMWOORD), (”dag”, B-NAAMWOORD)].

Ondersteuning voor Geavanceerde NLP-functies:
(Spacy, 2024) biedt ingebouwde ondersteuning voor complexe NLP-functies zoals topic modellering en vectorisatie. NLTK vereist vaak extra maatwerk of externe bibliotheken voor deze functionaliteiten, wat SpaCy meer geschikt maakt voor geavanceerde machine learning taken.

Gebruiksvriendelijkheid:
(Spacy, 2024) heeft een intuïtieve API met uitgebreide documentatie en kant-en-klare modules, waardoor het gemakkelijker is om mee te werken. NLTK vereist vaak meer configuratie en maatwerk, wat de leercurve kan verhogen en de implementatietijd verlengt.

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 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.

2.5.2. 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 die aan ACID voldoet.

ACID

Volgens (Kaur, 2024) zijn de ACID-eigenschappen—Atomiciteit, Consistentie, Isolatie, en Duurzaamheid—zijn fundamenteel voor het waarborgen van betrouwbare transacties in een Database Management Systeem (DBMS). Hieronder volgt een korte uitleg van elke eigenschap:

1. **Atomiciteit:** Atomiciteit zorgt ervoor dat een transactie wordt behandeld als één ondeelbare eenheid van werk. Dit betekent dat alle bewerkingen binnen de transactie volledig worden uitgevoerd of helemaal niet. Als een deel van de transactie mislukt, wordt de gehele transactie teruggedraaid naar de oorspronkelijke staat, waardoor dataconsistentie en integriteit worden gewaarborgd.

2. **Consistentie:** Consistentie verzekert dat een transactie de database van een consistente staat naar een andere consistente staat brengt. De database moet zowel voor als na de transactie in een consistente toestand verkeren. Dit houdt in dat integriteitsregels, zoals unieke sleutels en vreemde sleutels, behouden blijven om dataconsistentie te waarborgen.

3. **Isolatie:** Isolatie zorgt ervoor dat meerdere transacties gelijktijdig kunnen worden uitgevoerd zonder elkaar te beïnvloeden. Elke transactie moet geïsoleerd blijven van andere transacties totdat deze is voltooid. Deze isolatie voorkomt problemen zoals 'dirty reads', niet-herhaalbare leesacties, en 'phantom reads', wat bijdraagt aan de consistentie van de database.

4. **Duurzaamheid:** Duurzaamheid garandeert dat zodra een transactie is vastgelegd, de resultaten permanent zijn en overleven ondanks eventuele systeemfouten. De wijzigingen die door de transactie zijn aangebracht, worden blijvend opgeslagen in de database en blijven intact, zelfs bij systeemcrashes.

Deze eigenschappen vormen samen een kader voor het waarborgen van de consistentie, integriteit en betrouwbaarheid van gegevens in een DBMS. Ze zorgen ervoor dat transacties op een betrouwbare en consistente manier worden uitgevoerd, zelfs in het geval van systeemfouten, netwerkproblemen of andere complicaties. Dankzij de ACID-eigenschappen blijft een database een betrouwbaar en efficiënt hulpmiddel voor het beheren van gegevens in moderne organisaties.

SQL versus NOSQL

Bij het kiezen tussen SQL- en Nosql-databases is het belangrijk om de onderliggende architectuur en toepassingsmogelijkheden te begrijpen (Khan e.a., 2023). 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 ondersteunen horizontale schaalbaarheid en zijn geschikt voor het verwerken van grote hoeveelheden ongestructureerde data, waardoor ze ideaal zijn voor big data-analyse. De keuze tussen beide hangt grotendeels af van de specifieke behoeften van het onderzoek, zoals de focus op datastructuur of schaalbaarheid.
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 uitgevoerd is PostgreSQL gekozen voor dit 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 dit project te voldoen.

2.5.3. Unsloth.AI

Volgens Unsloth (2024) is unsloth een platform dat zich richt op het verbeteren van de snelheid en efficiëntie van het trainen van grote taalmodellen (LLMs). Traditioneel gezien kunnen deze trainingsprocessen lang duren en veel geheugen vereisen, maar Unsloth heeft technologie ontwikkeld die dit proces tot 30 keer sneller maakt dan gebruikelijke methoden zoals Flash Attention 2 (FA2). Bovendien verbruikt Unsloth hierbij tot 90% minder geheugen. Het platform werkt met verschillende soorten GPU's, waaronder die van NVIDIA, AMD en Intel, zonder dat er dure hardware-upgrades nodig zijn. Dit maakt het trainen van AI-modellen toegankelijker en goedkoper, met mogelijkheden variërend van gratis tot uitgebreide enterprise-oplossingen.

2.6. Uitdagingen en beperkingen

In dit hoofdstuk zullen enkele uitdagingen en beperkingen benoemt worden

2.6.1. 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.

2.6.2. 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.

2.7. Tekortkomingen in huidig onderzoek

Dit hoofdstuk geeft enkele tekortkomingen weer in het huidig onderzoek.

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. Fase 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). Deze fase zal 6 weken in beslag nemen en zal als resultaat een literatuurstudie zijn die van groot belang is om het verdere verloop van het onderzoek te begrijpen.

3.2. Fase 2 - Requirements analyse

Het doel van deze fase is het opstellen van de criteria waaraan de POC moet voldoen en het identificeren van de specifieke onderdelen die erin moeten zitten om als succesvol te worden beschouwd. Deze fase zal een week duren om succesvol te voltooien. Volgens autociteAchimigu2014 is de MoSCoW-techniek een methode

die kan worden gebruikt om prioriteit toe te kennen aan leveringen. Deze operatie wordt uitgevoerd met behulp van de MoSCoW-methode. D.w.z. er kan ook een prioriteit worden toegekend aan de verschillende benodigdheden als ze op deze manier worden georganiseerd. Het eindproduct van dit deel van het onderzoek was een geprioriteerde lijst van wat wel en niet nodig was om een succesvol proof-of-concept te genereren.

3.3. Fase 3 - POC

In deze fase wordt het Proof of Concept (PoC) uitgevoerd, waarbij er in verschillende stappen de haalbaarheid en effectiviteit van de voorgestelde oplossingen zullen testen. De belangrijkste stappen omvatten de implementatie van de geselecteerde technieken, het verzamelen en voorbereiden van de benodigde data, en het configureren van de database. Vervolgens wordt de PoC geëvalueerd op basis van de vastgestelde criteria om de resultaten te analyseren en verdere verfijningen aan te brengen. Deze fase zal naar verwachting 5 weken in beslag nemen.

3.3.1. Dataset creatie

Tijdens deze fase hebben wordt er een dataset gecreëerd 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.

3.3.2. Vergelijking technieken

In deze sectie wordt een gedetailleerde vergelijking gepresenteerd van technieken binnen de domeinen van Natuurlijke Taalverwerking (NLP), Machine Learning (ML), en Data Mining (DM). Elke techniek heeft haar eigen unieke voor- en nadelen, afhankelijk van het toepassingsgebied en de specifieke doelen die nagestreefd worden.

3.3.3. Databank

Voor dit project wordt PostgreSQL ingezet als database-oplossing om te voldoen aan de eisen voor gegevensbeheer. PostgreSQL biedt robuuste mogelijkheden voor het opslaan, verwerken en opvragen van grote hoeveelheden financiële gegevens. De database wordt geoptimaliseerd voor het effectief beheren van semi-gestructureerde gegevens die voortkomen uit de NLP-taken.

3.3.4. Implementatie

In dit gedeelte wordt het Proof of Concept (POC) uitgevoerd om de effectiviteit van verschillende benaderingen te evalueren voor het verwerken en analyseren van de financiële gegevens in de 13F-dossiers. Hierbij zullen er diverse tools en frameworks onderzoeken en testen om essentiële NLP-taken uit te voeren, zoals tekstextractie, Named Entity Recognition (NER), en informatieherwinning. Het POC richt zich op het demonstreren van de haalbaarheid en effectiviteit van de gekozen methoden in het project.

3.3.5. Analyse van de resultaten

In deze fase zullen de resultaten van het POC analyseren, waarbij de prestaties van de vergeleken technieken en de efficiëntie van de database-oplossing worden geëvalueerd. Deze analyse biedt inzicht in de verdere toepassing van de resultaten en mogelijke vervolgstappen binnen het project.

3.4. Fase - 4

In de loop van deze laatste fase zal er een extra week worden gereserveerd voor het opstellen van een conclusie en abstract over het verloop van het onderzoek. Daarnaast is er in de laatste twee weken van het onderzoek tijd voorzien om de scriptie voor deze bachelorproef af te ronden.

4

Benodigdheden

Dit hoofdstuk van deze bachelorproef bespreekt de vereisten voor de ontwikkeling van de proof of concept. De lijst van benodigdheden is opgesteld volgens de MoSCoW-methode, waarmee duidelijk kan worden bepaald wat absoluut noodzakelijk is en wat buiten het bereik van dit project valt.
De MoSCoW-methode, zoals onderzocht door Achimugu e.a. (2014), is een acroniem dat de prioriteiten van de vereisten aangeeft. De **M** staat voor ”must have”, de essentiële vereisten met de hoogste prioriteit. De **S** staat voor ”should have”, de vereisten die wenselijk zijn maar niet absoluut noodzakelijk. De **C** vertegenwoordigt ”could have”, de optie die een meerwaarde biedt indien mogelijk, maar niet cruciaal is. Tot slot staat de **W** voor ”won’t have”, wat verwijst naar de vereisten die niet zullen worden behandeld in dit project, hoewel ze mogelijk in de toekomst worden ontwikkeld.

4.1. Must Have

1. **De gekozen tools en resources moeten gratis toegankelijk zijn.**

   Alle software en tools die gebruikt worden in dit project moeten volledig gratis toegankelijk zijn. Dit betekent dat er geen kosten verbonden mogen zijn aan licenties of gebruik. Het gebruik van gratis tools garandeert dat het project toegankelijk blijft voor iedereen zonder financiële barrières, en zorgt ervoor dat de implementatie en het onderhoud kosteneffectief blijven. Voorbeelden van gratis tools zijn die welke onder open-source licenties vallen, zoals de GNU General Public License (GPL) of de MIT-licentie.

2. **Model of script voor header extractie**

   Er moet een model of script worden ontwikkeld dat in staat is om headers te extraheren uit tekstbestanden. Headers zijn vaak belangrijk voor het identifi-

ceren en structureren van gegevens in bestanden, zoals tabelkoppen of sectietitels. Het model of script moet betrouwbaar en accuraat headers kunnen identificeren en extraheren, zodat de gegevens correct kunnen worden verwerkt en opgeslagen.

3. **Model of script voor het extraheren van tabel data uit de tekstbestanden**

   Naast header extractie moet er een model of script beschikbaar zijn voor het extraheren van tabulaire gegevens uit tekstbestanden. Dit betekent dat het script in staat moet zijn om gestructureerde data, zoals rijen en kolommen in tabellen, correct te identificeren en te extraheren. Dit is essentieel voor het omzetten van ongestructureerde gegevens naar een gestructureerd formaat dat verder kan worden geanalyseerd of opgeslagen.

4. **Databank om ge-extraheerde data in op te slaan**

   Een database is nodig om de data die uit de tekstbestanden wordt geëxtraheerd op te slaan. Deze databank moet in staat zijn om de gestructureerde gegevens op een georganiseerde manier te bewaren, zodat ze eenvoudig kunnen worden geraadpleegd, geanalyseerd of verder verwerkt. De databank moet bovendien voldoende capaciteit en functionaliteit bieden om aan de vereisten van het project te voldoen.

4.2. Should Have

1. **Script dat de data (13F-meldingen) downloadt van de SEC en voorbereidt op data-extractie**

   Er moet een script beschikbaar zijn dat automatisch 13F-meldingen van de SEC downloadt. 13F-meldingen zijn rapporten die worden ingediend door institutionele beleggers en bevatten informatie over hun beleggingen. Dit script moet niet alleen de gegevens downloaden, maar ook voorbereiden op extractie door bijvoorbeeld de bestanden te parseren of te converteren naar een geschikt formaat.

2. **Er moet gebruik worden gemaakt van open-source software of diensten met een volledig gratis licenties.**

   Het is belangrijk dat alle software en diensten die worden gebruikt in dit project open-source zijn of een volledig gratis licentie hebben. Dit betekent dat de broncode beschikbaar moet zijn en er geen verborgen kosten of verplichtingen aan het gebruik van de software verbonden mogen zijn. Dit bevordert transparantie, aanpasbaarheid, en zorgt ervoor dat het project kosteneffectief blijft zonder juridische complicaties.

4.3. Won't Have

1. **Het ontwikkelen van een centrale interface voor het maken van een dashboard of vergelijkbare functionaliteiten**

   Het project omvat niet het ontwikkelen van een centrale interface voor het creëren van dashboards of andere visuele representaties van de gegevens. Dit betekent dat er geen functionaliteit wordt geïmplementeerd die zich richt op het presenteren van gegevens op een interactieve of visuele manier. De focus ligt puur op het extraheren en opslaan van gegevens, niet op hun presentatie of visualisatie.

2. **Betalende technieken**

   Er zullen geen betaalde technologieën of tools worden gebruikt in dit project. Alle gebruikte software, diensten, en tools moeten gratis zijn en mogen geen kosten met zich meebrengen. Dit sluit commerciële software en betaalde licenties uit, en garandeert dat het project volledig kosteloos blijft voor gebruikers en ontwikkelaars.

5

POC

5.1. Apparaten

In dit onderzoek zijn verschillende apparaten en omgevingen gebruikt om de benodigde taken uit te voeren en de prestaties te evalueren. Hieronder volgt een gedetailleerd overzicht van de specificaties van de gebruikte apparatuur, inclusief de laptop die als primaire werkstation fungeerde en de Google Colab-omgevingen die werden ingezet voor aanvullende rekenkracht en resources.
De laptop die voor de meeste van de berekeningen en data-analyse werd gebruikt, is de Dell XPS 15 9500.
Daarnaast zijn voor bepaalde taken en analyses Google Colab-omgevingen benut. Google Colab biedt zowel CPU- als GPU-resources. Deze omgevingen zijn vooral nuttig gebleken voor het uitvoeren van zwaardere berekeningen en experimenten.
Een overzicht van de specificaties van de gebruikte apparaten en omgevingen is weergegeven in Tabel 5.1.

Component	Specifications
1. 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
2. Google Colab Base	
OS	Ubuntu 20.04 TLS
RAM	13 GB
CPU	Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz
3. Google Colab GPU (free credits)	
GPU	Nvidia T4 15GB VRAM

Tabel 5.1: Specifications of Laptop and Google Colab Environments

5.2. Toegang en Libraries

Hier wordt er kort overlopen over wat er nodig is van toegangen en bibliotheken om de POC te kunnen uitvoeren

Toegang tot Llama

In bezit zijn van een Hugging Face aacount met toegang tot de Llama3(.1) modellen. Toegang kan verkregen worden de Llama model page van Hugging Face

Bibliotheken

1. **Python**: Zorg ervoor dat je een werkende installatie van Python hebt. Python is de programmeertaal die nodig is voor het uitvoeren van de code.
2. **Regex**: Voor reguliere expressies. Dit is standaard in Python en wordt vaak gebruikt voor patroonherkenning in tekst.
3. **Spacy**: Een krachtige NLP-bibliotheek voor tekstverwerking en natuurlijke taalverwerking.
4. **Pandas (pd)**: Voor gegevensmanipulatie en analyse.
5. **NumPy (np)**: Voor numerieke berekeningen en array-manipulatie.
6. **BeautifulSoup (bs4)**: Voor webscraping en het parseren van HTML en XML.
7. **psql**: De command-line interface voor PostgreSQL. Zorg ervoor dat je toegang hebt tot een PostgreSQL-database en dat je de juiste inloggegevens hebt.

8. **pyodbc**: Een ODBC-connector voor toegang tot databases via Python.
9. **Torch**: De PyTorch-bibliotheek voor machine learning en deep learning.

5.3. Data

In deze sectie zak er gesproken worden over de voorbereiding op de POC het gaat hier onder andere over de data verzamelen en voorbereiden.

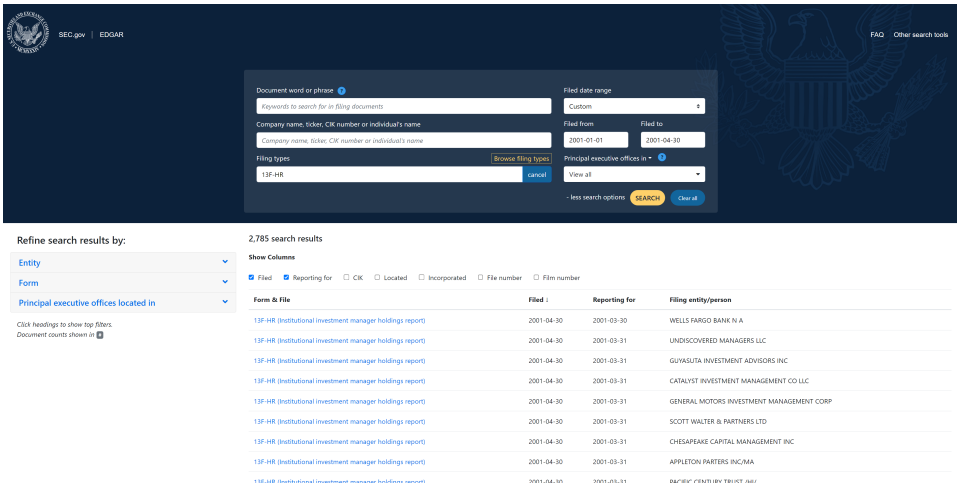
5.3.1. Data verzamelen

Als een subsectie van de POC behandelt dit segment het proces van het ophalen van 13F filings van voor 2013 met behulp van een web scraper. De scraper is ontworpen om het ophalen van deze historische deponeringen rechtstreeks uit SEC-archieven te automatiseren. Het primaire doel was om de bestanden efficiënt te vinden en te downloaden, ongeacht hun formaat (HTML, PDF, tekst). Door zich te richten op specifieke URL's en variaties in de bestandsstructuur te verwerken, haalde de scraper met succes de benodigde documenten op, zodat de gegevens vervolgens verwerkt en geanalyseerd konden worden.
Dit werd gedaan aan de hand van bs4 in volgende stappen:

1. **Stap 1** Begin met de basis-URL, https://www.sec.gov/edgar/search. Voeg parameters toe aan de URL om de zoekresultaten te verfijnen, zoals het type indiening en het datumbereik.
   Bijvoorbeeld: https://www.sec.gov/edgar/search/#/dateRange=custom#category=custom#star
   01-01#enddt=2001-04-30&forms=13F-HR.
   Dit wordt gedaan tot en met de einddatum 2012-12-31 vanaf deze datum word er met .xml gewerkt met incrementen van 4 maanden.
   Dit zorgt ervoor dat bestanden uit een specifiek datumbereik worden opgehaald. Omdat maximaal 10.000 bestanden per keer worden geretourneerd, worden de aanvragen opgesplitst in meerdere chunks.

   Gebruikte parameters:

   (a) **dateRange=custom**: Stelt een aangepast datumbereik in.
   (b) **category=custom**: Specificeert een aangepaste zoekcategorie.
   (c) **startdt=2001-01-01**: Geeft de startdatum van het datumbereik aan.
   (d) **enddt=2001-04-30**: Geeft de einddatum van het datumbereik aan.
   (e) **forms=13F-HR**: Filtert de resultaten op het type formulier 13F-HR.

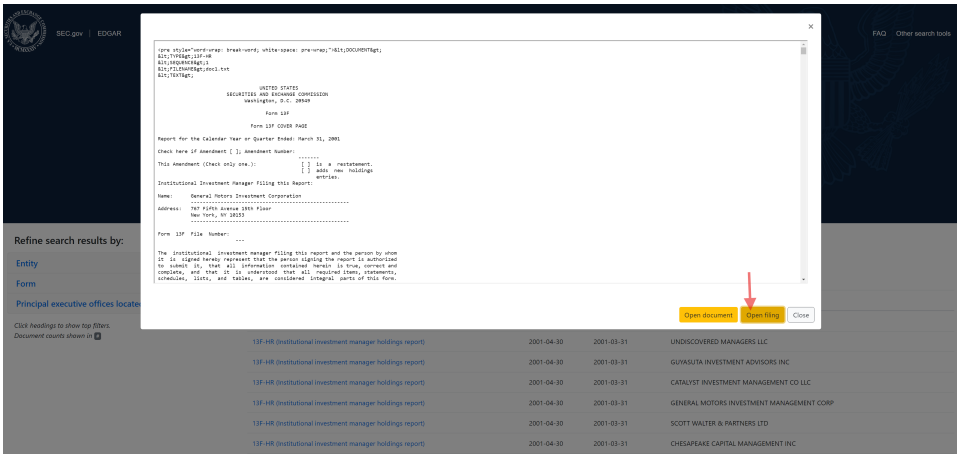


Figuur 5.1: Voorbeeld van de bovenkant van resultaatpagina na het uitvoeren van de zoekopdracht.

Figuur 5.2: Voorbeeld van de onderkant van resultaatpagina na het uitvoeren van de zoekopdracht.

Het resultaat van deze stap is de resultaatpagina van de query, zoals weergegeven in Figuur 5.1 en Figuur 5.2 , die meerdere pagina's bevat met elk maximaal 100 links naar unieke filings.

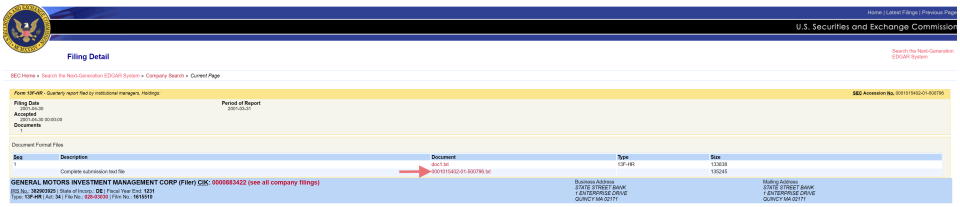
2. **Stap 2:** Op basis van het resultaat van de vorige stap wordt elke link op elke pagina gevolgd, wat resulteert in de weergave van een popup (zie Figuur 5.3) met de details van de filing. Vervolgens wordt op de knop in deze pop-up geklikt om toegang te krijgen tot het overzicht van de gerelateerde bestanden van de filing (zie Figuur 5.4).
   Resultaat: Het resultaat van deze stap is een overzicht van de bestanden die behoren tot de specifieke filing. In dit geval bevat het overzicht slechts twee bestanden, zoals geïllustreerd in Figuur 5.4.



Figuur 5.3: Pop-up van een filing.

Figuur 5.4: Voorbeeld van het overzicht van de bestanden gerelateerd aan één specifieke filing.

3. **Stap 3:** In deze stap wordt op de link geklikt die de tekst 'Complete submission file' bevat (zie **??**). Door op deze link te klikken, wordt het resulterende bestand gedownload, zoals weergegeven in **??**.



Figuur 5.5: Voorbeeld van het overzicht van de bestanden gerelateerd aan één specifieke filing.

```
-----BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twIDAQAB
MIC-Info: RSA-MD5,RSA,
 MVDtnhPAKXkTTseuRWB0ZBRbjfkx4RzjVIgQraZ0G++bEXVIQ+munShxLv+60MPg
 9Lu7WNZUSdce1dZiTvzELg==

<SEC-DOCUMENT>0001015402-01-500796.txt : 20010501
<SEC-HEADER>0001015402-01-500796.hdr.sgml : 20010501
ACCESSION NUMBER:               0001015402-01-500796
CONFORMED SUBMISSION TYPE:      13F-HR
PUBLIC DOCUMENT COUNT:          1
CONFORMED PERIOD OF REPORT:     20010331
FILED AS OF DATE:               20010430

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 GENERAL MOTORS INVESTMENT MANAGEMENT CORP
                CENTRAL INDEX KEY:                      0000883422
                STANDARD INDUSTRIAL CLASSIFICATION:      []
                IRS NUMBER:                             382903925
                STATE OF INCORPORATION:                 DE
                FISCAL YEAR END:                        1231

        FILING VALUES:
                FORM TYPE:              13F-HR
                SEC ACT:
                SEC FILE NUMBER:        028-03030
                FILM NUMBER:            1615510
```

Figuur 5.6: Voorbeeld van 13F bestand.

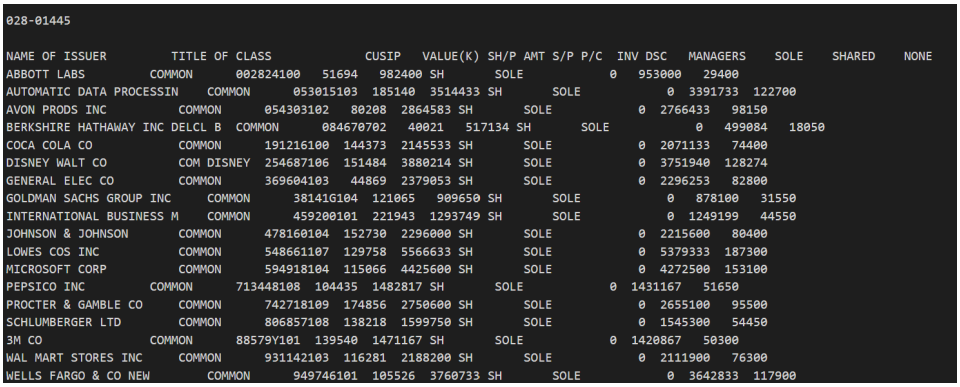
5.3.2. Data verwerking

De procedure voor het samenstellen van de dataset begon met het splitsen van de bestanden in twee afzonderlijke componenten: de koptekst Figuur 5.7 en de tabel Figuur 5.10 met als essentiële het unieke bestandsnummer. Vervolgens werd het tabelbestand opgeschoond, wat inhield dat HTML-tags werden verwijderd, entries die over meerdere lijnen waren verspreid op een enkele regel werden gezet, lege lijnen werden verwijderd, en het bestandsnummer werd toegevoegd aan de eerste rij van het bestand.

De tabelgegevens werden vervolgens op een methodische manier georganiseerd. Deze gestructureerde tabelgegevens werden, samen met het bijbehorende oorspronkelijke bestand, gebruikt om training sets te genereren. De training set bestond uit twee componenten: het originele (opgeschoonde) bestand Figuur 5.8 en een georganiseerd CSV-bestand Figuur 5.10 dat de opgeschoonde tabel bevatte. Alsook wordt er een 'Gouden Dataset gemaakt' deelparagraaf 2.4.3 om het model te kunnen evalueren. Deze methodologie garandeerde dat de dataset nauwkeurig gestructureerd was, waardoor verdere verwerking, analyse en modellering mogelijk was.

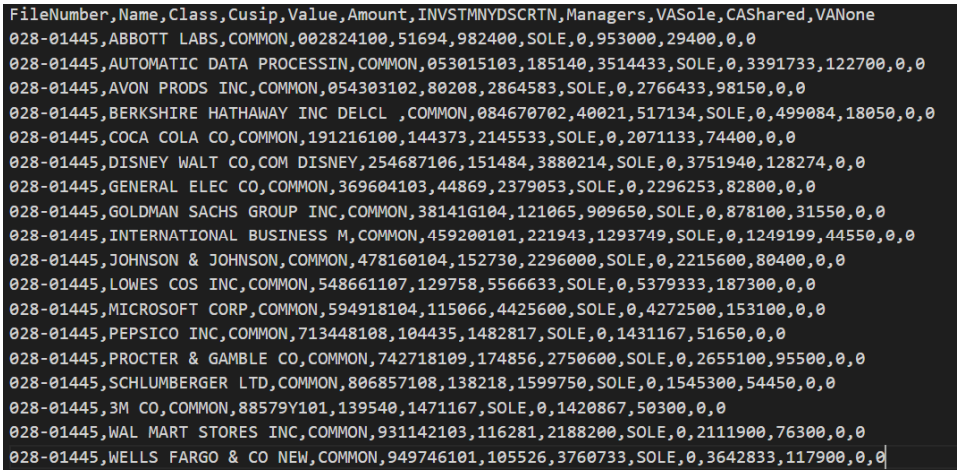
```
<DOCUMENT>
<TYPE>13F-HR
<SEQUENCE>1
<FILENAME>bkd2q09.txt
<DESCRIPTION>BDK WEALTH ADVISORS LLC
<TEXT>
                              UNITED STATES
                    SECURITIES AND EXCHANGE COMMISSION
                         Washington, D.C.  20549

                               Form 13F

                          Form 13F COVER PAGE

Report for the Calendar Year or Quarter Ended: June 30, 2009

Check here if Amendment [  ]; Amendment Number:
This Amendment (Check only one.): [  ] is a restatement.
                                  [  ] adds new holdings entries.

Institutional Investment Manager Filing this Report:

Name:    BKD Wealth Advisors, LLC
Address: 1700 Lincoln Street, Suite 1450
         Denver, CO  80203

13F File Number:  28-11934

The institutional investment manager filing this report and the person by whom
it is signed hereby represent that the person signing the report is authorized
to submit it, that all information contained herein is true, correct and
complete, and that it is understood that all required items, statements,
schedules, lists, and tables, are considered integral parts of this form.

Person Signing this Report on Behalf of Reporting Manager:

Name:      Tod Eastlake
Title:     Operations Manager
Phone:     417.831.7283

Signature, Place, and Date of Signing:

/s/  Tod Eastlake     Springfield, MO     July 23, 2009

Report Type (Check only one.):

[ X]         13F HOLDINGS REPORT.

[  ]         13F NOTICE.

[  ]         13F COMBINATION REPORT.

<PAGE>

                          FORM 13F SUMMARY PAGE

Report Summary:

Number of Other Included Managers:         0

Form13F Information Table Entry Total:     98

Form13F Information Table Value Total:     $90,214 (thousands)

List of Other Included Managers:

Provide a numbered list of the name(s) and Form 13F file number(s) of all
institutional managers with respect to which this report is filed, other
than the manager filing this report.

NONE
```

Figuur 5.7: Header van een filing



```
028-01445

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
DISNEY WALT CO          COM DISNEY  254687106  151484  3880214 SH       SOLE           0  3751940  128274
GENERAL ELEC CO         COMMON      369604103   44869  2379053 SH       SOLE           0  2296253   82800
GOLDMAN SACHS GROUP INC     COMMON      38141G104  121065   909650 SH       SOLE           0   878100   31550
INTERNATIONAL BUSINESS M    COMMON      459200101  221943  1293749 SH       SOLE           0  1249199   44550
JOHNSON & JOHNSON       COMMON      478160104  152730  2296000 SH       SOLE           0  2215600   80400
LOWES COS INC           COMMON      548661107  129758  5566633 SH       SOLE           0  5379333  187300
MICROSOFT CORP          COMMON      594918104  115066  4425600 SH       SOLE           0  4272500  153100
PEPSICO INC         COMMON      713448108  104435  1482817 SH       SOLE           0  1431167   51650
PROCTER & GAMBLE CO     COMMON      742718109  174856  2750600 SH       SOLE           0  2655100   95500
SCHLUMBERGER LTD        COMMON      806857108  138218  1599750 SH       SOLE           0  1545300   54450
3M CO               COMMON      88579Y101  139540  1471167 SH       SOLE           0  1420867   50300
WAL MART STORES INC     COMMON      931142103  116281  2188200 SH       SOLE           0  2111900   76300
WELLS FARGO & CO NEW        COMMON      949746101  105526  3760733 SH       SOLE           0  3642833  117900
```

Figuur 5.8: Voorbeeld van originele tabel van een filing.



```
FileNumber,Name,Class,Cusip,Value,Amount,INVSTMNYDSCRTN,Managers,VASole,CAShared,VANone
028-01445,ABBOTT LABS,COMMON,002824100,51694,982400,SOLE,0,953000,29400,0,0
028-01445,AUTOMATIC DATA PROCESSIN,COMMON,053015103,185140,3514433,SOLE,0,3391733,122700,0,0
028-01445,AVON PRODS INC,COMMON,054303102,80208,2864583,SOLE,0,2766433,98150,0,0
028-01445,BERKSHIRE HATHAWAY INC DELCL ,COMMON,084670702,40021,517134,SOLE,0,499084,18050,0,0
028-01445,COCA COLA CO,COMMON,191216100,144373,2145533,SOLE,0,2071133,74400,0,0
028-01445,DISNEY WALT CO,COM DISNEY,254687106,151484,3880214,SOLE,0,3751940,128274,0,0
028-01445,GENERAL ELEC CO,COMMON,369604103,44869,2379053,SOLE,0,2296253,82800,0,0
028-01445,GOLDMAN SACHS GROUP INC,COMMON,38141G104,121065,909650,SOLE,0,878100,31550,0,0
028-01445,INTERNATIONAL BUSINESS M,COMMON,459200101,221943,1293749,SOLE,0,1249199,44550,0,0
028-01445,JOHNSON & JOHNSON,COMMON,478160104,152730,2296000,SOLE,0,2215600,80400,0,0
028-01445,LOWES COS INC,COMMON,548661107,129758,5566633,SOLE,0,5379333,187300,0,0
028-01445,MICROSOFT CORP,COMMON,594918104,115066,4425600,SOLE,0,4272500,153100,0,0
028-01445,PEPSICO INC,COMMON,713448108,104435,1482817,SOLE,0,1431167,51650,0,0
028-01445,PROCTER & GAMBLE CO,COMMON,742718109,174856,2750600,SOLE,0,2655100,95500,0,0
028-01445,SCHLUMBERGER LTD,COMMON,806857108,138218,1599750,SOLE,0,1545300,54450,0,0
028-01445,3M CO,COMMON,88579Y101,139540,1471167,SOLE,0,1420867,50300,0,0
028-01445,WAL MART STORES INC,COMMON,931142103,116281,2188200,SOLE,0,2111900,76300,0,0
028-01445,WELLS FARGO & CO NEW,COMMON,949746101,105526,3760733,SOLE,0,3642833,117900,0,0
```

Figuur 5.9: Voorbeeld van een verwerkte tabel van een filing.

Tijdsinschatting

In deze sectie wordt de benodigde hoeveelheid trainingsdata en de tijd die nodig is om deze te verwerken, berekend. Voor de schatting worden de volgende parameters in overweging genomen:

1. **Verwerkingsduur per filing:** 5 minuten, dit houd in het opkuisen van de data en het opstellen van de
2. **Aantal filings per jaar:** 4
3. **Aantal bedrijven (incl. marge):** 505
4. **Aantal jaren:** 12
5. **Hoeveel percent van originele data**: 0,5% Zoals vermeld in subsectie: Data vereisten voor LLMs te trainen 2.4.4 van de literatuurstudie.

Het totaal aantal bestanden, niet meegerekend dat een bedrijf ene dubbelle filing kan doen:

Totaal Aantal Bestanden = 4 filings/jaar × 505 bedrijven × 12 jaren = 24240 bestanden

Het minimaal aantal bestanden nodig om een LLM effectieve te trainen:

Minimaal Aantal Bestanden = 24240 (Totaal bestanden)×0.005 (0.5% van totaal) ≈ 121.2 bestanden

De totale tijdsduur voor de verwerking van alle gegevens kan als volgt worden berekend:

$$\text{Totaal tijd} = \frac{122 \text{ bestanden, afgerond} \times 5 \text{ minuten per bestand}}{60 \text{ minuten per uur}} \approx 10 \text{ uur en} 10 \text{ minuten}$$

Dit geeft een indicatie van de hoeveelheid tijd die nodig is voor het verwerken van de gegevens voor de hele periode. Dit is essentieel voor het plannen van de benodigde middelen en het vaststellen van de haalbaarheid van het project.

5.4. 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 Al de technieken zijn gedaan geweest met het zelfde bestand voor zowel de header Figuur 5.7 als voor de informatie tabel Figuur 5.8.

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.

1. **Voordelen**:
   (a) Nauwkeurig voor kleine datasets
   (b) Simpel
2. **Nadelen**:
   (a) Niet praktisch voor grote datasets
   (b) Tijdrovend
   (c) menselijke fouten
   (d) Niet schaalbaar

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

```
ABBOTT LABS,COMMON,002824100,51694,982400,SOLE,0,953000,29400,0,0
AUTOMATIC DATA PROCESSIN,COMMON,053015103,185140,3514433,SOLE,0,3391733,122700,0,0
AVON PRODS INC,COMMON,054303102,80208,2864583,SOLE,0,2766433,98150,0,0
BERKSHIRE HATHAWAY INC DELCL ,COMMON,084670702,40021,517134,SOLE,0,499084,18050,0,0
COCA COLA CO,COMMON,191216100,144373,2145533,SOLE,0,2071133,74400,0,0
DISNEY WALT CO,COM DISNEY,254687106,151484,3880214,SOLE,0,3751940,128274,0,0
GENERAL ELEC CO,COMMON,369604103,44869,2379053,SOLE,0,2296253,82800,0,0
GOLDMAN SACHS GROUP INC,COMMON,38141G104,121065,909650,SOLE,0,878100,31550,0,0
INTERNATIONAL BUSINESS M,COMMON,459200101,221943,1293749,SOLE,0,1249199,44550,0,0
JOHNSON & JOHNSON,COMMON,478160104,152730,2296000,SOLE,0,2215600,80400,0,0
LOWES COS INC,COMMON,548661107,129758,5566633,SOLE,0,5379333,187300,0,0
MICROSOFT CORP,COMMON,594918104,115066,4425600,SOLE,0,4272500,153100,0,0
PEPSICO INC,COMMON,713448108,104435,1482817,SOLE,0,1431167,51650,0,0
PROCTER & GAMBLE CO,COMMON,742718109,174856,2750600,SOLE,0,2655100,95500,0,0
SCHLUMBERGER LTD,COMMON,806857108,138218,1599750,SOLE,0,1545300,54450,0,0
3M CO,COMMON,88579Y101,139540,1471167,SOLE,0,1420867,50300,0,0
WAL MART STORES INC,COMMON,931142103,116281,2188200,SOLE,0,2111900,76300,0,0
WELLS FARGO & CO NEW,COMMON,949746101,105526,3760733,SOLE,0,3642833,117900,0,0
```

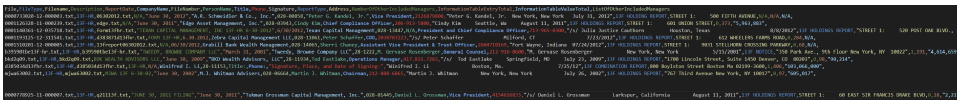
Figuur 5.10: Voorbeeld van een verwerkte tabel van een filing.

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 semi-gestructureerde gegevens.

1. **Voordelen**:
   (a) Flexibel: Regex laat toe om op maat gemaakte patronen maken die passen bij een grote verscheidenheid aan gegevens formaten.
   (b) Integratie: Regex is gemakkelijk te integreren in bestaande software
2. **Nadelen**:
   (a) Complex: naarmate patronen ingewikkelder worden, kan regex moeilijk te lezen en onderhouden beginnen worden.
   (b) 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.

Resultaat header:



Figuur 5.11: Regex header extraction

Resultaat tabel:

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). Bij het extraheren van data uit de tabel werd dit echter al zeer snel complex en niet meer onderhoudbaar en werd aldus stopgezet door de grote variëteit in opmaak en structuur, een missende/extra waarden, indentatie verschillen en extra spaties. Maar het REGEX deed het zeer goed in het extraheren van de data uit de header. Hierdoor werd regex niet gebruikt voor het extraheren van de tabel informatie maar wel voor de algemene informatie van het indienende bedrijf.

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

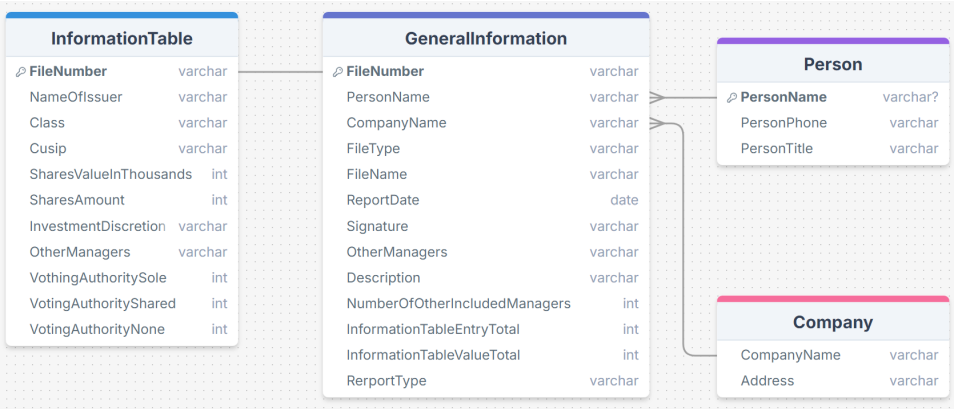
Llama presteerde het beste in vergelijking met alle andere technieken, maar ondervindt nog steeds moeilijkheden wanneer het wordt geconfronteerd met structuren die aanzienlijk afwijken van wat het eerder heeft gezien. Ondanks deze uitdaging, heeft Llama echter wel de capaciteit om effectief om te gaan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

heeft aangetroffen.

Conclusie
Na evaluatie van de technieken voor gegevensextractie blijkt dat REGEX en Llama de beste resultaten opleveren voor hun specifieke taken. REGEX is bijzonder effectief gebleken voor het extraheren van gegevens uit headers. Zijn vermogen om op maat gemaakte patronen te herkennen maakt het geschikt voor het accuraat extraheren van headerinformatie.
Llama blijkt de meest robuuste keuze voor het extraheren van gegevens uit tabellen. Het kan effectief omgaan met ontbrekende waarden en complexe structuren, wat essentieel is voor de verwerking van de tabellen in deze dataset. Hoewel Llama uitdagingen ondervindt bij ongebruikelijke structuren, biedt het de beste prestaties voor tabellen.
In samenvatting zal REGEX worden ingezet voor headerextractie, terwijl Llama zal worden gebruikt voor tabelgegevens. Deze aanpak benut de sterke punten van beide technieken en optimaliseert de efficiëntie van de gegevensverwerking.

5.5. Databank

In dit deel van de POC word de relationele databank ontworpen om de geëxtraheerde gegevens in op te beheren en organiseren. De databank zal bestaan uit 4 tabellen: 'Company', 'Person', 'InformationTabel' en 'GeneralInformation'. Dit schema zal dienen als basis voor gegevensextractie- en beheer.



Figuur 5.12: Regex header extraction

'Company' Tabel
De **Company** tabel slaat informatie op over verschillende bedrijven. Elke record in deze tabel vertegenwoordigt een afzonderlijk bedrijf.

1. **CompanyName**: De naam van het bedrijf (Primaire Sleutel).
2. **Address**: Het adres van het bedrijf.

'Person' Tabel
De **Person** tabel bevat gegevens over individuen die zijn gekoppeld aan de bestanden. Elke record vertegenwoordigt een afzonderlijk persoon.

1. **PersonName**: De volledige naam van de persoon (Primaire Sleutel).
2. **PersonPhone**: Het telefoonnummer van de persoon.
3. **PersonTitle**: De functietitel van de persoon.

'GeneralInformation' Tabel
De **GeneralInformation**tabel slaat uitgebreide informatie op die betrekking heeft op elk bestand. Dit omvat details over het bestand zelf, zoals het type, de naam, de handtekening, en aanvullende informatie.

1. **FileNumber**: 13F bestand nummer (Primaire Sleutel).
2. **PersonName**: De naam van de persoon die aan het bestand is gekoppeld (Verwijst naar `Person(PersonName)`).
3. **CompanyName**: De naam van het bedrijf dat aan het bestand is gekoppeld (Verwijst naar `Company(CompanyName)`).
4. **FileType**: Het type bestand (13F).
5. **FileName**: De naam van het bestand.
6. **ReportDate**: De datum waarop het rapport is opgesteld.
7. **Signature**: De handtekening op het bestand.
8. **OtherManagers**: Beschrijvende tekst over andere managers die aan het bestand zijn gekoppeld.
9. **Description**: Een beschrijving van het bestand.
10. **NumberOfOtherIncludedManagers**: Het aantal andere managers dat in het bestand is opgenomen.
11. **InformationTableEntryTotal**: Het totaal aantal vermeldingen in de informatie tabel.
12. **InformationTableValueTotal**: Het totaal van de waarden van de aandelen in de informatie tabel.
13. **ReportType**: Het type rapport.

'InformationTable' Tabel
De **InformationTable** bevat gedetailleerde informatie over financiële en beheersaspecten die verband houden met elk bestand.

1. **FileNumber**: De unieke identificator voor elk bestand (Verwijst naar `General-Information(FileNumber)`).
2. **NameOfIssuer**: De naam van de uitgever (van de aandelen).
3. **Class**: De klasse van de uitgever.
4. **Cusip**: Het CUSIP nummer.
5. **SharesValueInThousands**: De waarde van de aandelen in duizenden.
6. **SharesAmount**: Het aantal aandelen.
7. **InvestmentDiscretion**: De investeringsdiscretie.
8. **OtherManagers**: Beschrijvende tekst over andere managers.
9. **VotingAuthoritySole**: Stemautoriteit uitsluitend.
10. **VotingAuthorityShared**: Stemautoriteit gedeeld.
11. **VotingAuthorityNone**: Geen stemautoriteit.

5.6. Implementatie

In deze sectie van de proof of concept wordt het proces beschreven voor het extraheren van gegevens uit de 13F-rapporten, zoals eerder besproken in de sectie over gegevensverwerking. De implementatie is verdeeld in drie hoofdsecties: header-extractie, tabelextractie, en de invoer in de database.

5.6.1. Header data extractie

De header van elk bestand bevat essentiële informatie, zoals het bestandnummer, de naam van het bedrijf en andere meta-informatie die cruciaal is voor verdere verwerking en organisatie. Voor de extractie van deze gegevens zijn reguliere expressies (regex) gebruikt, een techniek die bijzonder effectief blijkt te zijn voor het extraheren van gestructureerde informatie uit de headers vanwege het voorspelbare patroon van de informatie.
De eerste stap in dit proces is het identificeren van de relevante metadata. Dit omvat gegevens zoals de naam van het bedrijf, algemene informatie over het bestand en extra details die zich bevinden rondom de informatie tabel. Het zorgvuldig bepalen van deze gegevens is van groot belang om te verzekeren dat alle relevante informatie correct wordt verzameld en verwerkt.

Na het identificeren van de relevante metadata worden patronen voor extractie ontwikkeld door middel van reguliere expressies. Deze patronen zijn zorgvuldig ontworpen om te matchen met de verschillende gegevensvelden die in de header worden verwacht. Het gebruik van regex maakt het mogelijk om flexibel en efficiënt door de tekst te navigeren en specifieke informatie te isoleren op basis van het gestructureerde formaat van de gegevens.
Vervolgens worden de gedefinieerde patronen toegepast op de tekst van de header. De reguliere expressies worden uitgevoerd op elke headertekst om de gegevens te identificeren en te extraheren. Dit proces houdt in dat gezocht wordt naar overeenkomsten in de tekst die voldoen aan de gedefinieerde structuren, wat het mogelijk maakt om snel en nauwkeurig de gewenste informatie te vinden.
Wanneer de regex-patronen overeenkomsten vinden, worden de relevante gegevens uit de tekst gehaald en verzameld. In gevallen waar geen tekstuele overeenkomsten worden gevonden, worden deze gevallen genoteerd als N/A (Not Applicable). Deze aanpak helpt om een compleet overzicht te krijgen van de gegevens die niet konden worden geëxtraheerd, wat belangrijk is voor latere analyses en correcties.
De verzamelde gegevens worden vervolgens opgeslagen in een CSV-bestand. Dit bestand fungeert als een tussenoplossing en maakt het mogelijk om de gegevens op een gestructureerde manier te bewaren. Later kunnen deze gegevens in bulk naar de databank worden geschreven, waar ze verder kunnen worden verwerkt en geïntegreerd in de bestaande gegevenssystemen. Het gebruik van een CSV-bestand zorgt ervoor dat de gegevens gemakkelijk kunnen worden gecontroleerd en beheerd voordat ze definitief worden overgedragen aan de databank.
Door deze gestructureerde aanpak kan de data-integriteit worden gewaarborgd en kan een efficiënte verwerking van gegevens worden gegarandeerd. Dit draagt uiteindelijk bij aan een beter beheer en een verbeterde toegang tot cruciale informatie.

5.6.2. Table data extractie

1. **Voorbereiding van de Trainingsdata**: De eerste stap betrof de voorbereiding van de trainingsdata. Hiervoor werden zowel de oorspronkelijke gegevens als de opgeschoonde versies gebruikt. Deze combinatie van datasets biedt een uitgebreid spectrum aan voorbeelden, wat het model helpt om robuust te leren en te generaliseren over verschillende datastijlen en formaten.

2. **Model Laden**: Het LLaMA 3.18B-model werd gedownload vanuit de Unsloth-repository. Tijdens het laden van het model werd de parameter `params_max_seq_length` ingesteld op basis van de vereiste sequentielengte voor de specifieke data. Deze instelling is cruciaal om ervoor te zorgen dat het model effectief omgaat met de lengte van de inputgegevens.

3. **Model Initialiseren**: Na het laden van het model werd het geconfigureerd met de nodige parameters die specifiek zijn afgestemd op de tabelextractie-taak. Dit omvatte het afstemmen van hyperparameters en andere instellingen om de prestaties van het model te optimaliseren voor de taak van gegevensextractie.

4. **Dataset Structureren**: De dataset werd gestructureerd volgens een format dat lijkt op de Alpaca-prompt. Deze prompt zorgt voor een gestructureerde en herhaalbare manier om instructies, context en verwachte antwoorden te definiëren. Het format is als volgt:

```
alpaca_prompt = """Hieronder staat een instructie die een taak beschrijft, same

### Instructie:
{}

### Invoer:
{}

### Reactie:
{}"""
```

   Dit formaat helpt om duidelijke en consistente voorbeelden te genereren voor het trainen van het model, wat bijdraagt aan een betere prestaties bij het uitvoeren van de extractietaak.

5. **Model Training**: Het model werd getraind gedurende 60 epochs met een batchgrootte van 2. Tijdens het trainen werd de AdamW-optimalisator met 8-bits precisie gebruikt, met een leerschema ingesteld op 2e-4. Zowel trainings- als validatiesets werden ingezet om de modelprestaties te optimaliseren en overfitting te voorkomen. De training werd zorgvuldig gevolgd om te waarborgen dat het model goed presteerde op zowel bekende als nieuwe data.

6. **Inferentie Uitvoeren**: Na de training werd het model opgeslagen voor latere toepassing. Dit betekent dat het model kan worden geladen om inferentie uit te voeren. Tijdens de inferentie wordt het model toegepast op nieuwe gegevens om voorspellingen te doen of gegevens te extraheren op basis van de getrainde kennis.

7. **Input Formatteren voor Inferentie**: De input voor inferentie werd geformatteerd met behulp van de tokenizer volgens het volgende sjabloon:

```
inputs = tokenizer(
```

```
[
alpaca_prompt.format(
"Extraheer de gegevens uit de informatie tabel als CSV", # Instructie
f"{input}", # Invoer
"", # Output - laat dit leeg voor generatie!
)
], return_tensors = "pt").to("cuda")
```

Deze formattering zorgt ervoor dat de gegevens correct worden geïnterpreteerd door het model, wat essentieel is voor het verkrijgen van nauwkeurige resultaten.

8. **Evaluatie van het Model**: De evaluatie van het model werd uitgevoerd door het gebruik van een gouden dataset, een zorgvuldig samengestelde dataset met een hoge kwaliteit en nauwkeurigheid. Deze gouden dataset fungeerde als een referentie om de prestaties van het model te meten. De resultaten van het model werden vergeleken met deze gouden standaard om te beoordelen hoe goed het model de gegevens kon extraheren en verwerken.

9. **Resultaten Opslaan**: De uiteindelijke resultaten van de extractie werden opgeslagen in CSV-bestanden. Elk bestand werd benoemd op basis van het bestandnummer en opgeslagen in een aangewezen map. Deze gestructureerde aanpak zorgt ervoor dat de resultaten gemakkelijk toegankelijk en controleerbaar zijn voor verdere analyse en rapportage.

10. **Toevoegen aan de databank:** Na het opslaan van de resultaten werden alle CSV-bestanden in een opgegeven map doorgelopen en toegevoegd aan de database. Dit werd bereikt met behulp van een script dat elke CSV in de map doorloopt, de gegevens inleest en deze gegevens vervolgens naar de database pushte. Het script zorgde ervoor dat alle gegevens op een gestructureerde manier in de database werden opgeslagen, wat de latere toegang en analyse vergemakkelijkte.

5.7. Conclusie

TODO

6

Conclusie

6.1. Conclusie

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

A

Onderzoeksvoorstel

A.1. Inleiding

A.1.1. Achtergrond en Context

13F-meldingen, bij de SEC zijn ingediend, bevatten essentiële informatie over de beleggingsportefeuilles van institutionele investeerders en zijn van cruciaal belang voor financieel onderzoek en investeringsanalyse. Maar voorafgaand aan 2013 vertonen 13F-rapporten vaak inconsistenties in formaat en structuur, waardoor handmatige verwerking extreem tijdrovend en foutgevoelig is.
AI-technologieën zoals NLP en ML kunnen helpen deze oudere documenten te standaardiseren en vervolgens te integreren in een gestructureerde databank. Dit zou de efficiëntie van gegevensverwerking verbeteren en de toegankelijkheid van historische financiële data vergroten. Een proof-of-concept applicatie die deze AI-technieken toepast, zal niet alleen de analyse van historische beleggingstrends vergemakkelijken, maar ook het ontwikkelen van voorspellende modellen eenvoudiger maken.

A.1.2. 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. Dit onderzoek gaat opzoek naar hoe AI-technologieën zoals NLP en ML, ingezet kunnen worden om deze meldingen te extraheren, te structuren en te integreren in een databank, wat als gevolg het gebruik en de toegankelijkheid van historische financiële gegevens te verbeteren.

A.1.3. Hoofonderzoeksvraag

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?

A.1.4. Deelonderzoeksvragen

1. Wat zijn de potentiële voordelen en beperkingen van het gebruik van AI-technologieën voor dit doel vergeleken met traditionele methoden?
2. Wat zijn de belangrijkste uitdagingen bij het standaardiseren van de verschillende formaten en structuren van 13F-meldingen?
3. Hoe kan de ontwikkelde proof-of-concept worden gevalideerd en geëvalueerd op basis van nauwkeurigheid, efficiëntie en bruikbaarheid?

A.1.5. 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.

A.2. Literatuurstudie

SEC Textmining LLM

A.3. Methodologie

Dit onderzoek richt zich op het ontwikkelen van een proof-of-concept applicatie die AI- technologieën, zoals Natural Language Processing (NLP) en Machine Learning (ML), gebruikt om 13F-meldingen van vóór 2013 te standaardiseren en te integreren in een relationele databank. De methodologie omvat vier hoofdfasen: literatuurstudie, systeemontwikkeling, evaluatie, en implementatie.
In de eerste fase zal de literatuurstudie worden voorbereid, deze zal zich focussen op het analyseren van bestaande technieken en benaderingen te analyseren. Dit zal bestaan uit het verkennen van relevante NLP-technieken zoals Named Entity Recognition (NER), tekstclassificatie en tokenisatie, die nuttig kunnen zijn voor het extraheren van de nodige gegevens. Alssok zal er een analyse gedaan worden naar al bestaande modellen en bibliotheken zoals BERT, GPT en Spacy.

Op basis van de bevindingen uit de eerste fase zal er een proof-of-concept systeem ontwikkeld met de volgende stappen:

1. **Data Voorbereiding:** Verzamelen en voorbereiden van een dataset van 13F-meldingen van vóór 2013. Dit kan bestaan uit het downloaden van historische rapporten en het opschonen van gegevens om consistentie en kwaliteit te waarborgen.

2. **NLP- en ML-implementatie:** Het tepassen van NLP-technieken voor het extraheren van relevante informatie zoals bedrijfsnamen, aandelen en aantallen. Vervolgens worden ML-modellen getraind om patronen en structuren te herkennen, en om de gegevens te classificeren en te structureren.

3. **Integratie:** Integreren van deze gegevens in een relationele databank die ontworpen is voor efficiënte opslag en toegang.

In de derde fase zal het systeem worden geëvalueerd op basis van enkele criteria: accuraatheid en efficiëntie en kwaliteit van de geëxtraheerde gegevens.
De resultaten van de gegevensextractie worden vergeleken met handmatig gecodeerde gegevens en de verwerkingstijd om de efficiëntie en nauwkeurigheid te evalueren.
De kwaliteit van de gegevens wordt gemeten door fouten en inconsistenties in de geëxtraheerde en genormaliseerde gegevens te vinden, naast de consistentie en volledigheid van de gestandaardiseerde gegevens.
Na evaluatie van het proof-of-concept systeem, worden de bevindingen gepresenteerd en aanbevelingen gedaan voor verdere verbeteringen en mogelijke toepassingen. Dit kan ook aanbevelingen omvatten voor bredere implementatie, zoals integratie met andere financiële analysetools en verdere verfijning van de AI-modellen op basis van feedback en aanvullende gegevens.
Deze gestructureerde aanpak zorgt ervoor dat het proof-of-concept systeem effectief en efficiënt de historische 13F-meldingen kan verwerken, waardoor de toegankelijkheid en analyse van historische financiële gegevens wordt verbeterd.

A.4. Verwachte resultaten, conclusie

Het verwachte resultaat van het onderzoek is een werkende proof-of-concept applicatie te ontwikkelen die AI-technologieën gebruikt, waaronder NLP en ML- technologieën, om alle 13f-meldingen van voor 2013 te standaardiseren en integreren in een relationele databank. De applicatie die wordt ontwikkeld moet de gegevens binnen ene acceptabele tijd extraheren en verwerken naar een uniform formaat en vervolgens naar een databank weg te schrijven. Het gevolg hiervan is dat de toegankelijkheid en analyse van de historische financiële gegevens worden verbeterd en vergemakkelijkt. Hierdoor kunnen onderzoekers met minder inspanning

en kosten diepere inzichten verkrijgen in historische beleggingstrends en gemakkelijker voorspellende modellen maken.

Kortom, AI-technologieën zoals NLP en ML kunnen een machtige oplossing bieden bij het standaardiseren en normaliseren van historische 13F-meldingen. Het systeem zal automatisch de inconsistenties in dergelijke documenten op. Het daaropvolgende bewijs-of-concept systeem zal een waardevolle input zijn voor financieel onderzoek en investeringsanalyse en zal fungeren als basis voor toekomstige toepassingen in de analyse van historische financiële data-analyse en de ontwikkeling van voorspellende modellen.

B

Bijlagen

Bibliografie

Achimugu, P., Selamat, A., Ibrahim, R., & Mahrin, M. N. (2014). A systematic literature review of software requirements prioritization research. *Information and Software Technology*, *56*(6), 568–585. https://doi.org/https://doi.org/10.1016/j.infsof.2014.02.001

Amade, D., Chandra, R., Sinha, V. K., & Anand, D. (2024). Automatic Text Summarization Using NLTK & Spacy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4742012

AWS. (2024). *What is the difference between structured and unstructured data?* Verkregen augustus 22, 2024, van https://aws.amazon.com/compare/the-difference-between-structured-data-and-unstructured-data/

Baker, B. (2022, augustus 23). *6 important SEC filings every stock investor should know about*. Verkregen augustus 22, 2024, van https://www.bankrate.com/investing/sec-filings-stock-investors-should-know/

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*(3), 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), 42–45. https://doi.org/10.5120/14937-3507

Ganesh, V. (2024, juli 10). *SEC filing*. Verkregen augustus 22, 2024, van https://en.wikipedia.org/wiki/SEC_filing

Graph, R. (2024, augustus 9). *Llama 3.1 405B vs GPT-4o: Which model is better?* Verkregen augustus 22, 2024, van https://medium.com/@researchgraph/llama-3-1-405b-vs-gpt-4o-which-model-is-better-659662234b3e

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

Hayes, A. (2024, april 27). *What Is a CUSIP Number, and How Do I Find a Stock or Bond CUSIP?* Verkregen augustus 22, 2024, van https://www.investopedia.com/terms/c/cusipnumber.asp

Huang, J. (2024, augustus 22). *Evaluating Large Language Model (LLM) systems: Metrics, challenges, and best practices*. Verkregen maart 5, 2024, van https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

IBM. (2024). *What is text-mining*. Verkregen augustus 22, 2024, van https://www.ibm.com/topics/text-mining

Javija, R. (2024, juli 16). *Difference between Information Retrieval and Information Extraction*. Verkregen augustus 22, 2024, van https://www.geeksforgeeks.org/difference-between-information-retrieval-and-information-extraction/

Kaur, A. (2024, juli 25). *ACID Properties in DBMS*. Verkregen augustus 22, 2024, van https://www.geeksforgeeks.org/acid-properties-in-dbms/

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. *Big Data Cogn. Comput.*, *7*(2), 97. https://doi.org/10.3390/bdcc7020097

Kinter, P. (2024, februari 12). *Text mining: applications and techniques*. Verkregen augustus 22, 2024, van https://www.alexanderthamm.com/en/blog/text-mining-basics-methods-and-application-cases/

Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., & Valencia, A. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews*, *117*(12), 7673–7761. https://doi.org/10.1021/acs.chemrev.6b00851

Martinez, A. R. (2012). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*. https://doi.org/https://doi.org/10.1002/wics.195

Meta. (2024, juli 23). *Introducing Llama 3.1: Our most capable models to date*. Verkregen augustus 22, 2024, van https://ai.meta.com/blog/meta-llama-3-1/

Nagarjoon, B. (2022). What are regular expressions, and why should you use them? *Medium*. Verkregen augustus 22, 2024, van https://medium.com/@nagarjoon.b/what-are-regular-expressions-and-why-should-you-use-them-26140fe52bbe

openai. (2024). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*. Verkregen augustus 22, 2024, van https://openai.com/index/gpt-4/

Ray, I. (2024). *Validation of Large Language Models (LLMs)*. Verkregen augustus 22, 2024, van https://rayislam.medium.com/validation-of-large-language-models-llms-d934e1373d78

Saadani, T. (2024). Exploring NLP Techniques: Tokenization, POS Tagging, and NER. *Medium*. Verkregen augustus 22, 2024, van https://medium.com/ubiai-nlp/exploring-nlp-techniques-tokenization-pos-tagging-and-ner-bc9ead3f0843.

SaturnCLoud. (2024). *Stemming in Natural Language Processing*. Verkregen augustus 22, 2024, van https://saturncloud.io/glossary/stemming/

Scispace. (2024). *How much data is needed for fine-tuning a llm?* Verkregen augustus 22, 2024, van https://typeset.io/questions/how-much-data-is-needed-for-fine-tuning-a-llm-8liu5om85s

Securities, U., & Commission, E. (2023). *Frequently Asked Questions About Form 13F*. Verkregen augustus 22, 2024, van https://www.sec.gov/divisions/investment/13ffaq

Spacy. (2024). *Facts & Figures*. Verkregen augustus 22, 2024, van https://spacy.io/usage/facts-figures

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*, *7*(11). https://doi.org/https://doi.org/10.14569/IJACSA.2016.071153

Team, C. (g.d.). *SEC Filings*. Verkregen augustus 21, 2024, van https://corporatefinanceinstitute.com/resources/valuation/sec-filings/

Unsloth. (2024). unsloth homepage. Verkregen augustus 22, 2024, van https://unsloth.ai/

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