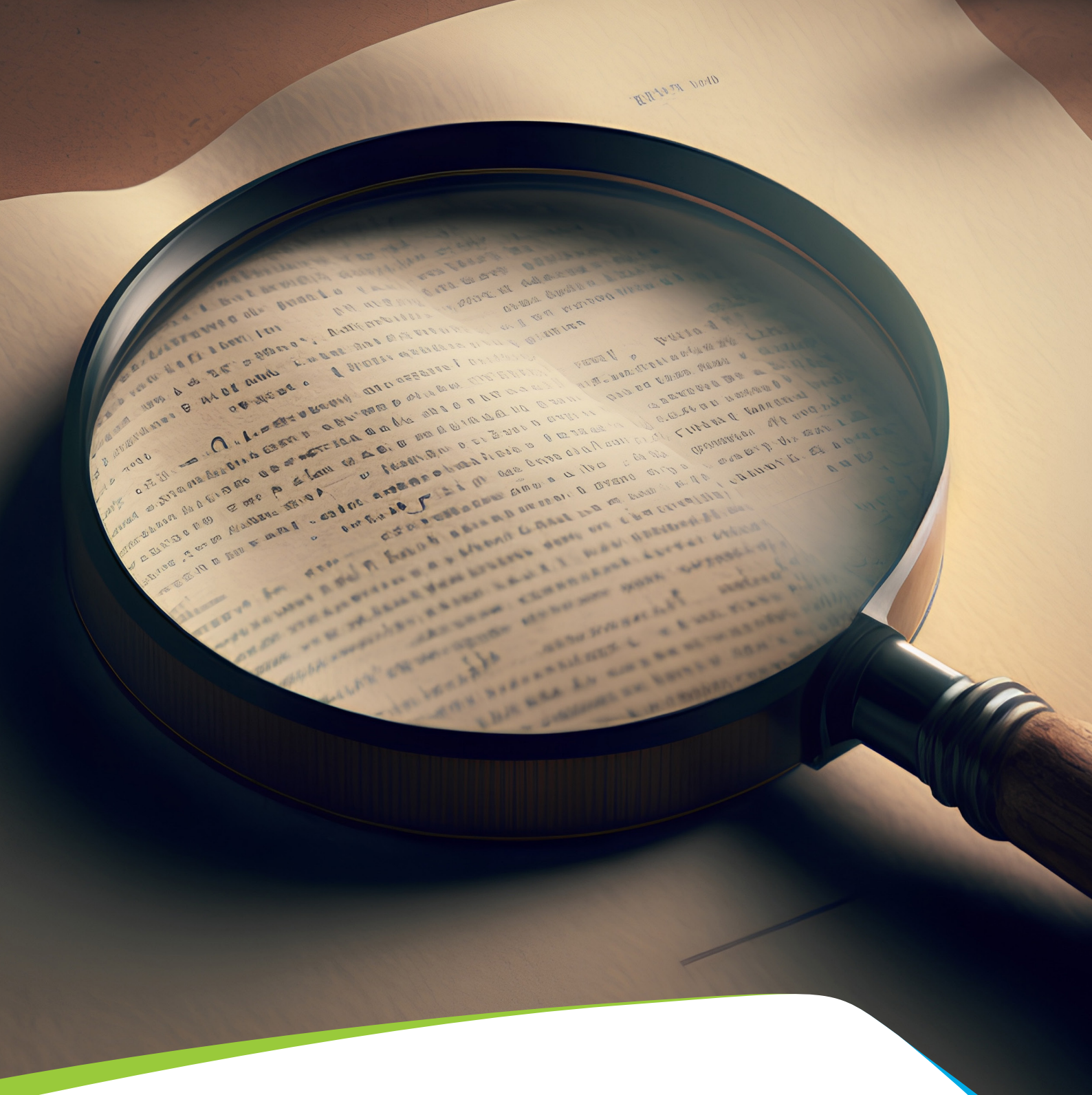

Tekstanalyse: gewoon doen!

Instituut van
Internal Auditors
Nederland

OPDRACHTGEVER
IIA Nederland

AUTEURS
Van Vugt, H.C., Van Duinen, W. en Van Uden, L.

Instituut van
Internal Auditors
Nederland

Inhoudsopgave

Inleiding

Beleidsdocumenten, overeenkomsten en contracten, rapporten, werkinstructies, e-mails, notulen, klachten, omschrijvingsvelden, krantenartikelen, Twitter en websites: ze bevatten allemaal teksten die worden gebruikt in de dagelijkse praktijk van organisaties. En dus ook binnen de auditfunctie. De hoeveelheid teksten die organisaties produceren en verzamelen is gigantisch en neemt elke dag toe; men kan zelfs spreken van een explosie in datavolume. Het lezen en gebruiken van deze teksten in onderzoeken is vaak tijdrovend en het objectief beoordelen is lastig. Deze teksten met behulp van een computer analyseren (tekstanalyse) kan een uitkomst bieden voor deze beperkingen van traditionele technieken. IIA onderschrijft dat tekstanalyse kansrijk is en verwacht dat het in snel tempo in belang zal toenemen in de auditpraktijk; voor IIA de aanleiding deze handreiking te publiceren. De auteurs van deze handreiking hebben in de eigen auditpraktijk ervaren dat tekstanalyse veel potentieel biedt om audits en onderzoeken efficiënter en effectiever te maken.

Doel van deze handreiking

Met deze handreiking willen we het auditors gemakkelijker maken om te starten met het toepassen van tekstanalyse; zowel in de planningsfase, bij het afbakenen, als tijdens de audit: om bewijs te verzamelen, om relevante documenten te filteren, om de gehele populatie te onderzoeken, et cetera. We starten met voorbeelden en de impact die onze tekstanalyses in onze eigen auditcontext hebben gehad. Vervolgens gaan we in op vier uitdagingen om tekstanalyse tot een succes maken. Begin 2022 hebben wij een enquête uitgezet onder auditors als opmaat naar deze handreiking. Hieruit bleek dat veel auditors weliswaar het nut en de potentie van tekstanalyse inzien, maar ook uitdagingen ervaren. Daarom dragen wij oplossingsrichtingen aan om allerhande uitdagingen te lijf te gaan. Met andere woorden: Wat is nodig voor het welslagen van tekstanalyse-projecten?

Belang: Voorbeelden, impact, en concrete analyses

E-mails die automatisch in je spamfolder gezet worden. E-mails van klanten aan een bedrijf die automatisch worden toegedeeld aan de juiste afdeling. Klantverslagen waar automatisch ongewenste, persoonlijke data gemaskeerd wordt zodat dit niet opgeslagen wordt in systemen van een organisatie. En, nog een stap verder: de ChatGPT bot die een tiener ‘helpt’ met huiswerk maken door relevante teksten te produceren (OpenAI, 2023). Of, een model dat op basis van de inhoud voorspelt hoe vaak een artikel, bijvoorbeeld op LinkedIn, wordt gelezen en geliket; en de inhoud zelfs aanpast om de kwaliteit te verhogen. Allemaal voorbeelden van tekstanalyse die we in de praktijk tegenkomen. Tekstanalyse kan worden omschreven als een verzameling technieken die leidt tot het extraheren van kennis en bruikbare inzichten uit tekstuele gegevens.

Ook uit de auditpraktijk zijn er allerhande voorbeelden van vraagstukken waarvoor tekstanalyse kan worden toegepast. Wordt een verplicht omschrijvings- of toelichtingsveld altijd ingevuld? Wordt duurzaamheid besproken in gesprekken met klanten? Welke personen zijn vaak aan- dan wel afwezig bij vergaderingen (blijkend uit de notulen) en wie zijn daarom logische sleutelpersonen om te interviewen in de audit? Welke transacties bevatten namen of landcodes behorende bij sanctielanden? Welke documenten zijn relevant om te lezen voor een onderzoek? De analyses om te komen tot antwoorden op deze vragen zijn relatief simpel in uitvoering en ook door auditors met beperkte kennis van tekstanalyse uit te voeren.

Natuurlijk zijn er ook voorbeelden waar complexere analyses voor nodig zijn. Worden wetten goed vertaald in intern beleid? In welke klachten wordt gesproken over hoge bedragen, en vormen daarom mogelijk een hoog risico? Is de door een maskeersysteem gebruikte woordenlijst compleet, en kunnen we deze aanvullen met synoniemen en andere gerelateerd woorden? Zijn kredietrapporten door de 1e lijn juist of onjuist geclassificeerd in een risicocategorie? Wat is de kwaliteit van Know Your Customer (KYC-)rapportages? Welke documenten uit een gigantische database zijn gemist maar zijn wel cruciaal in een grootschalig onderzoek? Zie hiervoor de casus in Box 1. Voor dit soort analyses is de inzet van een data scientist meestal nog onvermijdelijk, omdat de analyses complexer zijn en kennis van machine learning nodig kan zijn. Het is voor auditors echter goed zich te realiseren dat dit soort analyses mogelijk zijn, en dat de impact groter is dan wellicht gedacht.

Leeswijzer

Met het oog op hanteerbaarheid hebben we de uitdagingen geclusterd:

- **Uitdaging 1: Tekstuele data**: Hoe om te gaan met persoonsgegevens en gevoelige gegevens? Welke stappen zijn er nodig om een tekstanalyse uit te voeren? Wat zijn veel gebruikte technieken?
- **Uitdaging 2: Gereedschap**: Welke tools zijn er voorhanden? Waar kan relevante kennis verkregen worden om de tools te leren gebruiken? Focus is op de tooling die wij al succesvol toepassen in ons werk.
- **Uitdaging 3: Organisatie**: Wat is in een auditafdeling nodig om tekstanalyse tot een succes te maken?
- **Uitdaging 4: Succesvol aan de slag**: Welke additionele tips hebben wij om de kans op het succesvol toepassen van tekstanalyse in de auditpraktijk te vergroten? De tips gaan met name over uitdagingen rondom de projectaanpak.

Data-analyse begint zijn plek te vinden in de controlemethodiek van de accountant (Gold & Majoor, 2022; Liew, Boxall, & Setiawan, 2022; NBA, 2019) en ook steeds meer in de auditfunctie (Johnson, Wiley, Moronay, Campbell, & Hamilton, 2021; Gartner, 2020; Wang & Cuthbertson, 2015; Eilifsen, Kinserdal, Messier, & McKee, 2020; Deloitte, 2013). Wij hopen met deze handreiking eraan bij te dragen dat ook tekstanalyse niet meer weg te denken is uit de auditpraktijk. Ons motto is daarom: ***Tekstanalyse: gewoon doen*!**

Impact van tekstanalyse

Auditors die een vorm van tekstanalyse hebben ingezet in hun werkzaamheden zijn veelal enthousiast. Ze hebben gemerkt dat hun werk effectiever en efficiënter wordt. Effectiever door het zicht op de gehele populatie (alle documenten bekijken i.p.v. sampling); door het verkrijgen van meer en andere inzichten leidend tot verhoogde kwaliteit van assurance; door inzicht in de werking van beheersmaatregelen, door de selectie van de juiste sleutelpersonen voor de audit. Efficiënter door besparing in leestijd en/of door de afname in te analyseren documenten, wat weer kan leiden tot meer focus en een verkleinde of duidelijkere scope; door het makkelijker periodiek toepassen van dezelfde analyse. Kort gezegd, de beschikbare teksten worden benut ten gunste van de audit.

Een bijkomend voordeel is dat de resultaten na analyse vaak direct toonbaar zijn aan de auditee. Dit maakt op feiten gebaseerde besprekingen eenvoudiger. Net als andere vormen van data-analyse zal tekstanalyse ervoor zorgen dat er eerder actie wordt ondernomen naar aanleiding van bevindingen (Gartner, 2017), omdat ze datagedreven en inzichtelijk zijn. Vaak zijn analyses ook relevant voor de 2e lijn om controlerende werkzaamheden uit te voeren. Soms zijn inzichten zo bruikbaar, relevant en/of verrassend voor de auditee en/of de 2e lijn dat eigen analisten verder gaan bouwen op de door audit uitgevoerde tekstanalyses en deze zelfs gaan standaardiseren. Audit kan vervolgens daarop voortbouwen in de toekomst en maakt hierdoor een omslag van gegevensgericht auditen met tekstanalyse naar systeemgericht auditen van tekstanalyse.

Concrete analyses

De legio voorbeelden van tekstanalyses toegepast in auditafdelingen zijn veelal relatief simpel in uitvoering. Deze zijn ook door auditors met beperkte kennis van data-analyse eenvoudig toe te passen. Natuurlijk zijn er ook meer complexe analyses, waar meer kennis voor nodig is. Denk aan voorspelmodellen op basis van tekstkenmerken. Voor de meer complexe analyses is de inzet van een data analist/scientist meestal nog onvermijdelijk; laat deze voorlopig nog maar aan data scientists over.

De onderstaande analysevoorbeelden zijn (grofweg) geordend van relatief simpel naar meer complex. We starten steeds met de vraag, en gaan dan kort in op de analyse. Bij elke analyse geldt dat de auditor, met zijn ‘auditor-bril’ op, de uitkomsten inspecteert, interpreteert, en beoordeelt, d.w.z., nadenkt over wat de analyse betekent in de context van de audit, en welke vervolgvragen het oproept die mogelijk weer met vervolganalyses beantwoord kunnen worden.

LENGTE VAN TEKSTEN

- Wordt een (verplicht) omschrijvings- of toelichtingsvelden altijd ingevuld? Zo ja, worden er meer dan drie woorden gebruikt? Een inhoud van minder dan 4 woorden is vaak niet betekenisvol, denk aan 'zie dossier', 'niet van toepassing', '-', of '..'.

Analyse: Tellen van niet gevulde velden en het tellen van het aantal woorden in een veld. Sorteren op frequentie.

ZOEKEN NAAR WOORDEN

- Welke documenten zijn relevant om te lezen voor een onderzoek?

Analyse: Maken van een woordenlijst met interessante woorden voor het onderzoek (bv. 'Covid' en 'Corona'). Tellen van deze woorden in de documenten. Sorteren op frequentie om vervolgens de meest relevante, interessante documenten te selecteren.

- Wordt duurzaamheid besproken in gesprekken met klanten?

Analyse: Maken van een woordenlijst met woorden gerelateerd aan duurzaamheid. Tellen van deze woorden in de gespreksverslagen. Tellingen doen per woord en per verslag. Sorteren op frequentie om snel te zien in welke gesprekken er niet of juist veel is gesproken over duurzaamheid.

- Wordt persoonlijke informatie goed gemaskeerd door het maskeersysteem? Is de door het maskeersysteem gebruikte woordenlijst compleet?

Analyse: Synoniemen zoeken van woorden uit de bestaande woordenlijst, gebruik makend van een bestaand synoniemenwoordenboek van de Nederlandse taal. Zoeken naar alle woorden uit de bestaande woordenlijst en naar de relevante synoniemen in de documenten na verwerking door het maskeersysteem. Sorteren op frequentie om te zien welke woorden vaak door het systeem 'glippen'.

- Welke transacties bevatten namen of landcodes behorende bij sanctielanden (en zijn mogelijk risicovol)?

Analyse: Landenlijst met landcodes gebruiken als woordenlijst. Zoeken in transactievelden (van, naar, opmerkingen) om transacties op te sporen die mogelijk gelinkt zijn aan sanctielanden.

ZOEKEN NAAR PATRONEN

- Wie zijn sleutelpersonen in een bepaald traject of project? Dit helpt bij een juiste selectie van personen voor interviews, en voor reconstructie van een proces.
  - Wie hebben e-mails verstuurd en ontvangen (in een bepaalde context)?

Analyse: Zoeken naar de patronen 'van:', 'cc:' 'bcc:' in e-mails (.msg bestanden).

  - Veranderde dat over de tijd?

Analyse: Zoeken naar 'datum:', en naar een patroon als '1 of 2 cijfers, spatie, 1 woord, spatie, 4 cijfers' om de datums uit tekst te halen.

  - Wie waren aanwezig en afwezig in vergaderingen van het afgelopen jaar?

Analyse: Zoeken naar patronen als 'aanwezig:' en 'afwezig:' in notulen, en vervolgens de namen die erachter staan filteren en tellen.

- In welke klachten wordt gesproken over hoge bedragen, en vormen daarom mogelijk een hoog risico?

Analyse: zoeken naar patronen als '4 of meer cijfers, spatie, euro'.

VERGELIJKEN VAN TEKSTEN

- Worden teksten steeds gekopieerd en geplakt in opmerkingenvelden? Dit kan een indicatie zijn voor de kwaliteit van het werk van de eerste lijn.

Analyse: Groeperen van tekstvelden en tellen hoe vaak elk tekstveld voorkomt. Groeperen a.d.h.v. 100% gelijke teksten, of a.d.h.v. 'near-matches' (bijvoorbeeld 90% gelijke teksten).

- Worden wetten goed vertaald in intern beleid? Dit is bijvoorbeeld relevant om te zien of nieuwe of gewijzigde wetgeving goed door vertaald wordt in het interne beleid.

Analyse: Tellen van alle woorden, of beter: lemma's, in zowel wetten als beleidsdocumenten. Vergelijken van de woorden/lemma's uit elke wet met woorden/lemma's uit elk beleidsdocument. Indien woorden/lemma's veel voorkomen in een wet maar niet in het beleidsdocument, kunnen er vraagtekens bij gezet worden.

SENTIMENT-ANALYSE

- Komt de risicocategorie waarin een klant is ingedeeld overeen met het sentiment in de gespreksverslagen?

Analyse: Zoeken naar en tellen van positieve woorden en woordcombinaties (bv. 'goed, positief', 'bekwaam', 'verbeterd', 'niet slecht', 'geen aandachtspunten') en negatieve woorden en woordcombinaties (bv. 'slecht', 'verslechterd', 'probleem', 'onvoldoende', 'niet goed', 'weinig verbetering'). Selecteren van een relevante sample, bijvoorbeeld die klanten ingedeeld in de laag-risicocategorie en veel negatieve woorden in de tekst; Dit zijn mogelijk klanten die in een hogere risicocategorie thuishoren gezien het aantal negatieve woorden in de tekst.

NETWERKANALYSE

- Wie is met wie in contact in de organisatie, en in welke mate?

Analyse: In kaart brengen van netwerken door emailanalyse: wie mailt naar wie en hoe vaak?

CLASSIFICEREN VAN TEKSTEN

- Welke teksten zijn juist of onjuist geclassificeerd door de eerste lijn? Zijn teksten met gelijksoortige inhoud op een zelfde manier geclassificeerd?

Analyse: zoeken naar onderwerpen in de teksten (onderwerpmodellering). De tekst of documenten met behulp van algoritmes classificeren of categoriseren aan de hand van de onderwerpen.

- Welke documenten uit een gigantische database zijn gemist maar zijn wel cruciaal in een grootschalig onderzoek?

Analyse: Op basis van publieke bronnen, bijvoorbeeld: krantendatabase, middels tekstanalyse relevante zoektermen generen en daarmee zoeken naar relevante documenten in een database.

In navolgende casus (Auditdienst Rijk, 2021) wordt een voorbeeld gegeven van de toepassing van tekstanalyse in een onderzoek.

CASUS

Tekstanalyse voor een parlementaire enquête Gaswinning Groningen

Bij een parlementaire enquête vraagt een parlementaire commissie via vorderingsvragen dossiers uit bij de opdrachtgever. De opdrachtgever wil graag een volledige en transparante beantwoording geven en vraagt daarom de Auditdienst Rijk (ADR) naar de bevindingen bij de totstandkoming daarvan.

Een auditor kan kijken naar de getroffen beheersmaatregelen en of deze gedurende het zoekproces goed gewerkt hebben. Maar met die aanpak kan de auditor niet vaststellen of het resultaat van het zoekproces volledig is. En indien de resultaten niet volledig waren, welke documenten er ontbreken.

Door tekstanalyse toe te passen kan op een objectieve en onafhankelijke manier worden vastgesteld of de gehanteerde aanpak volledig is en welke documenten mogelijk gemist zijn. De tekstanalyse geeft antwoord op onder andere de volgende onderzoeksvragen:

1. Zijn de opgestelde zoekvragen toereikend om relevante documenten te vinden?
2. Zijn er documenten gemist tijdens de beantwoording van de vorderingsvragen?

Om een antwoord te geven op deze twee onderzoeksvragen is tekstanalysetechniek “onderwerpmodellering” toegepast. Voor het opstellen van de set van zoekvragen is gebruik gemaakt van publieke bronnen, namelijk kamerstukken en krantenartikelen. Met onderwerp modellering zijn de hoofdonderwerpen vastgesteld voor alle relevante kamerstukken. In een krantendatabase is vervolgens op deze hoofdonderwerpen gezocht om tot een verrijkte set van zoekvragen te komen.

Vervolgens is de verrijkte set van zoekvragen gebruikt om in een documentenmanagementsysteem naar relevante documenten te zoeken. Dit leverde grofweg twee miljoen documenten op, die vervolgens zijn ingedeeld in verschillende onderwerpen. Er zijn tien onderwerpen geclassificeerd als meest relevant. Voor elk van deze onderwerpen zijn documenten gestratificeerd geselecteerd die door het team van de opdrachtgever gevonden zijn en documenten die niet door het team gevonden zijn. Op basis van de massa ‘niet gevonden door het team’ is iets te zeggen over de volledigheid van de zoekvragen (onderzoeksvraag 1) en welke documenten er mogelijk zijn gemist bij de gehanteerde zoekmethode (onderzoeksvraag 2).

Box 1. Casus toepassing van tekstanalyse voor de parlementaire enquête Gaswinning Groningen.


var b,d=this,e=this
undelegateEvent
reviewDeviceBut
maybeRequestFile
announceS
function(){c.overl
render:

Uitdaging 1: Tekstuele data

Uitdagingen rondom tekstuele data kunnen we clusteren in drie groepen:

I. Verzamelen van persoonsgegevens en gevoelige gegevens
II. Verwerken van ongestructureerde data
III. Analyseren met technieken voor tekstanalyse

Elk van deze groepen brengt zijn eigen uitdagingen met zich mee.

I. Verzamelen van persoonsgegevens en gevoelige gegevens

Tekstuele data komen ofwel uit de organisatie zelf (denk aan e-mails, interne documenten en evaluaties) ofwel van buiten de organisatie (denk aan social media, kranten en websites). Er kunnen andere regels gelden voor interne en externe bronnen.
Voor interne bronnen is het belangrijk te bekijken op welke manier en met welk doel de data verkregen is. Vaak komen in tekstuele data persoonsgegevens voor die herleidbaar zijn tot een individu, zoals namen, paspoortnummers en adressen. Ook kunnen gevoelige gegevens of gegevens van persoonlijke aard voorkomen, zoals gegevens die informatie bevatten over religieuze of politieke overtuigingen, etnische afkomst, gezondheid, genetische gegevens, en vakbond-lidmaatschappen. Sinds 2018 geldt in de hele EU dezelfde privacywetgeving: de Algemene Verordening Gegevensbescherming (AVG). De Autoriteit Persoonsgegevens schrijft hierover: "*Elke keer als u persoonsgegevens verwerkt, is dat een inbreuk op de privacy van de mensen over wie het gaat. Daarom mag u alleen persoonsgegevens verwerken als het echt niet anders kan. Dus: als u zonder deze gegevens uw doel niet kunt bereiken*." (Autoriteit Persoonsgegevens, 2022). Bepaal van tevoren dus goed of er persoonsgegevens en/of gevoelige data aanwezig is en hoe daarmee om te gaan. Een aantal beginselen kunnen in overweging worden genomen voor het verwerken van de data (Schermer, Hagenauw, & Falot, 2018):

- **Noodzaak**: Beperk het gebruik van gegevens tot wat noodzakelijk is om een onderzoek uit te voeren (data-minimalisatie). De verwerking moet gebonden zijn aan specifieke doelen.
- **Relevantie**: Verwerk alleen de relevante persoonlijke data. Vaak kan een analyse ook uitgevoerd als persoonlijke gevoelige gegevens uit datasets verwijderd worden (anonimiseren), vervangen door gerandomiseerde codes (pseudonimiseren), of op zo'n manier geaggregeerd die zinvol is voor de controle. Bewaar de data niet langer dan nodig.
- **Veiligheid**: Gegevens moeten goed beveiligd zijn. Bijvoorbeeld, plaats het niet op een gedeelde schijf waar onbevoegden zomaar bij kunnen, of analyseer de data niet in een treinreis.
- **Vertrouwelijkheid**: Gegevens moeten vertrouwelijk blijven. Deel de persoonlijke data niet met anderen zonder geldige reden.

Net als bij interne bronnen, is doelbinding ook belangrijk bij het gebruik van data uit externe bronnen. Dat informatie vrijelijk op internet te vinden is, maakt het nog niet automatisch gerechtvaardigd dit voor allerlei doeleinden te gebruiken. Er zijn verschillende rechtsgebieden, zoals het databankenrecht, het auteursrecht, het recht op eerbiediging van de persoonlijke levenssfeer (privacy) en het recht op bescherming van persoonsgegevens, die beperkingen of zelfs verboden kunnen stellen aan het verzamelen van externe (tekstuele) data voor eigen gebruik (Wiseman Advocaten, 2020). Ook kunnen er specifieke voorwaarden bestaan voor verwerking van data van een externe partij. Het is belangrijk om hiernaar te kijken alvorens enthousiast aan de slag te gaan met, bijvoorbeeld, het scrapen (het afhalen van informatie van websites) van nieuwsberichten of reviews.

Steeds meer organisaties hebben een data protection officer, ook wel privacy officer of functionaris gegevensbescherming genoemd, in dienst die meer weet over het interne beleid om te voldoen aan (privacy)wetgeving. Ga dus in gesprek met de privacy officer van je organisatie om te achterhalen of en hoe bepaalde data mag worden verwerkt in het licht van het doel dat wordt nagestreefd.

II. Verwerken van ongestructureerde data

Kenmerkend is dat tekstuele gegevens ongestructureerd zijn; de data heeft (vrijwel) geen vast formaat, is niet gelabeld en zinnen en woorden zijn niet gestructureerd in rijen en kolommen. Het is natuurlijke taal, geen computertaal. Ongestructureerde data kan afkomstig zijn van allerlei typen bestanden, denk aan *.doc*, *.txt*, *.pdf*, *.msg*, *.ppt* en *.xls*. In deze bronnen is de tekst vaak al op zó'n, gedigitaliseerde, manier aanwezig dat het direct kan worden ingelezen in de tekstanalyse software. Soms hebben we te maken met spraak, bijvoorbeeld in het geval van interviews, of met handgeschreven tekst in gescande documenten. In beide gevallen is eerst voorbewerking nodig om de input in het juiste, digitale, formaat te krijgen. Spraak kan door speciale software automatisch worden omgezet in gedigitaliseerde tekst (Turner, 2022). Handgeschreven tekst kan door speciale software met behulp van de techniek Optical Character Recognition (OCR) worden ingelezen; Letters in afbeeldingen worden herkend en vervolgens wordt de handgeschreven tekst omgezet naar gedigitaliseerde tekst.

Om tekstanalyse toe te passen, wordt structuur aangebracht in de ongestructureerde data: de tekstuele gegevens worden omgezet naar gestructureerde gegevens in rijen en kolommen. Een voorbeeld van een gestructureerde vorm van tekst is een matrix met in de rijen zinnen, en in de kolommen een item zoals een woord; van elk woord wordt aangegeven of het wel of niet aanwezig is in de zin (één of geen; zie Tabel 1). Om hiertoe te komen worden één of meerdere verwerkingsstappen toegepast. Het is belangrijk om goed na te denken over de (ir)relevantie van elke stap voor het doel van de analyse. Immers geldt over het algemeen: hoe beter de kwaliteit van de data (en, de kwaliteit van de datapreparatie), des te beter de resultaten van de analyse.

	Er	is	een	betalingsregeling	De	aanvrager	verzoekt	om	krediet
Er is een betalingsregeling	1	1	1	1	0	0	0	0	0
De aanvrager verzoekt om een krediet	0	0	1	0	1	1	1	1	1

Tabel 1. Voorbeeldmatrix

Verwerkingsstappen kunnen zijn:

VERWIJDEREN VAN IRRELEVANTE WOORDEN

Om een vereenvoudigde, kortere, weergave van de tekst te verkrijgen, worden leestekens zoals punten, komma's, en vraagtekens, vaak verwijderd uit de tekst. Ook worden stopwoorden vaak verwijderd. Dit levert een vereenvoudigde en kleine matrix op (zie Tabel 2). Onder stopwoorden verstaat men veelgebruikte woorden zonder op zichzelf staande betekenis, zoals lidwoorden (de, het, een) en bijwoorden (die, dat, aan), en woorden die opmerkelijk vaak gebruikt worden zonder uitdrukkelijke betekenis, zoals 'zeg maar'. Er bestaan stopwoordenlijsten voor verschillende talen en het is ook mogelijk een eigen stopwoordenlijst te maken. Ook worden hoofdletters vaak vervangen door kleine letters, zodat de computer 'Vandaag' en 'vandaag' als identieke woorden beschouwt. Het is natuurlijk altijd raadzaam om te bedenken of voor jouw specifieke analyse bepaalde woorden inderdaad verwijderd en hoofdletters vervangen kunnen worden, of dat het beter is de woorden te houden, bijvoorbeeld, als je geïnteresseerd bent in het precieze aantal gebruikte woorden.

	betalingsregeling	aanvrager	verzoekt	krediet
betalingsregeling	1	0	0	0
aanvrager verzoekt krediet	0	1	1	1

Tabel 2. Voorbeeldmatrix waarbij irrelevante woorden verwijderd zijn

OPKNIPPEN VAN DE TEKST IN TOKENS

"Tokenization" is het opknippen van de tekst in zinnen, en zinnen in losse woorden. Er bestaan verschillende tokenizers die elk leiden tot een iets ander resultaat. Je geeft in de tool aan welke tokenizer je wilt gebruiken, afhankelijk van de taal waarin de tekst die je gaat analyseren geschreven is. Een Engels woord-tokenizer splitst het Engelse 'she's' op in de tokens 'she' en 'is'. In een matrix staan de woorden 'she' en 'is' dan in aparte kolommen. Een Whitespace-tokenizer zal een tekst opknippen in stukjes die met een spatie beginnen en eindigen: 'she's' wordt dan niet opgesplitst maar gezien als één token. In een matrix staat 'she's' dan in één kolom. Het Nederlandse 'Anna's fiets' wordt door de Whitespace-tokenizer opgesplitst in de twee tokens 'Anna's' en 'fiets'. Een Nederlandse tokenizer weet dat ''s' gezien kan worden als apart derde token (dat bezit aangeeft). Niet elke tool heeft alle mogelijke tokenizers geïmplementeerd. Gebruik voor de Nederlandse taal dan een Engelse, Duitse of Whitespace Tokenizer. Deze geven vrij goede resultaten.

GROEPEREN VAN TOKENS MET EEN GELIJKENDE BETEKENIS

Een token als 'fiets', 'fietsen' en 'gefietst' zijn gerelateerd aan elkaar. De computer ziet deze tokens pas als gerelateerde woorden als we ervoor zorgen dat de tokens teruggebracht worden tot hun 'stam' of 'lemma'.

De stam is, simpel gezegd, het deel van een woord dat overblijft wanneer men de buigingsuitgangen weghaalt. Bijvoorbeeld, enkelvouden en meervouden van een woord (fiets/fietsen) hebben dezelfde stam (fiets). Een lemma is iets verfijnder, want het houdt rekening met de betekenis. Bijvoorbeeld, het woord 'beter' heeft 'goed' als lemma (en 'bet' als stam). Om dit te bepalen zijn woordenboeken nodig om een woord te koppelen aan zijn lemma.

De stammen of lemma's worden vervolgens gegroepeerd zodat ze als één item kunnen worden geanalyseerd. In een matrix staan alle stammen of lemma's dan in één kolom. Over het algemeen levert lemmatization betere groeperingen op dan stemming, maar het is moeilijker te implementeren omdat er meer kennis van de taal nodig is (en niet elk woordenboek is even goed).

III. Analyseren met technieken voor tekstanalyse

Veel organisaties worstelen in de beginfase dan ook met het kiezen van de juiste aanpak om de gewenste uitkomsten uit een analyse te krijgen (Expert.AI, 2023). Bij het uitvoeren van een tekstanalyse is het belangrijk dat de juiste techniek wordt gekozen om de doelen van een analyse te bereiken. Probeer dus van te voren het analysedoel goed scherp te hebben, en kies dan een techniek die daarbij past. Tekstextractie en tekstclassificatie, containerbegrippen voor een waaier aan technieken, worden veel gebruikt:

TEKSTEXTRACTIE

Met tekstextractie wordt een bepaalde term of stuk tekst uit de data gehaald. Dit kunnen bepaalde kernwoorden zijn maar ook bijvoorbeeld e-mailadressen, personen, of zelfstandig naamwoorden. Bekende technieken die hierbij worden ingezet zijn 'bag of words', term frequency, regex, named entity recognition, en part of speech tagging. Een 'bag of words' omvat alle individuele woorden uit een tekst. Hierin kan vervolgens gezocht worden naar specifieke woorden of woorden uit een woordenlijst(je). Term frequency (TF) geeft de frequentie aan van een woord in een tekst. Met regex kan gezocht worden naar een stukje tekst die een bepaalde structuur volgt, zoals een e-mailadres (een tekst dat een '@' bevat) of rekeningnummer. Named entity recognition kan worden ingezet om entiteiten zoals personen, organisaties en locaties in een tekst te herkennen en te extraheren. Part of speech tagging kan worden gebruikt om woorden aan te duiden als zelfstandig naamwoord, werkwoord, et cetera.

TEKSTCLASSIFICATIE

Soms zijn teksten in een database geclassificeerd in groepen of gelinkt aan thema's. Bijvoorbeeld, kredietrapporten worden in de bank door de 1e lijn geclassificeerd in een risicocategorie. Een auditvraag kan zijn: is dit classificeren juist? Door de teksten te onderzoeken, kan een relatie ontdekt worden tussen de teksten en de classificatie en worden beoordeeld of de classificatie juist was.

Soms zijn teksten nog niet geclassificeerd en kan het nuttig zijn om teksten automatisch te laten categoriseren. In welke groep of thema een niet geclassificeerde tekst het beste past, wordt bepaald door de (on)gelijkheid tussen deze tekst en bestaande teksten in de verschillende groepen. Als gelijkheid met een groep groot genoeg is, kan de tekst automatisch worden geclassificeerd in die groep. Op deze manier worden bijvoorbeeld e-mails in een klantenservice automatisch doorverwezen aan de juiste afdeling.

Onderwerpmodellering (topic modeling), een machine learning techniek, kan ingezet worden door de meer ervaren data-analist om gemeenschappelijke onderwerpen te zoeken in de documenten. Deze techniek werd toegepast door Auditdienst Rijk in de casus in Box 1. Ook kunnen teksten worden geclassificeerd in termen van sentiment: is de emotionele lading positief of negatief? Sentimentanalyse wordt bijvoorbeeld gebruikt om reviews te analyseren, en kan door audit worden gebruikt om klachten en klachtenmanagement te onderzoeken.


esc
F1
Tab
Q
@
A
Shif
Fn

Uitdaging 2: Gereedschap

Uit onze enquête bleek dat ondanks de wil om tekstanalyse toe te passen in auditwerkzaamheden, de kennis vaak ontbreekt om het daadwerkelijk te doen. Sommige organisaties nemen data-analisten of data-scientists in dienst die zorgen voor de benodigde kennis van tekstanalyse en software. Maar steeds meer auditors willen zelf datagedreven werken en tekstanalyses uitvoeren. Er bestaan veel tekstanalyse tools en recente overzichten van gratis en commerciële oplossingen zijn te vinden op het internet (Pat research, 2022; Datamation, 2021; Mousa, 2020). Van sommige tools die hierin staan viel het ons op dat deze gebruikers niet zelf aan de knoppen draaien. In deze handreiking hebben we ervoor gekozen te focussen op de tools die wij zelf gebruiken, of gebruikt hebben, voor tekstanalyse en daardoor de meeste kennis van hebben.

I. Basis functionaliteiten in Microsoft Excel

Microsoft Excel heeft een aantal functies waarmee simpele tekstanalyses uitgevoerd kunnen worden, denk aan een functie als 'Find' om te checken of een woord in een tekst in een kolom voorkomt. De functies die Excel biedt zijn onderverdeeld in categorieën, en de categorieën 'Text' en 'Lookup & Reference' bevatten mogelijk interessante functies. Natuurlijk omvat Excel bij lange na niet de functies die de tools hebben die hieronder worden besproken. Wij zijn niet goed bekend met andere tooling die Microsoft heeft ontwikkeld. Denk hierbij aan Azure die mogelijkheden voor tekstanalyse biedt (Text Analytics | Microsoft Azure). Azure biedt waarschijnlijk meer mogelijkheden dan Excel.

II. No-code omgevingen: KNIME en Alteryx

Voor KNIME en Alteryx is geen programmeerkennis vereist, het zijn een 'no-code' omgevingen. Ze hebben een visuele interface die men snel kan leren te gebruiken en men kan op deze manier goed kennismaken met de mogelijkheden en toepassingen van tekstanalyse. Met beide tools kunnen allerlei soorten simpele en complexere tekstanalyses worden uitgevoerd, zoals de voorbeelden genoemd in het eerste deel van deze handreiking.

KNIME kan gratis worden gedownload, inclusief een extensie waarmee tekstanalyses gedaan kunnen worden. Er bestaat een betaalde versie die extra functionaliteiten biedt (zie https://www.knime.com/knime-software/knime-hub-pricing). Dit gebruiken wij zelf niet. Aan Alteryx zjin meer kosten verbonden. Er kan een gratis trial versie worden gedownload van Alteryx Designer Free Trial | Alteryx.

KNIME biedt een leerplatform (KNIME Learning | KNIME) waar ondersteunend materiaal te vinden is om aan de slag te gaan met KNIME: boeken, cursussen (online, op locatie en in eigen tempo), technische documentatie, certificering en meer. From words to wisdom (https://www.knime.com/knimepress/from-words-to-wisdom) is een relevant boek toegespitst op tekstanalyse. Ook

bestaan er online video's op YouTube (bijvoorbeeld Text Mining Techniques - YouTube) en zijn er voorbeeld-analyses die kunnen worden gedownload op de KNIME-hub (https://hub.knime.com/). Alteryx biedt ook allerhande mogelijkheden (video's, trainingen en voorbeelden) om de tool te leren gebruiken op Alteryx Academy - Alteryx Community.

III. Programmeercode: Python of R

Python en R zijn programmeertalen. Om dit te gebruiken voor tekstanalyse is programmeerervaring nodig. Vaak kunnen data-scientists goed met Python of R overweg, maar is er voor auditors een te hoge drempel om dit te leren. Python en R kunnen worden gebruikt in allerhande applicaties, zoals Anaconda die gebruik maakt van zogenaamde Jupyter notebooks (Using Jupyter Notebook — Anaconda documentation). Voor tekstanalyse kan in Python gebruik worden gemaakt van de packages NLTK (Natural Language ToolKit; https://www.nltk.org/) en spaCy (https://spacy.io). Er bestaan allerhande cursussen om deze packages te leren gebruiken (bijvoorbeeld via https://www.datacamp.com/).


IN
LIFTING
OTHERS
WE
RISE

Uitdaging 3: Organisatie

Het structureel, maar ook incidenteel toepassen van tekst- en andere data-analyses in audits betekent waarschijnlijk een behoorlijke verandering in de manier van werken. Als auditafdelingen analytische inspanningen daadwerkelijk willen volhouden en in de praktijk willen zien, moeten ze verandering omarmen (Deloitte, 2013). En dat in een professie waar grote waarde wordt gehecht aan bestaande methodologieën, en een conservatieve houding en scepsis ten aanzien van data-analyse veel voorkomt (Li, 2022). Over het tot een succes maken van veranderingen als deze zijn legio boeken, cursussen en websites geschreven. De acht veranderstappen[1] van Kotter (Kotter, 1996) geven naar ons beeld een goed overzicht van wat nodig is. Hier zullen we deze stappen niet uitgebreid beschrijven. Wel zullen we – merendeels in lijn met het Kotter model - een aantal tips bespreken die door deelnemers van de IIA Professional Practice dag in 2022 genoemd zijn en die we ook zelf hebben ervaren als cruciaal:

I. Vraag expliciet om steun van management

Het geloof dat data-analyse in audit een belangrijke plek zal (blijven) innemen in de toekomst is de kern om het tot een succes te maken. Het is belangrijk dat management de verandering naar meer datagedreven auditen omarmt, om er zelf als auditor ook echt mee aan de slag te gaan en om zich een nieuwe, datagedreven aanpak eigen te maken. Immers, hoe voelt de auditor zich anders gesteund om er tijd en moeite in te investeren? De meeste auditors hebben geen data-achtergrond en hebben tijd nodig om kennis van (tekst)analyse op te bouwen, deze kennis toe te passen en ervan te leren. Als cursussen gevolgd moeten gaan worden of tools worden aangeschaft, kan het ook een investering in geld betekenen. Zonder de expliciete steun van management is dit lastig. Verder kan positieve waardering van management voor de meer innovatieve, datagedreven auditor helpen als stimulans om meer en meer datagedreven te gaan werken.

II. Maak uitdagingen bespreekbaar

Elke organisatie heeft te maken met uitdagingen om tekstanalyse tot een succes te maken. Het expliciteren en bespreken van zowel doelen als de uitdagingen om deze doelen te bereiken helpt om verder te komen (Van Vugt, 2022). Belangrijke vragen zijn: 1) Welk volwassenheidniveau heeft de organisatie of het organisatieonderdeel nu? Is tekstanalyse afwezig, beperkt aanwezig, is het

1 1. Voelbaar, zichtbaar maken van de noodzaak ('sense of urgency': 'gevoel van urgentie')
2. Instellen van een krachtige stuurgroep met voldoende middelen om de noodzakelijke verandering te leiden
3. Ontwikkelen van een richtinggevende visie annex strategieën om die visie te realiseren
4. Communiceren van de nieuwe visie
5. Stimuleren en mogelijk maken conform de nieuwe visie te handelen
6. Zorgen voor zichtbare kortetermijnsuccessen
7. Consolideren van verbeteringen en blijven doorvoeren van veranderingen
8. Veranderingen verankeren in de bedrijfscultuur

opkomend, of wordt tekstanalyse al op gestructureerde wijze ingezet? 2) Welk volwassenheidsniveau streeft de organisatie na? Wat wil en kan de organisatie bereiken door het inzetten van tekstanalyse? 3) Welke belemmeringen worden ervaren om tot het gewenste ambitieniveau te komen? Hoe kunnen belemmeringen worden weggenomen?

Uitdagingen, of belemmeringen, kunnen van allerlei aard zijn (Shahim, van Praat, Harmzen, & Matthijsse, 2018; Stuurman, 2022): een gebrek aan kennis, beperkte datakwaliteit en -bruikbaarheid, zorgen om datasecurity en privacy, het ontbreken van een eenduidig perspectief op de toepassing van data-analyse, een hoge inspanningsverwachting (bijvoorbeeld, om data te verkrijgen, het doorgronden van de techniek en het juist interpreteren van de uitkomsten), beperkte faciliterende omstandigheden (het gebrek aan tijd, kennisdeling en gereedschap) en de complexiteit van informatiesystemen. Met de juiste instelling blijken uitdagingen minder groot dan gedacht en kunnen ze plaats maken voor mogelijkheden.

Een hulpmiddel bij het bespreken van zulke uitdagingen is het Optimaal Digitaal Spel, dat is aangepast aan de auditcontext (Bruinenberg, 2022). Het kan worden gespeeld in de opstartfase van audits waar verschillende soorten uitdagingen spelen. Het op een gestructureerde en leuke manier bespreken van uitdagingen helpt het audit-team enorm. Betrokkenen komen op één lijn en vinden samen praktische oplossingen, waardoor vervolgens de productiviteit in de audit verhoogt.

III. Werk samen met 1e en 2e lijn

Data die relevant is in auditcontext wordt veelal geproduceerd in de 1e lijn. Vaak maakt de 1e lijn ook al dashboards op deze data, bijvoorbeeld door een BI-afdeling. 2e en 3e lijn hebben interesse in dezelfde data, maar zoeken vaak naar een andere kijk op deze data. Ze hebben eigen inzichten nodig om hun controlerende werkzaamheden te kunnen verrichten. Samenwerking met 2e lijn is belangrijk om voort te bouwen op daar uitgevoerde (tekst)analyses en te leren van elkaars aanpak en focus. Samenwerking met 1e lijn is belangrijk om te weten welke data beschikbaar is, wat de data betekent, hoe de data geproduceerd wordt, hoe het gebruikt wordt in processen en dashboards en wat analyses zijn die op de wensenlijst staan maar waar men nog niet aan toe is gekomen. Een goede verhouding met business owners is ook van belang, omdat dit het verkrijgen van toestemming om data te gebruiken in analyses, natuurlijk binnen de kaders van het beleid, makkelijker en sneller maakt.

IV. Krijg de IT-afdeling mee

Het kunnen gebruiken van analyse-software is vaak afhankelijk van het beleid van een IT-afdeling in de organisatie. Vaak worden programmeeromgevingen ondersteund, maar auditors zijn veelal geen programmeurs. Zogenaamde 'no-code' omgevingen maken het voor auditors laagdrempeliger om zelf aan de slag te gaan, maar deze worden vaak nog niet ondersteund of zelfs gekend door de IT-afdeling.

Als auditors op een IT-afdeling stuiten die bepaalde analyse-software niet ondersteunt of waarvan het beleid het zelfs niet toestaat het te gebruiken, is dat lastig. Om niet meteen afhankelijk van IT te zijn, kan een optie zijn om een eigen lab-omgeving te creëren, en dit in een meer gevorderd stadium te integreren in de corporate architectuur. Voor het opzetten hiervan kan extern hulp ingeroepen worden als de kennis intern ontbreekt.

Mogelijk kunnen auditafdelingen van verschillende organisaties elkaar ook helpen door het delen van software-assessments (een analyse dat software veilig is om te gebruiken). In onze ervaring helpt het om het gesprek aan te gaan met de IT-afdeling om de behoefte en oplossing te bespreken binnen de kaders van het beleid.

Uitdaging 4: Succesvol aan de slag

Wij hebben ervaren dat zelfs als de organisatie data-analyse omarmt en de juiste tooling en kennis beschikbaar is, projecten niet automatisch daadwerkelijk succesvol zijn. Projecten worden te groot, duren te lang, data is niet beschikbaar, de focus verdwijnt of de impact is marginaal. Om de kans op succes te vergroten, hebben we een aantal relevante praktische tips, waarvan wij hebben ervaren dat ze werken:

I. Genereer ideeën en prioriteer op impact

In het ene onderzoek heeft tekstanalyse meer potentie om het efficiënter of beter te maken dan in het andere. Pas tekstanalyse toe in een onderzoek waarin door het toepassen ervan relatief veel winst te behalen valt. Bijvoorbeeld, omdat er erg veel documenten mee gemoeid zijn, of omdat belangrijke audit(deel)vragen aan de hand van tekstanalyse beantwoord kunnen worden. Hierbij is het vanzelfsprekend van belang te weten welke vragen het belangrijkst voor de branche, strategie en prioriteiten van de eigen organisatie zijn (Deloitte, 2013).

Een workshops en een brainstorm zijn een goed middel om met een groep, bijvoorbeeld het hele auditteam, te komen tot ideeën hoe tekstanalyse toe te passen op beschikbare teksten in de context van een audit. Relevante vragen zijn: Welke (periodieke) audit heeft veel te maken te teksten? Welke auditvragen zijn essentieel om te beantwoorden? Welke documenten of teksten zijn hierbij van belang? Waar steekt een auditor veel tijd in en zou tekstanalyse kunnen helpen? Hier kunnen ook 1e en 2e lijn bij betrokken worden, zodat een uitgevoerde analyse ook tot voor hen belangrijke inzichten leidt.

Als er ideeën zijn gegenereerd (bijvoorbeeld, op papieren of digitale post-its) kunnen deze worden geplot op een impact-moeite matrix om te prioriteren. Uiteindelijk zijn we op zoek naar een idee dat met relatief weinig moeite veel impact kan hebben. Als kennis van tekstanalyse vergroot over de tijd, zal de moeite steeds minder worden om de lastigere tekstanalyses uit te voeren.

II. Start op tijd

Het klinkt als een open deur, en is waar voor vrijwel elke vorm van data-analyse. Toch is deze tip relevant: start op tijd. Bijvoorbeeld, als tekstanalyse gebruikt wordt voor een steekproef, is het noodzaak de analyse meteen aan het begin of net voor de officiële start van de audit of onderzoek te doen. In het algemeen is het belangrijk om voor de start van de uitvoering van een onderzoek te starten met:

1. **het vergaren van de data**. Ervaring leert dat dit vaak langer duurt dan gedacht. Ten eerste is er toestemming nodig om data te analyseren voor het doel dat voor ogen is. Het is handig om de data protection officer vanaf het begin mee te nemen in de plannen om data te verkrijgen, zodat zij in de meedenkmodus komt. Denk hierbij ook aan overkoepelende afspraken om niet voor elk onderzoekje opnieuw in de papiermolen terecht te komen. Ten tweede dient de juiste aanleverende partij gezocht te worden die de data uit (bron)systemen haalt en die de data, waar nodig, anonimiseert. Ook deze partij werkt vaak met backlogs, waar je niet meteen bovenaan staat. Zowel het proces als de datavergaring zelf kunnen tijdrovend zijn.

2. **een analyse-aanpak**. Bedenk hoe de analyse moet worden uitgevoerd als de data beschikbaar is. Hierbij kunnen hypothesen worden geformuleerd om te toetsen in de analyse-fase. Bijvoorbeeld, voor de hypothese "*Beschrijvingsveld x is altijd gevuld met meer dan drie woorden*" ligt het tellen van woorden in het beschrijvingsveld voor de hand. Een hypothese-gerichte aanpak maakt het interpreteren van de uitkomsten makkelijker doordat focus wordt aangebracht op dat wat belangrijk is.

3. **het inlezen en exploreren van de data**. Voorafgaand aan de daadwerkelijke analyse is het nuttig om een idee te krijgen bij de kwaliteit, het formaat en de grootte van de teksten. Dit geeft inzicht in mogelijke lastige punten over de data (bijvoorbeeld, wat betekenen bepaalde codes?) waarvoor wellicht de business geconsulteerd moet worden.

III. Gebruik de data die er is en check de datakwaliteit

In de ideale wereld doen we analyses op een recente, complete dataset van goede kwaliteit, waarvan ook de metadata beschikbaar is zodat we precies weten wat alle kolommen betekenen. Echter, de werkelijkheid is vaak weerbarstig. We hebben te maken met documenten-verzamelingen die zo groot zijn dat het onmogelijk is om alle relevante documenten (snel) te analyseren, met niet goed ingevulde velden, met data uit verschillende systemen die moeilijk te matchen zijn, met verouderde data, enzovoorts. Het is roeien (analyseren) met de riemen (data) die we hebben. Laat je niet ontmoedigen. Toets, voorafgaand aan de analyse, of de datakwaliteit geschikt is voor het doel waarvoor je het gebruikt (fit-for-purpose). Slechte kwaliteit van data betekent niet automatisch dat je niet tot relevante inzichten kan komen. Focus op welke (deel)vraagstukken wel beantwoord kunnen worden met de data en welke inzichten wel verkregen kunnen worden. Het is altijd goed om, naast heldere conclusies, de beperkingen van het onderzoek aan te geven, zodat ook de auditee begrijpt hoe de vork in de steel zit.

De staat van de data kan overigens iets over de organisatie zelf zeggen: slechte datakwaliteit kan een indicatie zijn voor een niet goed werkend proces of een verouderd systeem. En dit kan dan weer gebruikt worden om te bespreken met de auditee: hoe goed zijn de processen en de systemen op orde? Het gebruiken van (tekst)analyse in de audit kan meehelpen om de organisatie stappen te laten zetten ten aanzien van de datakwaliteit.

IV. Start met simpele analyses

De kennis in tekstanalyse-cursussen kan behoorlijk diep gaan en het aantal toepassingen en mogelijkheden kan overweldigend zijn. De kunst is om, uitgaande van de relevantie voor een vraagstuk, simpel te starten in analyses, bijvoorbeeld met het tellen van of zoeken naar woorden. Mocht een eerste analyse niet het gewenste effect hebben, is het ook de kunst om niet op te geven en te bedenken wat de eerste analyse wel gebracht heeft (bijvoorbeeld, kennis over de tool) en te starten met een tweede simpele analyse die potentie heeft om impact te genereren. Hierbij komen we bij de laatste tip.

V. Begin klein en evalueer

Het 'think big, start small' principe is hier van toepassing: begin met een *proof of concept* (PoC), en breid pas uit als de waarde ervan is aangetoond of als anderen ervan zijn overtuigd. Vroeg in het proces een eerste presentatie maken over het doel en de aanpak, en eventueel wat eerste resultaten, helpt om het nut van de analyse te bespreken, andere auditors of 1e lijn te enthousiasmeren, en verder te komen met ideeën voor impactvolle analyses. Het kan een voordeel zijn met de makkelijkere analyses te beginnen waarbij snel resultaat geboekt wordt, zodat anderen snel de meerwaarde zien. Het met anderen bespreken van de uitkomsten van de PoC en de waarde ervan voor het onderzoek kan leiden tot meer gedragenheid. Scepsis van auditors kan overwonnen worden door concrete positieve resultaten.

Dit alles is in lijn met de 'Lean-startup-methode', dat steeds meer wordt toegepast in organisaties waar geïnnoveerd wordt. Een Lean startup experimenteert veel, haalt meteen feedback op en verbetert aan de hand van feedback (Sixsigma, 2023). Fouten mogen gemaakt worden, zorgen ervoor dat er geleerd wordt om vervolgens verbeteringen door te voeren. Ook in het proces om tekstanalyse naar een hoger volwassenheidsniveau te krijgen, zullen fouten gemaakt worden. Het is dan zaak te evalueren:
Wat werkt wel, wat werkt niet en wat moet er anders? En vervolgens door te gaan om verbeteringen door te voeren.

Er zijn steeds meer voorbeelden van auditafdelingen waar tekstanalyse wordt toegepast, met goede resultaten. Hier zijn belangrijke belemmeringen blijkbaar opgelost, en is een juiste manier gevonden om tekstanalyse in de audit toe te passen. In onze optiek zijn een relevant idee, gezonde interesse in tekstanalyse, en de juiste (gratis) tooling genoeg om te starten. Iemand met kennis van verandermanagement kan zeker zaken helpen versnellen.

Verder is het een kwestie van: *gewoon doen!*

Bibliografie

Auditdienst Rijk. (2021). *Onderzoek vastleggingen binnen project PEGA.* Opgehaald van https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2021/12/28/onderzoek-vastleggingen-binnen-project-pega/2021-0000269209+Onderzoek+vastleggingen+binnen+project+PEGA.pdf

Autoriteit Persoonsgegevens. (2022). *Mag u persoonsgegevens verwerken?* Opgehaald van Autoriteit persoonsgegevens: https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/algemene-informatie-avg/mag-u-persoonsgegevens-verwerken

Bruinenberg, R. (2022, June 15). *Het spel leidt tot een hogere productiviteit*. Opgehaald van https://optimaaldigitaal.gebruikercentraal.nl/het-spel-leidt-tot-een-hogere-productiviteit/

Datamation. (2021, April 9). *Text Analysis Tools*. Opgehaald van https://www.datamation.com/big-data/text-analysis-tools/

Deloitte. (2013). *Adding Insight to Audit: Transforming Internal Audit through Data Analytics.* Detroit, MI. Opgehaald van https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/audit/ca-en-audit-adding-insight-to-audit.pdf

Eilifsen, A., Kinserdal, F., Messier, W. J., & McKee, T. (2020). An Exploratory Study into the Use of Audit Data Analytics on Audit Engagements. *Accounting Horizons 34 (4)*, 75–103.

Evelson, B. (2022, Juni). *The Forrester Wave™: People-Oriented Text Analytics Platforms, Q2 2022.* Opgehaald van Forrester: https://www.forrester.com/report/the-forrester-wave-people-oriented-text-analytics-platforms-q2-2022/RES176358

Expert.AI. (2023, Januari). *The 2023 Expert NLP Survey Report: Trends driving NLP Investment and Innovation.* Opgehaald van Expert.AI: https://www.expert.ai/resource/the-2023-nlp-survey-report/

Gartner. (2017, September 11). *Advancing Audit's Use of Data Analytics.* Opgehaald van https://www.gartner.com/document/3938780

Gartner. (2020, January 06). *Deliver Data-Driven Insights.* Opgehaald van https://www.gartner.com/smarterwithgartner/deliver-data-driven-insights

Gold, A., & Majoor, B. (2022). Data-analyse is here to stay. *Maandblad voor Accountancy en Bedrijfseconomie 96(1/2)*, 1-3.

Johnson, R., Wiley, L., Moronay, R., Campbell, F., & Hamilton, J. (. (2021). *Auditing: a practical approach with data analytics.* John Wiley & Sons.

Kotter, J. P. (1996). *Leading Change.* Boston: Harvard Business School Press.

Li, X. (2022). Behavioral challenges to professional skepticism in auditors' data analytics journey. *Maandblad voor Accountancy en Bedrijfseconomie 96(1/2)*, 27-36. Opgehaald van https://mab-online.nl/article/78525/list/8/

Liew, A., Boxall, & Setiawan, D. (2022, May 3). The transformation to data analytics in Big-Four financial audit: what, why and how? *Pacific Accounting Review*. Opgehaald van https://www.emerald.com/insight/content/doi/10.1108/PAR-06-2021-0105/full/html

Mousa, H. (2020). *8 Open-source/ Free Text Mining and Text Analysis solutions.* Opgehaald van https://medevel.com/text-mining-and-text-analysis-solutions/

NBA. (2019). *NBA-handreiking 1141. Data-analyse bij de controle: uitdagingen en vooral kansen.*

OpenAI. (2023). Introducing ChatGPT. Opgehaald van https://openai.com/blog/chatgpt

Pat research. (2022). *Top 26 Free Software for Text Analysis, Text Mining, Text Analytics in 2022 - Reviews, Features, Pricing, Comparison.* Opgehaald van https://www.predictiveanalyticstoday.com/: https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

Schermer, B., Hagenauw, D., & Falot, N. (2018, Januari). *Handleiding Algemene verordening gegevensbescherming en Uitvoeringswet Algemene verordening gegevensbescherming.* Opgehaald van https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/handleidingalgemeneverordeninggegevensbescherming.pdf

Shahim, A., van Praat, J., Harmzen, P., & Matthijsse, R. (Red.). (2018). *Research in IT-Auditing; A Multidisciplinary View.* Amsterdam: Vrije Universiteit SBE. Opgehaald van https://research.vu.nl/en/publications/research-in-it-auditing-a-multidisciplinary-view

Sixsigma. (2023). https://www.sixsigma.nl/artikelen/wat-is-een-lean-startup. *Wat is een Lean Startup?*

Stuurman, H. (2022). Belemmeringen voor toepassing van data analytics in mkb-bedrijven. *Master's Thesis*.

Turner, B. (2022, November 01). Best speech-to-text apps of 2023. Opgehaald van https://www.techradar.com/news/best-speech-to-text-app

Van Vugt, H. (2022). Meer effect met data-analyse. *Audit magazine 21*, 62-65.

Wang, T., & Cuthbertson, R. (2015). Eight Issues on Audit Data Analytics We Would Like Researched. *Journal of Information systems (29)*, 155-162. Opgehaald van https://publications.aaahq.org/jis/article-abstract/29/1/155/1013/Eight-Issues-on-Audit-Data-Analytics-We-Would-Like

Wiseman Advocaten. (2020, Oktober 19). *Mag ik data scrapen van het internet?* Opgehaald van https://www.wisemen.nl/nl/artikelen/mag-ik-data-scrapen-van-het-internet-/

Instituut van
Internal Auditors
Nederland