Onderzoek je metadata

www.data-detox.nl

DATA
DETOX
KIT

Omschrijving workshop

We genereren constant metadata. Als we e-mails versturen, een bericht op het internet plaatsen of gewoon een stukje lopen met de mobiele telefoon in onze zak. In deze les onderzoeken we metadata: wat is het, waarom is het belangrijk en hoe kan je dagelijkse leven worden voorspeld als deze gegevens worden geanalyseerd.

Doelen

Kennis

• Wat is metadata?
• Hoe je met behulp van alleen maar metadata iets te weten kunt komen over iemand.

Vaardigheden

• Basisvaardigheden voor het werken met metadata.
• Maak het boeiend voor mensen met relevante, herkenbare voorbeelden.

Redenen

• Iemand kan aan de hand van de metadatasporen die je achterlaat heel veel over je te weten komen, zelfs als je de content van je activiteiten (e-mail, foto's, enz.) verbergt.
• Er zijn manieren om controle te krijgen over de metadata die je produceert.

Benodigde materialen en apparatuur

• Uitgeprinte browsermetadataset
• Scherm of projector om de videocirkel te laten zien

Referenties

• Visuele weergave van de metadata van Joel https://public.tableau.com/profile/publish/DataPatterns/Dashboard1#!/publishconfirm
• Handleiding data decoderen (Het onzichtbare zichtbaar maken) https://exposingtheinvisible.org/guides/decodingdata/
• Balthasar Glättli (animatie) https://vimeo.com/160231751
• https://apps.opendatacity.de/vds/index_en.html
• http://opendata.zeit.de/widgets/dataretention/
• Mehlis-rapport http://www.un.org/press/en2005/051021_Mehlis_Briefing.doc.htm
• Immersion https://immersion.media.mit.edu/
• Metadata strippen https://mat.boum.org/
• http://codewelt.com/stripper
• Tor https://www.torproject.org/
• Re:log videocirkel (animatie) https://myshadow.org/resources/relog?locale=en

Verschillende soorten metadata

Afbeeldingen

• De locatie (lengte- en breedtecoördinaten) waar de foto is genomen als er een apparaat met ingebouwde gps, zoals een smartphone, is gebruikt.
• Camera-instellingen, zoals ISO-snelheid, sluitertijd, brandpuntsafstand, diafragma, witbalans, type lens, enz. Let op, sommige camera's voegen de locatiecoördinaten toe.
• Merk en model van de camera of smartphone.
• Datum en tijd waarop de foto is genomen.
• Naam van het programma dat is gebruikt om de foto te bewerken.

Pdf-bestanden
• De naam van de auteur, meestal is dat de naam die is toegewezen toen het programma voor het eerst na de installatie werd gebruikt om een bestand te maken.
• De versie en naam van het programma dat is gebruikt om het bestand te maken.
• Titel van het document.
• Bepaalde trefwoorden.
• Datum en tijd van bestandsaanmaak en laatste wijziging.
• Tekstbestanden - afhankelijk van het programma dat is gebruikt om het document te maken, kunnen de volgende data zichtbaar zijn:
• de namen van alle verschillende auteurs.
• zinnen en opmerkingen die in eerdere versies van het document zijn gewist.
• aanmaak- en wijzigingsdatums.

Video's
• Metadata in videobestanden kan worden opgesplitst in twee delen:
• automatisch gegenereerde metadata: aanmaakdatum, omvang, opmaak, codecs, tijdsduur, locatie.
• handmatig toegevoegde metadata: informatie over de beelden, teksttranscriptie, tags, meer informatie en opmerkingen voor editors, enz.
• Aanbevolen lectuur: "A thorough overview on video metadata and working with it" van WITNESS.

Audio
Audiometadata is vergelijkbaar met videometadata, maar wordt op grotere schaal gebruikt, met name om te registreren van wie het bestand is. Daarnaast kan de metadata het volgende bevatten: aanmaakdatum, omvang, opmaak, codecs, tijdsduur en een reeks handmatig toegevoegde data zoals tags, informatie over de artiest, grafisch materiaal, opmerkingen, tracknummer op een album, genre, enz.

Router
• Naam van het apparaat
• MAC-adres van het apparaat
• IP-adres van het apparaat
• Registreren domeinnaam
• Naam, adres, e-mailadres, telefoonnummer

Communicatiemetadata
Dit is afhankelijk van het soort communicatie dat je gebruikt (bijv. e-mail, ouderwets mobieltje, smartphone, enz.). Maar over het algemeen is het volgende te zien (als er geen tools worden gebruikt om de metadata te verbergen):
• ID's van verzender en ontvanger
• datum en tijd van de communicatie
• locatie
• wijze van communiceren, enz.

Stappen
Stap 1: Inleiding tot de workshop en de relevantie met betrekking tot de Glass Room (10 min.)
1. Vertel kort iets over jezelf en de les, doe een voorstelrondje en laat de deelnemers vertellen wat ze in deze les willen leren.
2. Houd de verwachtingen van de deelnemers in gedachten als je een kort overzicht geeft van wat je gaat behandelen (en wat niet), noem de doelstellingen en vertel hoeveel tijd jullie daarvoor hebben.
3. Stel de spelregels vast. Voorbeelden: wees respectvol (er is telkens één persoon aan het woord, zorg dat iedereen meedoet, speel niet met je telefoon of

laptop als er iemand praat), privacy is belangrijk (neem geen foto's zonder het te vragen), domme vragen bestaan niet.

Stap 2: Content versus metadata (10 min.)

Wat is het verschil tussen content en metadata? Maak gebruik van de kennis in de groep, stel vragen:

- Wie kan een omschrijving geven van wat metadata is?
- Welk soort metadata wordt gecreëerd in een afbeelding, pdf, browser? (zie boven)
- Schrijf alle punten op een whiteboard en voeg ontbrekende soorten metadata toe.

Opmerking: metadata is essentieel voor digitale beveiliging en privacy.
Bovendien gaat het om "data over data, gegevens over gegevens". Het is een onderdeel van de gegevens die we produceren en die zoveel over ons vertellen. En veel mensen versturen zonder dat ze het weten een heleboel informatie (metadata) over zichzelf. Het is belangrijk dat je je hiervan bewust bent en dat je weet hoe dit ten goede en ten kwade gebruikt kan worden.

Stap 3: Hoe wordt metadata gebruikt? (10 min.)

Je kunt de deelnemers een overzicht geven van de manieren waarop metadata wordt gebruikt en vragen of ze zelf met voorbeelden willen komen.
Opmerking: hieronder staan voorbeelden die je kunt gebruiken. Als je zelf voorbeelden hebt, kun je die er natuurlijk bij zetten.
Het doel van datawetenschap is de mogelijkheid creëren om voorspellingen te formuleren. Er wordt op basis van ingevoerde data een model gebouwd voor het voorspellen van een resultaat.

Voorbeelden:

- Bij voorspellend politiewerk wordt gebruik gemaakt van diverse voorspellende factoren, zoals specifieke gegevens over de wijk, het weer, het schoolrooster, criminaliteitsstatistieken, etniciteit en dergelijke, om statistische modellen te maken die voorspellen waar en wanneer bepaalde misdaden waarschijnlijk gaan plaatsvinden. Dat is het scenario dat Philip K. Dick heeft geschetst in zijn roman Minority Report.
- De politie in New York begint met een experiment om te stoppen met het beleid van preventief fouilleren en aanhouden, ook al wordt deze techniek door veel andere politiekorpsen in de VS gebruikt. Om het aanhoudingsbeleid te kunnen veranderen, wordt er fors geïnvesteerd in analytische vaardigheden en dataminingtechnieken.
- Als je een projector hebt, kun je een van de video's laten zien om uit te leggen hoe de metadata van telefoons werd verzameld om de bewegingen van personen op een congres te kunnen volgen.
- https://myshadow.org/resources/relog?locale=en
- https://myshadow.org/resources/lifeofbalthasarglattli?locale=en
- https://www.youtube.com/watch?v=J1EKvWot3c

We kunnen met dergelijke gegevens ook dingen te weten komen over relaties en structuren.

Voorbeeld:

- Snowden bracht naar buiten dat de NSA (de Amerikaanse geheime dienst) op grote schaal metadata verzamelde. De NSA-programma's waren onder meer opgezet om de onderlinge relaties tussen mensen te ontdekken en communicatienetwerken te schetsen. Met het verzamelen en analyseren van data wilde de NSA informatie achterhalen waar ze nog niet van op de hoogte waren. In dit scenario wordt data-analyse gebruikt om nog onbekende feiten te ontdekken en op basis van deze feiten besluiten te nemen.

Als dergelijke modellen kunnen worden gemaakt, kunnen we de input dan op een gecontroleerde manier wijzigen, zodat we een bepaald resultaat kunnen afdwingen?

Voorbeelden:
• Om het netwerk van Facebook aantrekkelijker te maken voor adverteerders, biedt het de adverteerders een scala aan invoervariabelen. Een van die variabelen is ras; dat betekent dat Facebook adverteerders de mogelijkheid geeft gebruikers uit te sluiten op basis van ras. Dit wordt etnisch verwantschap genoemd.
• Adverteerders hadden 60 jaar geleden al door dat het gedrag van consumenten kan worden gestuurd. Data-analyse intensiveert deze kennis en biedt de mogelijkheid tot meer systematische benaderingen.
• En het werkt nog ook. Met de hedendaagse technische mogelijkheden kunnen we op zo'n grote schaal data verzamelen dat de analysemodellen de resultaten met grote nauwkeurigheid kunnen voorspellen.

Deze modellen zijn een middenweg tussen variantie en afwijking.
• Variatie: als we ons model maken op basis van bepaalde input, hoeveel zou ons model dan veranderen als we andere input hadden gebruikt om het te maken?
• Afwijking: een foutpercentage dat wordt veroorzaakt wanneer een reallife probleem, dat uiterst gecompliceerd kan zijn, door een veel simpeler model wordt nagebootst.
• Modellen kunnen het ook mis hebben door een verkeerde interpretatie van de gegevens, als gevolg van de menselijke factor, expres of onbedoeld.

Stap 4: Het analyseren van een browserdataset (20 min.)

Deze oefening geeft de deelnemers een idee van hoe data eruitziet en hoe het kan worden geanalyseerd. Geef de deelnemers de browserdataset. Vertel waar deze dataset vandaan komt. Maak kleine groepen en vraag ze om zoveel mogelijk uit te zoeken over deze persoon.
Vraag ieder groepje om verslag te doen.

Laat het beeldmateriaal van het onderzoek naar de browsermetadata zien. Hiermee wordt gevisualiseerd wat er kan worden gevonden als je gaat graven in die dataset.
Het beeldmateriaal is hier te vinden: https://public.tableau.com/profile/publish/DataPatterns/Dashboard1#!/publishconfirm

Opmerkingen:
• Een journalist heeft de browserdataset aan Tactical Tech gegeven. De media waren geïnteresseerd in wat er kan worden gevonden in de sporen van basisdata die je achterlaat op het internet. Tactical Tech heeft drie maanden van zijn browsegeschiedenis gekopieerd (tijdsaanduidingen, URL, bezoekersfrequentie, enz.). Met alleen deze gegevens kun je al een nauwkeurig beeld van iemand schetsen. De dataset is bijna 100% origineel, Tactical Tech heeft alleen de naam van de journalist geredigeerd.
• Het principe van dataverzameling is dat het niet gaat om dat ene op zichzelf staande gegevenspunt, maar dat je kunt gaan voorspellen wat mensen gaan doen, wat hun gedrag is, enz.

Stap 5: Afronding (10 min.)

• Waarschijnlijk zijn we al langer dan een uur bezig, dus misschien kun je de deelnemers beter naar de Bar of andere medewerkers verwijzen als ze nog vragen hebben. En je kunt ze verwijzen naar het appcenter in de Data Detox Bar.
• Je kunt ze ook de Data Detox Kit meegeven als ze hun digitale sporen willen uitwissen.