

Handleiding

OpenRefine



Robin van Rijn, Inge Vrencken, Joris Scholtenkamp, Ryquenco Echteld

Minor Datavisualisatie & Infographics

Inhoudsopgave

Inleiding	03
Leerdoelen: wat kan je verwachten?	03
Introductie tot de tool	04
Kernfuncties	04
Downloaden van OpenRefine	05
Problemen met Mac	06
Openen van OpenRefine	07
Cellen omzetten naar datum en toepassen van een filter	08
Custer en edit de dataset	09
Filteren van kolom en edit clusters	10
Exporteren van de dataset	12
Nawoord	13

Inleiding

Op de website zelf geeft OpenRefine, voorheen Google Refine, aan een “free, open source, powerful tool for working with messy data” te zijn. De tool is gratis en te gebruiken voor iedereen.

OpenRefine werkt goed voor het opschonen van rommelige data, en daar is het ook voor bedoeld. Wat er ook mee kan, is het organiseren van data.

Rommelige data is over het algemeen data wat door iemand met de hand geschreven is, of handmatig digitaal is opgeschreven door mensen, waardoor het gevoelig is voor spelfouten of verschillende formats.

Samengevat, Open Refine helpt met het onderzoeken van grote datasets. Dus wanneer je gebruik maakt van een dataset vanuit verschillende bronnen, kan je deze makkelijker op schonen. Ook kan je met OpenRefine data gemakkelijk transformeren

Leerdoelen: wat kan je verwachten?

In deze handleiding bespreken we een aantal leerdoelen. We beginnen met het downloaden en openen van het programma, waarna we vervolgens de volgende punten gaan leren:

Cellen omzetten naar datum en toepassen van een filter

Custeren en editen van de dataset

Filteren van kolom en edit clusters

Exporteren van de dataset

Introductie tot de tool

Open Refine is een software die oorspronkelijk geschreven is door David Huyn en ontwikkeld door Metaweb Technologies. Het bedrijf dat Open Refine heeft ontwikkeld werd in juli 2010 overgekocht door Google. Met deze overname vond Google het ook een goed idee om de software in een ander jasje te steken en de software Google Refine te noemen en marketen.

De eerste versie van Google Refine was een open source project, met de intentie om een database te creëren voor een community waar zij data kunnen opschonen, uploaden en deze ook kunnen vergelijken. Google deed dit doormiddel van een desktopapplicatie waarmee zowel online als offline gewerkt kon worden. In de twee jaar dat Google naamdrager was van Refine zijn er drie versies gemaakt van het programma, 2.0, 2.1 en 2.5.

Nadat de software de stap heeft gemaakt om afstand te doen van Google Refine is de software, onderdeel geworden van Code for Science & Society. CS&S is een non-profit joint venture van experts. CS&S streeft naar een toekomst waarin research, data en technologische innovaties draagkrachtig zijn en impact hebben binnen community.

In oktober 2012 ging Refine zich logischerwijs meer focussen op haar community. Dit betekende dat Google afstand deed van de software. Momenteel is Open Refine een programma dat door liefhebbers en kenners wordt ontwikkeld in samenwerking met CS&S. Momenteel zijn er over de hele wereld ontwikkelaars die het hun taak hebben gemaakt om te zorgen dat de software blijft draaien. Doordat het een open source software is, komt de software in 15 talen. Om ervoor te zorgen dat alle gebruikers gemakkelijk datasets op kunnen schonen, verdiepen, transformeren en in context te plaatsen. Wanneer je aan de slag gaat met een project, maakt de software geen gebruik van cloud-opslag.

De combinatie van de desktopapplicatie en lokale opslag geeft de gebruiker alle verantwoordelijkheid over haar projecten en data. De ontwikkelaars hebben geen toegang tot deze data en kunnen dit ook niet krijgen. Verder in deze handleiding zal er een verdieping zijn, waarin je stapsgewijs enkele functies door loopt binnen de software.

Kernfuncties

Explore data

Als je gebruik maakt van grote data sets is het soms gewenst om deze gemakkelijk op te schonen omdat publieke data vaak niet consistent en een troep zijn. In Open Refine wordt dit gedaan doormiddel van verschillende "Facet". Een veelgebruikte "Facet" is de "Text Facet" Deze functie clustert kolommen en geeft weer hoeveel cellen in deze rij dezelfde tekst hebben. Hierdoor is het gemakkelijk om kolommen die op elkaar lijken met elkaar te koppelen. Spelfouten, synoniemen en nog veel meer zijn met deze "Text Facet" gemakkelijk op te schonen.

Transform data

Naast het opschonen van rommelige data zal je misschien een dataset hebben die schoon is maar qua waardes niet is geschikt voor jou doeleinde. Open Refine heeft handige functies om data te transformeren in de gewenste vormen. Van lijst naar tabel bijvoorbeeld. In deze functie geef je een input en een output om de software te laten zien welk deel je wilt transformeren.

Reconcile and match data

Maak je gebruik van Extensies dan is Open Refine de software die een gemakkelijke workflow creëert. Maar ook als je data mist kan je gemakkelijk via Open Refine andere extensies en plug-ins laten zoeken naar de overige data. Ook heeft de software een functie waarin het zelf relevante onderwerpen toevoegt aan je data. Heb je een data set over sport maar wil je ook atleten beschrijven dan zal de functie "Reconcile" zelf op zoek gaan naar de atleten die relevant zijn.

Downloaden van OpenRefine

Stap 1:

Ga naar <https://openrefine.org/>

Stap 2: Klik rechts op "Download"

Stap 3: Download de laatste versie. In dit geval is dat OpenRefine 3.6.2.

Kies de juiste kit.

Gebruik je Windows? > Kies de Windows Kit. (Afhankelijk of je al Java hebt kies je voor de download met of zonder Java)

Gebruik je Mac? > Kies de Mac Kit

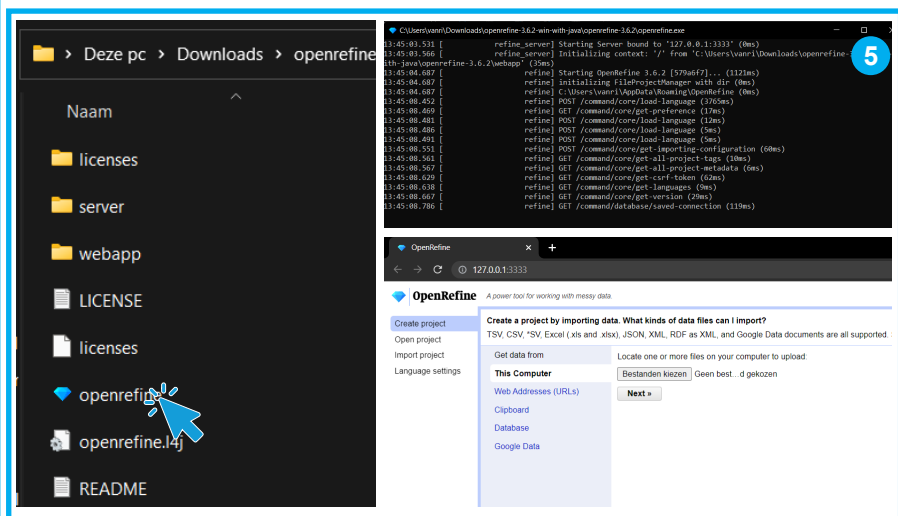
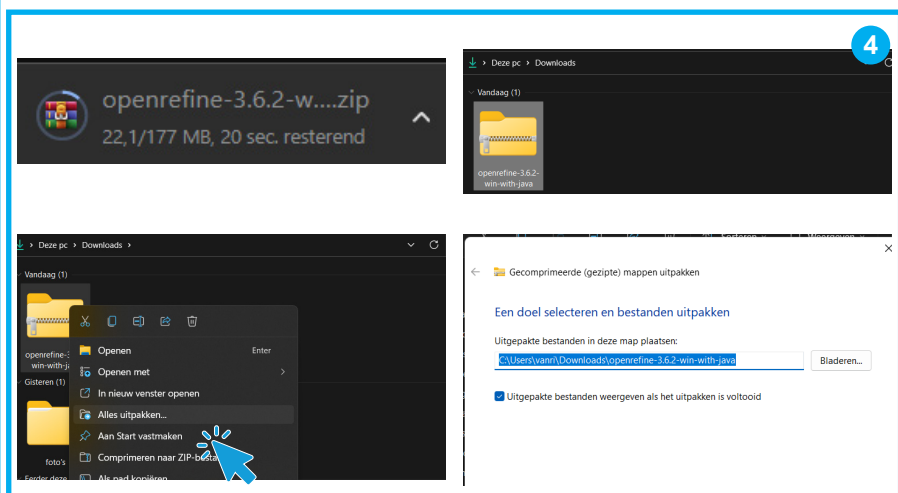
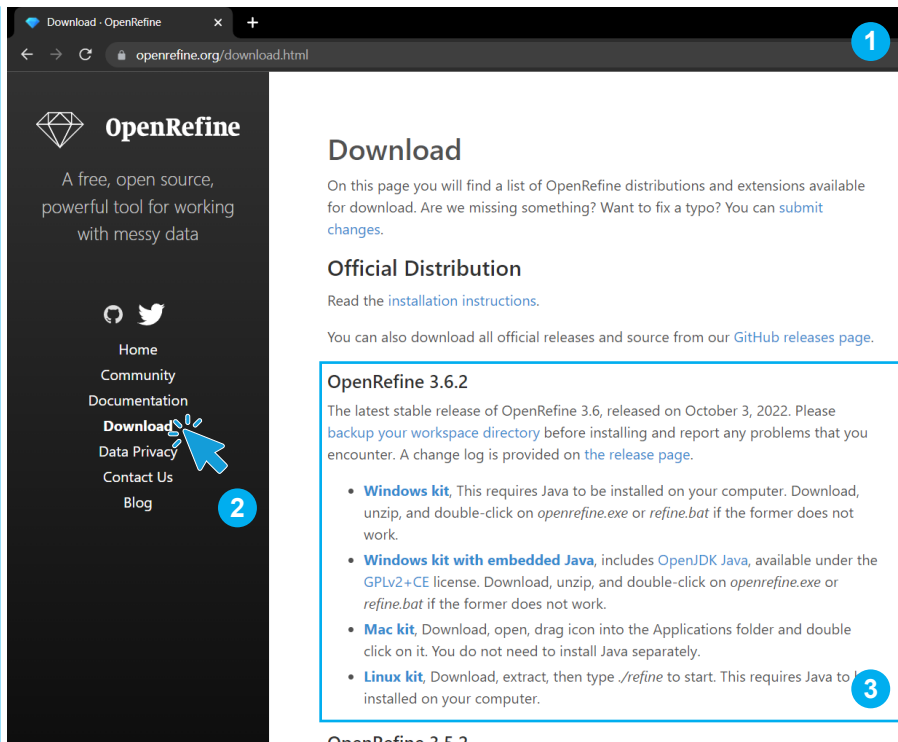
Gebruik je Linux? > Kies de Linux Kit

Stap 4: Als het programma is gedownload vind je deze in je downloads. Pak het ZIP bestand uit door met de rechtermuis op het pakket te klikken en kies voor "Alles uitpakken". Geef de bestanden een locatie waarin jij ze gemakkelijk terug kan vinden.

Stap 5:

In de map vind je nu een hoop bestanden waaronder de toepassing van OpenRefine (te herkennen aan het diamantje).

OpenRefine start nu met het opstarten van het programma. Dat kan je zien doordat er een command prompt (zwarte vlak met witte tekst) wordt gestart. Voor het werken met OpenRefine heb je geen internetverbinding nodig. OpenRefine draait via de command prompt een kleine webserver op je computer. Je hebt toegang tot deze webserver via je browser. Dit kan je zien doordat er een IP adres is geopend in je adresbalk. Wanneer je dus werkt met OpenRefine werk je dus lokaal. Je hoeft geen (gevoelige) bestanden de uploaden in een cloud.



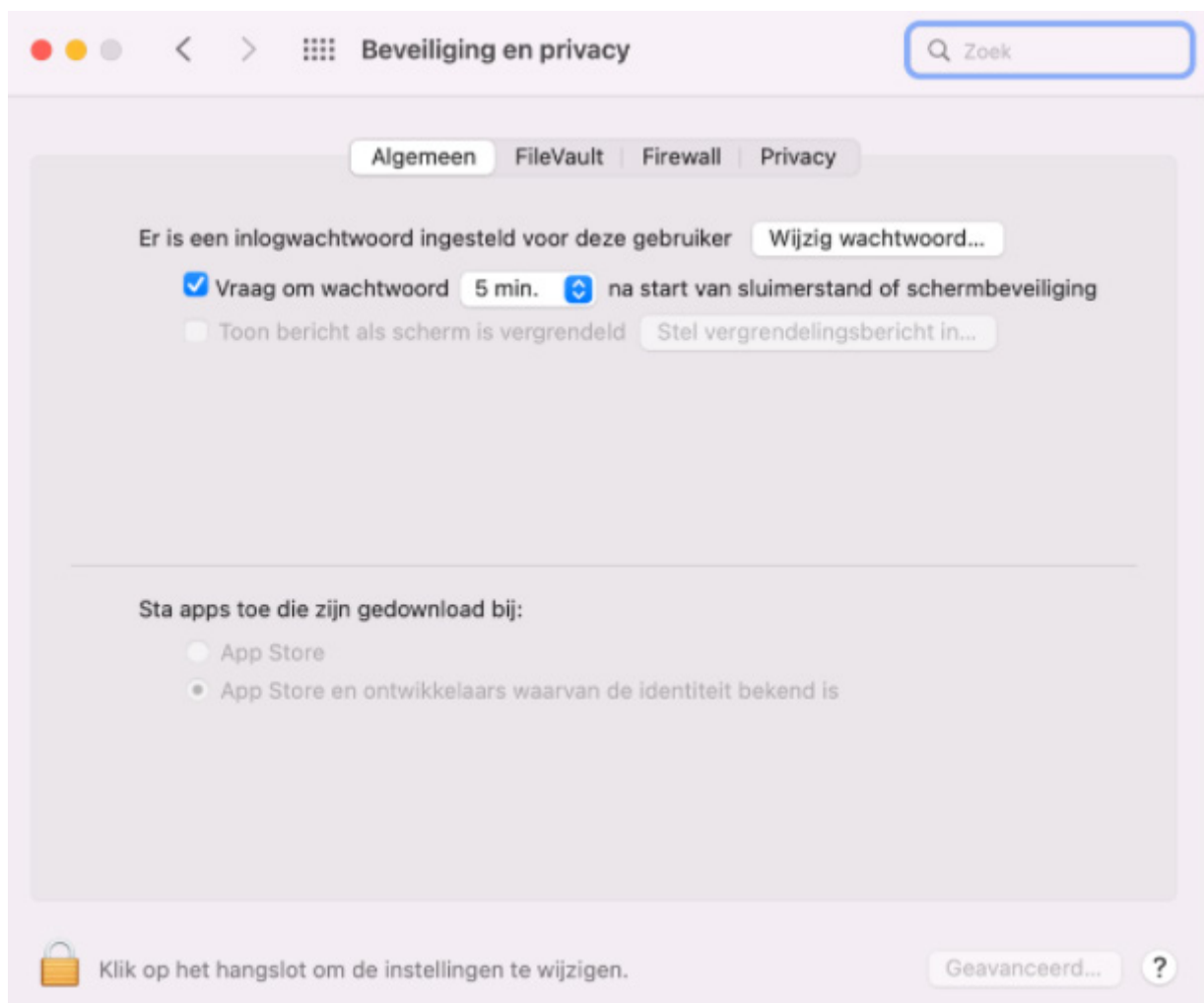
Problemen met Mac

Ga naar <https://openrefine.org/download.html> en kies de meest recente download die aansluit op jouw laptop (Macbook, Windows, Linux)

Op een Mac kan het voorkomen dat het programma in eerste instantie niet geopend kan worden omdat de ontwikkelaar niet geverifieerd kan worden.

Om dit op te lossen moet je naar **Systeemvoorkeuren > Beveiliging en privacy**. Hier krijg je de mogelijkheid om het hangslot te openen en toestemming te geven aan OpenRefine.

Als dit is gedaan kan het programma geopend worden en zal deze nu wel opstarten.



Openen van OpenRefine

Stap 1: Wanneer je voor het eerst werkt met OpenRefine moet je een project aanmaken. Je kan de volgende soorten datafiles importeren:

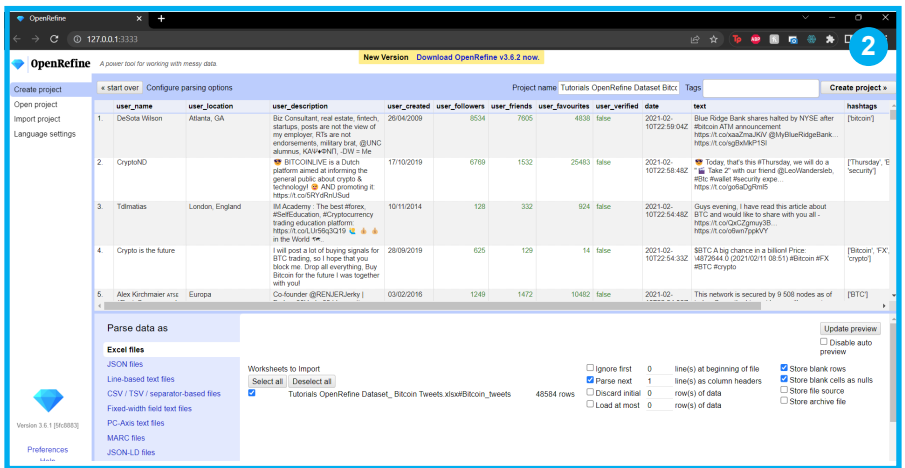
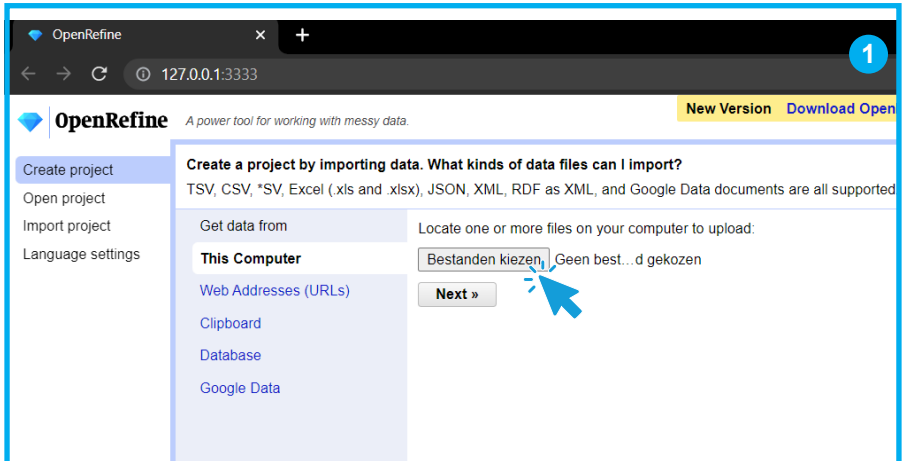
TSV, CSV, *SV, Excel (.xls en .xlsx), JSON, XML, RDF als XML, en Google Data documents.

Je kan deze bestanden importeren op verschillende manier:

Een lokaal bestand vanaf de computer uploaden, via een URL, kopiëren naar het clipboard, via een database of via Google Data (spreetsheet).

In deze tutorial uploaden we een Excel bestand vanaf de computer. Klik op "This Computer" en kies je bestand.

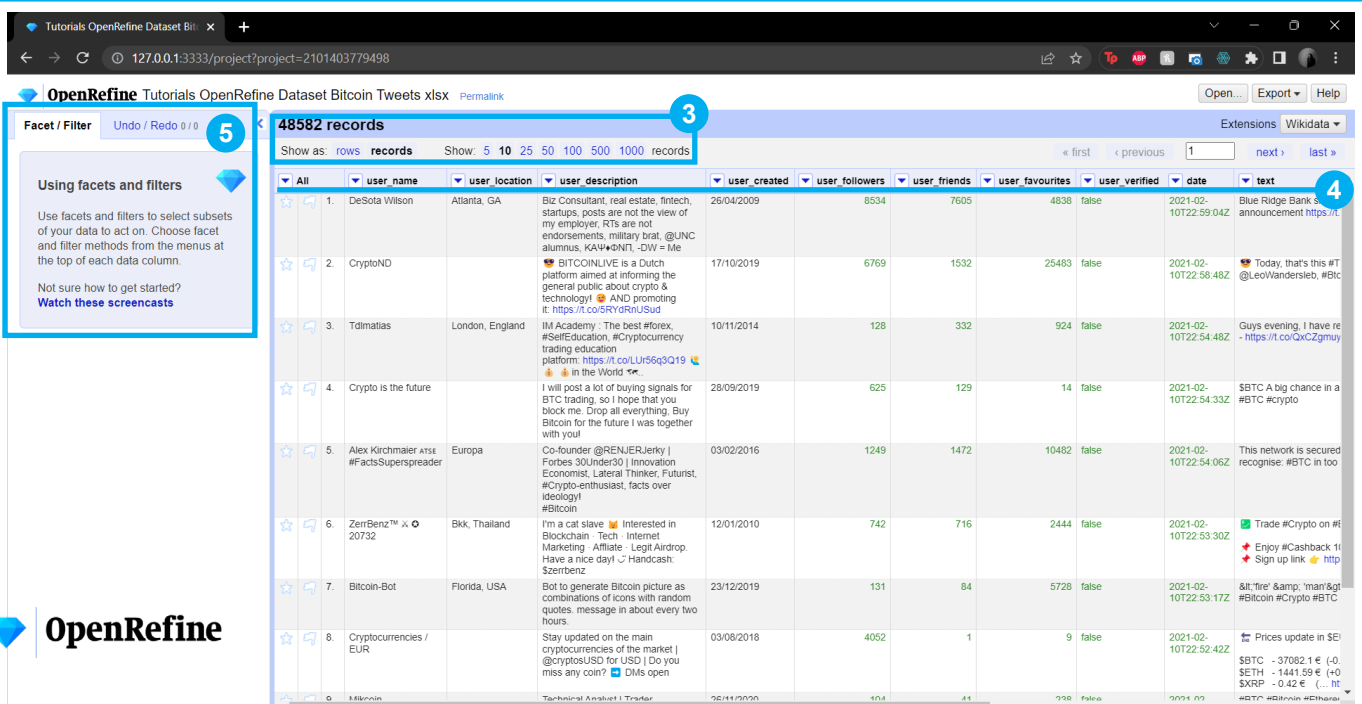
Stap 2: OpenRefine gaat nu de data uploaden en inspecteren. Wanneer dat is gebeurd zie je een preview van de data. Hier kan je nog verschillende opties aanpassen. Heeft een Excel bestand bijvoorbeeld meerdere worksheets, dan kan je gemakkelijk aanvinken in welke je wil werken. Wanneer alle opties naar wens zijn ingevuld klik je op de knop "Create project" rechts bovenin.



Stap 3: OpenRefine verteld ons dat er 48582 records zijn gevonden in deze dataset. Bij het kopje "Show" kan je kiezen hoeveel records OpenRefine je laat zien.

Stap 4: Bovenaan de data set vind je de beschrijving van de verschillende data zoals dit ook al in het Excel bestand stond. Bij elke titel vind je een pijltje naar beneden. Hiermee zullen we werken om verschillende acties op toe te passen.

Stap 5: Links vind je de Facet / Filter. Hier zien we straks alle filters die we gaan toepassen. In het tapje ernaast zie je Undo / Redo. Hier houd het programma de al je acties bij. Heb je iets verkeerd gedaan, kan je dat in dit tapje gemakkelijk ongedaan maken.



Cellen omzetten naar datum en toepassen van een filter

Stap 1: In deze handleiding gebruiken we een dataset van tweets die gaan over Bitcoin als voorbeeld. Allereerst gaan we werken met het kopje “user_created”. Zoals je kan zien zijn de kolommen rechts van deze rij groen. Dat betekent dat OpenRefine al heeft herkend dat het cellen waar nummers en datums in staan. Voor de kolom “user_created” gaan we dit ook aanpassen.

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified
1.	DeSota Wilson	Atlanta, GA	Biz Consultant, real estate, fintech startups, posts are not the view of my employer. RTs are not endorsements, military brat. @UNC alumnus. KAN+QNTL, -DW = Me	26/04/2009	8534	7605	4838	false
2.	CryptoND		BITCOINLIVE is a Dutch platform aimed at informing the general public about crypto & technology! AND promoting it. https://t.co/R9YR9H1Sud	17/10/2019	6769	1532	25483	false
3.	Tdimatias	London, England	IM Academy - The best forex, #SelfEducation, #Cryptocurrency trading education platform: https://t.co/LU56q3Q19	10/11/2014	128	332	924	false
4.	Crypto is the future		I will post a lot of buying signals for BTC trading, so I hope that you block me. Drop all everything, Buy Bitcoin for the future I was together with you!	28/09/2019	625	129	14	false
5.	Alex Kirchnermaier Arse #FactsSuperspreader	Europa	Co-founder @RENIERJerky Forbes 30Under30 Innovation Economist, Lateral Thinker, Futurist, #Crypto-enthusiast, facts over ideology! #Bitcoin	03/02/2016	1249	1472	10482	false

Stap 2: Klik op het pijltje naar beneden. Ga naar “Edit Cells” > “Common transforms” (hier vind je ook de andere transforms zoals To number, To text etc.) > “To date”.

(Datums worden weergegeven volgens de Amerikaanse datumnotatie.)

Stap 3: Bovenin ontvang je de melding dat de cellen in de kolom “user_created” zijn veranderd. Zoals je kan zien zijn de cellen ook groen geworden.


Als we nu een filter willen toepassen op de datums, is dat nu mogelijk omdat OpenRefine nu herkent dat de cellen datums zijn.

	user_description	user_created	user_verified
1.	Biz Consultant, real estate, fintech startups, posts are not the view of my employer. RTs are not endorsements, military brat. @UNC alumnus. KAN+QNTL, -DW = Me	2009-04-26T00:00:00Z	
2.	BITCOINLIVE is a Dutch platform aimed at informing the general public about crypto & technology! AND promoting it. https://t.co/LU56q3Q19	2019-10-17T00:00:00Z	
3.	IM Academy - The best forex, #SelfEducation, #Cryptocurrency trading education platform: https://t.co/LU56q3Q19	2014-10-10T00:00:00Z	
4.	I will post a lot of buying signals for BTC trading, so I hope that you block me. Drop all everything, Buy Bitcoin for the future I was together with you!	2019-09-28T00:00:00Z	
5.	Co-founder @RENIERJerky Forbes 30Under30 Innovation Economist, Lateral Thinker, Futurist, #Crypto-enthusiast, facts over ideology! #Bitcoin	2016-02-03T00:00:00Z	

Stap 4: Klik weer op het pijltje naar beneden. Ga naar “Facet” > “Timeline facet”. In het kopje Facet / Filter is nu een nieuw filter toegepast op de dataset. Met de sliders kan je de gewenste datum aanpassen. De dataset past zich automatisch aan op de waarden van het filter. Om het filter te verwijderen klik je op het kruisje.

Custer en edit de dataset

Stap 1: In de dataset van het voorbeeld is de kolom "user_location" ingevuld door de mens zelf. Wat betekent dat er (zeker in z'n grote dataset) veel menselijke fouten te vinden zijn. Een handige toepassing van OpenRefine is "Cluster and Edit". Hiermee analyseert OpenRefine alle ingevulde cellen en clustert het programma alle tekst die op elkaar lijkt. Op deze manier is het mogelijk om de dataset op te schonen.

Stap 1: Klik in de kolom "User location" het pijltje  naar beneden. Ga naar "Edit cells" > "Cluster and edit".

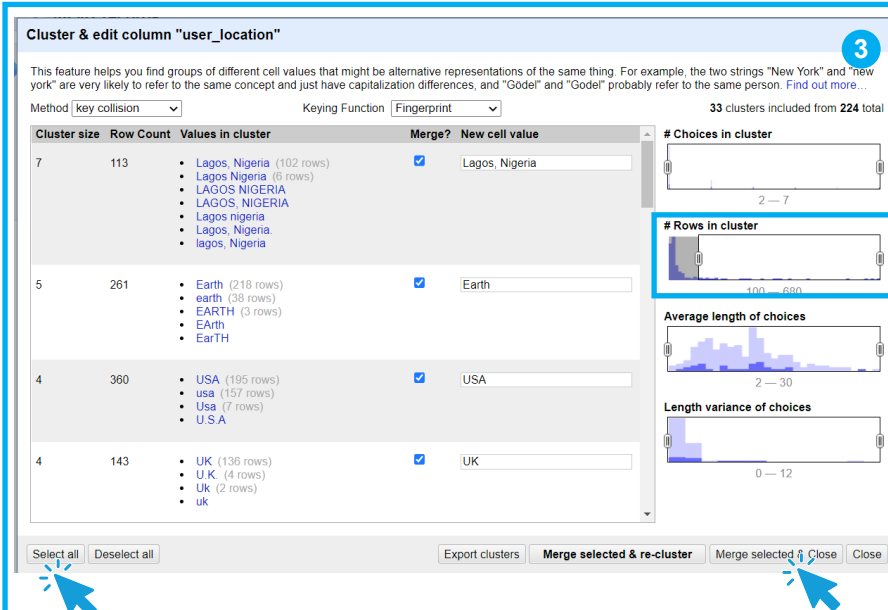
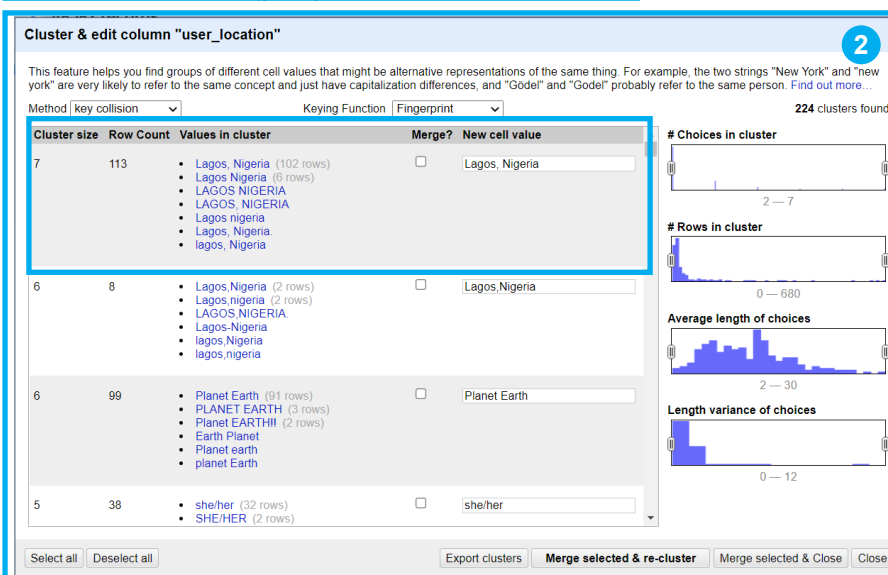
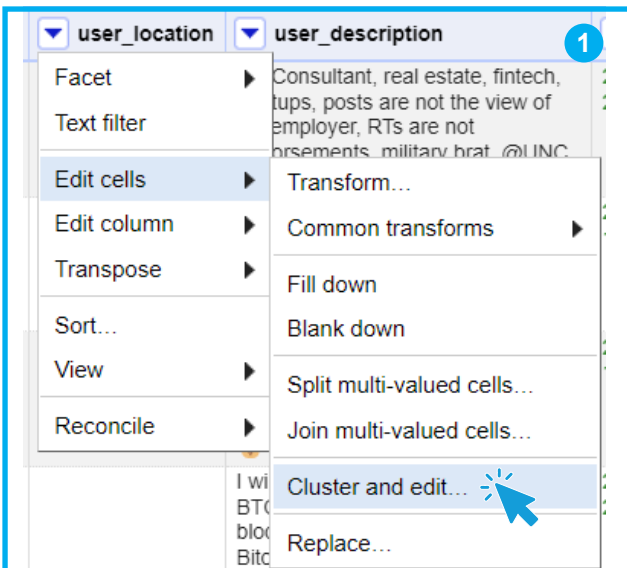
Stap 2: OpenRefine geeft ons nu een lijst met suggesties voor een nieuwe cel waardes. Zoals te zien is in de eerste rij is *Lagos, Nigeria* op 7 verschillende manieren geschreven in de dataset. Deze verschillen worden onderverdeeld in clusters. Deze clusters bestaan samen uit 113 regels in de dataset. OpenRefine geeft een suggestie voor een nieuwe cel waarde. Dit betekent dat alle 7 clusters de nieuwe schrijfwijze krijgen die je invult in de nieuwe cel waarde.

Rechts zijn er nog verschillende filters te vinden. Wil je bijvoorbeeld alleen werken met de clusters met de meeste rijen kan je er voor kiezen om dat bij "# Rows in cluster" te filteren.

Stap 3: In het voorbeeld clusteren we de dataset op de volgende manier:

"Pas het filter "# Rows in cluster" toe op 100 - 600 regels per cluster" > "Selecteer alle regels via "Select all" (Bij de kolom Merge? worden nu de regels geselecteerd)" > "Klik op Merge selected & Close"

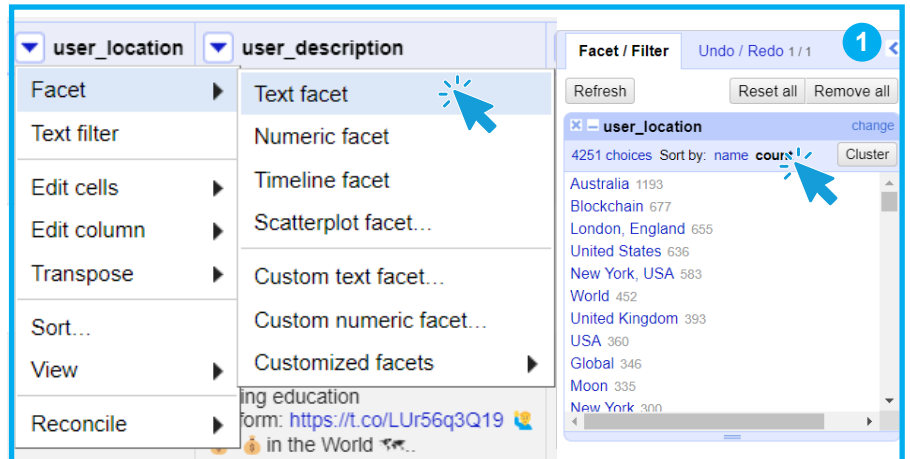
De melding verteld ons dat we de dataset hebben geclusterd.



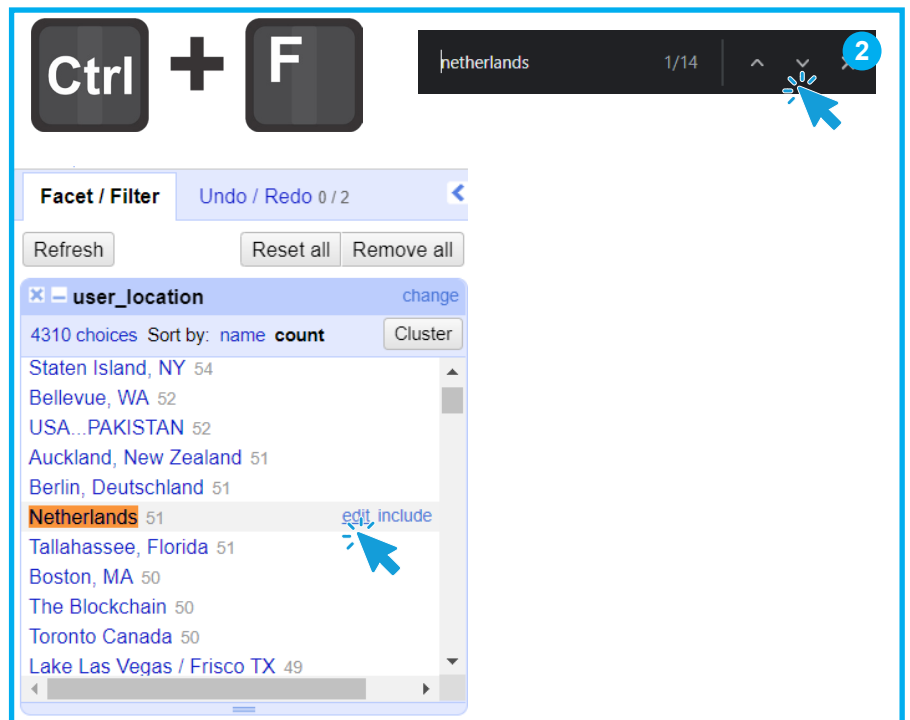
Mass edit 8538 cells in column user_location Undo

Filteren van kolom en edit clusters

Stap 1: Klik op het pijltje  naar beneden. Ga naar **“Facet” > “Text facet”**. Links verschijnt nu bij het kopje Facet / Filter een tekstfilter van de kolom “user_location”. Sorteer de resultaten op aantallen door bij “Sort by” te klikken op “count”. De locaties met de hoogste aantallen staan nu bovenaan.

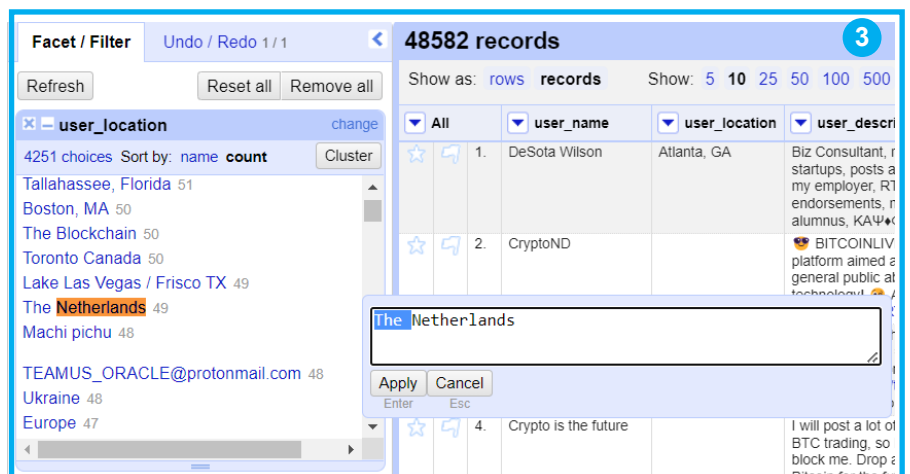


Stap 2: In het voorbeeld willen we graag alle Nederlandse clusters vinden. Met Ctrl of Cmd + F zoeken we op de pagina naar Netherlands. Met oranje wordt in het filter nu het cluster Netherlands opgelicht. We krijgen 14 resultaten. Ga met de pijltjes van de zoekfunctie door de lijst heen.



Stap 3: Elk cluster die de locatie *Netherlands* moet krijgen passen we aan.

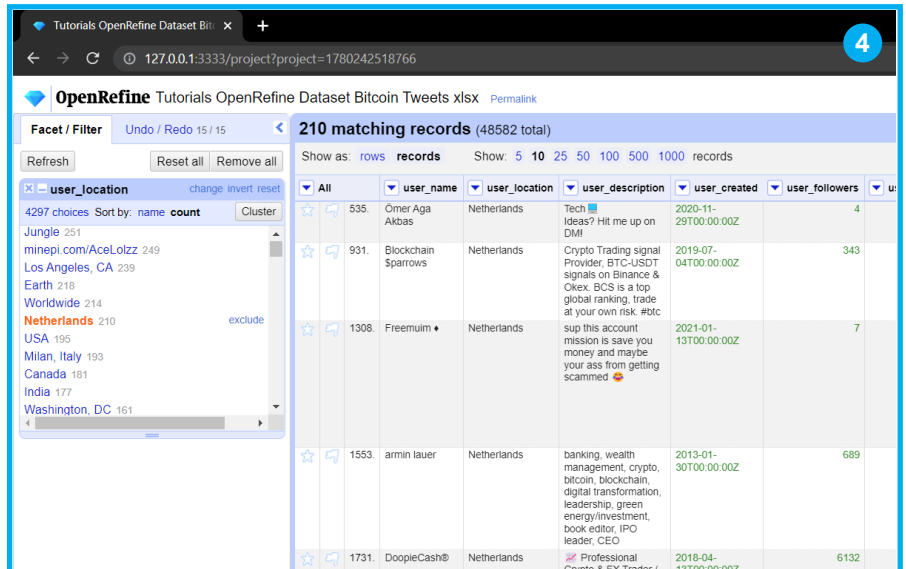
Om de naam van een cluster aan te passen klik je op “Edit”. Je kan nu in het tekstvlak de tekst aanpassen zoals jij die zou willen. Als je exact de zelfde schrijfwijze toepast, voegt OpenRefine de clusters automatisch samen.



Mass edit 49 cells in column user_location Undo

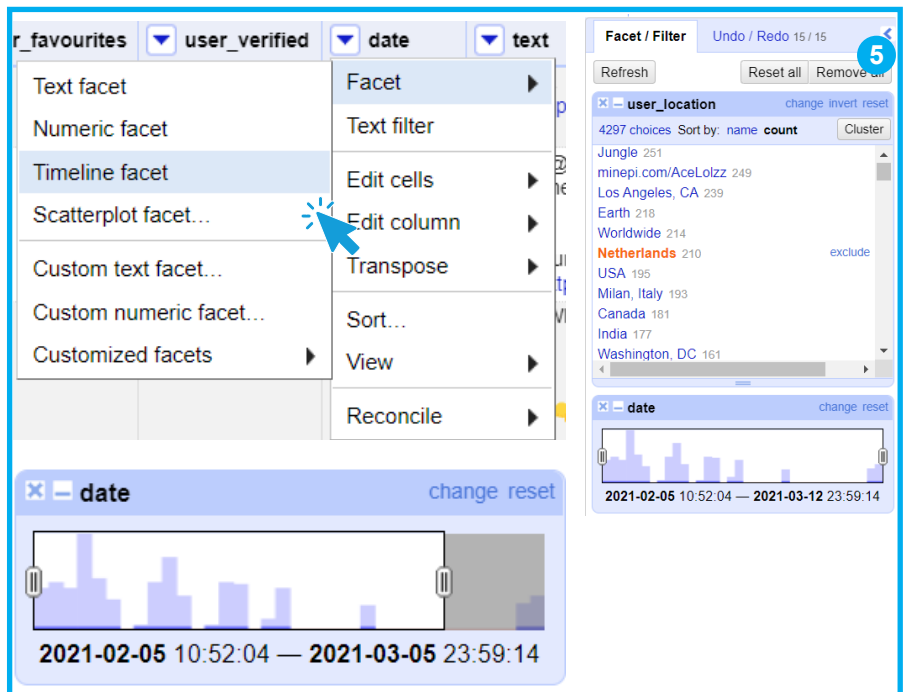
Filteren van kolom en edit clusters

Stap 4: Door de schrijfwijze aan te passen van de 14 resultaten naar één schrijfwijze bestaat het cluster "Netherlands" nu uit 210 regels. Klik op het cluster. Nu zien we in de preview alleen 210 regels.



The screenshot shows the OpenRefine interface with a project named 'Tutorials OpenRefine Dataset Bitcoin Tweets.xlsx'. The 'Facet / Filter' panel on the left shows a list of locations, with 'Netherlands' selected and highlighted in orange, indicating it has 210 records. The main table displays 210 matching records, with columns for 'user_name', 'user_location', 'user_description', 'user_created', and 'user_followers'. The table shows records for users from the Netherlands, such as 'Omer Aga Akbas', 'Blockchain Sparrows', 'Freemuim', 'armin lauer', and 'DoopleCash@'. A blue circle with the number '4' is in the top right corner.

Stap 5: Als laatste willen we graag de resultaten nog filteren op de datum van de posts. Klik weer op het pijltje naar beneden. Ga naar "Facet" > "Timeline facet". Het filter wordt toegevoegd in het kopje Facet / Filter. Met de twee sliders passen we het bereik aan van de datums die we willen zien. In het voorbeeld kiezen we voor 2021-02-05 t/m 2021-03-05.



The screenshot shows the OpenRefine interface with the 'Facet / Filter' menu open. The 'Timeline facet' option is highlighted, and a blue arrow points to it. The 'date' column is selected in the menu. Below the menu, a 'Timeline facet' for the 'date' column is visible, showing a histogram of data points. The date range is set to '2021-02-05 10:52:04' to '2021-03-05 23:59:14'. A blue circle with the number '5' is in the top right corner.

Exporteren van de dataset

The screenshot shows the OpenRefine interface with a dataset of 197 records. The 'user_location' facet is set to 'Netherlands' and the 'date' facet is set to a range from 2021-02-05 to 2021-03-05. The 'Export' menu is open, and the 'Excel (.xlsx)' option is highlighted. The main table displays columns for user_name, user_location, user_description, user_created, user_followers, user_friends, user_favorites, and user_verified.

Stap 1: We hebben nu een dataset die geclusterd en gefilterd is naar onze wensen. Als resultaat hebben we nu 197 regels geselecteerd uit het totaal van 48582 regels. Deze willen we gaan exporteren naar Excel zodat we er daar verder mee kunnen werken. Rechtsboven klik je op "Export" > "Excel (.xls)". Er download nu een Excelbestand waarin alleen de Nederlandse tweet staan, gefilterd op een bepaalde periode waarin je weer verder kan werken.

The screenshot shows an Excel spreadsheet with the following columns: user_name, user_location, user_description, user_created, user_followers, user_friends, user_favorites, user_verified, date, text, hashtags, source, and is_retweet. The data is sorted by user_location, showing only records from the Netherlands. The first few rows are:

user_name	user_location	user_description	user_created	user_followers	user_friends	user_favorites	user_verified	date	text	hashtags	source	is_retweet
Omer Aga Akbas	Netherlands	Tech Ideas? Hit me up on DM!	2020-11-29T00:00:00Z	4	40	94	ONWAAR	2021-02-10	loading...	['Bitcoin', 'BTC']	Twitter for iPhone	ONWAAR
Blockchain Sparrows	Netherlands	Crypto Trading signal Provider. BTC-USDT signals	2019-07-04	343	264	755	ONWAAR	2021-02-10	Thanks @cryptocurrency, 'bitcoin', 'crypt	['Bitcoin', 'BTC', 'crypto']	Twitter for iPhone	ONWAAR
Freemuim	Netherlands	sup this account mission is save you money and r	2021-01-13	7	47	2	ONWAAR	2021-02-10	People w [dogecoin, 'BUYANDHOLD', 'Bit	['Bitcoin', 'BTC', 'dogecoin']	Twitter for iPhone	ONWAAR

Nawoord

In deze handleiding zijn verschillende onderwerpen besproken die mogelijk zijn binnen OpenRefine. Mocht je nieuwsgierig zijn naar meer mogelijkheden, vanuit het programma zelf is er uitgebreide documentatie te vinden op <https://openrefine.org/documentation.html>

