Panama und Paradise Papers

Die Bedeutung von Text und Data Mining für den investigativen Journalismus

Hannes Munzinger, Süddeutsche Zeitung

Die großen Leaks

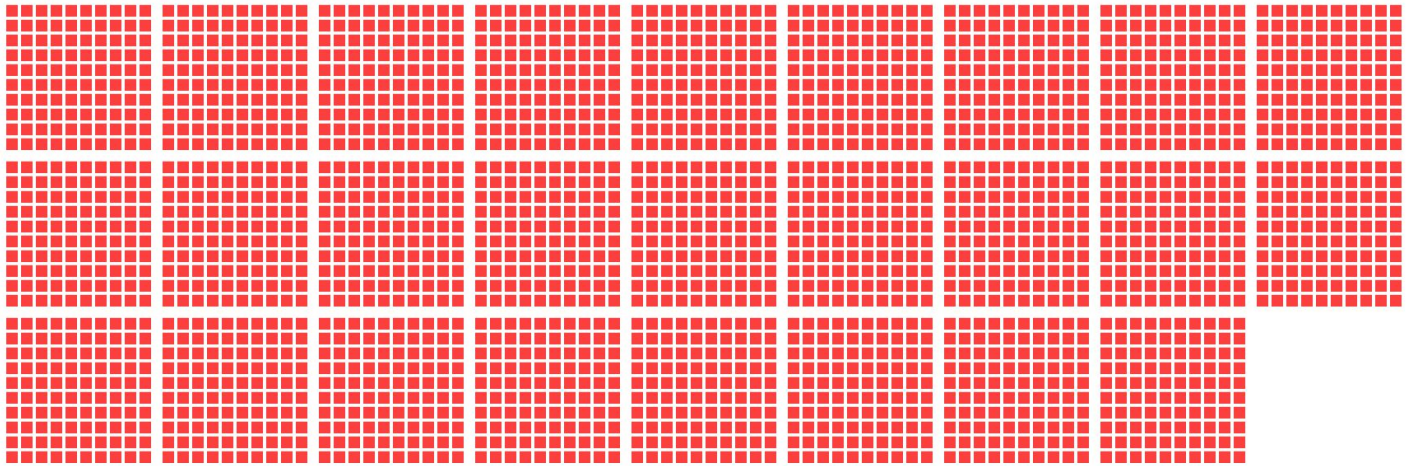
1,7 GB
Cablegate/Wikileaks (2010)

260 GB
Offshore-Leaks/ICIJ (2013)

4 GB
Luxemburg-Leaks/ICIJ (2014)

3,3 GB
Swiss-Leaks/ICIJ (2015)

≈ 2,6 TB
Panama Papers/ICIJ (2016)



HELLO. THIS IS JON DOE.

- Süddeutsche Zeitung wurde von einer **anonymen Quelle** kontaktiert
- Die Quelle übermittelte **interne Dokumente** der Kanzlei Mossack Fonseca
- Die Quelle wollte weder finanzielle Kompensation, noch irgendetwas anderes außer einiger **Sicherheitsvorkehrungen**

Jon Doe

«For the record: I do not work for any government or intelligence agency and I never have»

- Die Kommunikation verlief über verschiedene verschlüsselte Wege
- Warum tut die Quelle das? Sie will “diese Verbrechen öffentlich machen”

Die Firma – Mossack Fonseca

Die Firma – Mossack Fonseca

Kunden können für einige hundert Dollar anonyme Firmen kaufen

Mehr bezahlen? Mehr Geheimhaltung!

Unter Mossack Fonsecas Kunden waren:

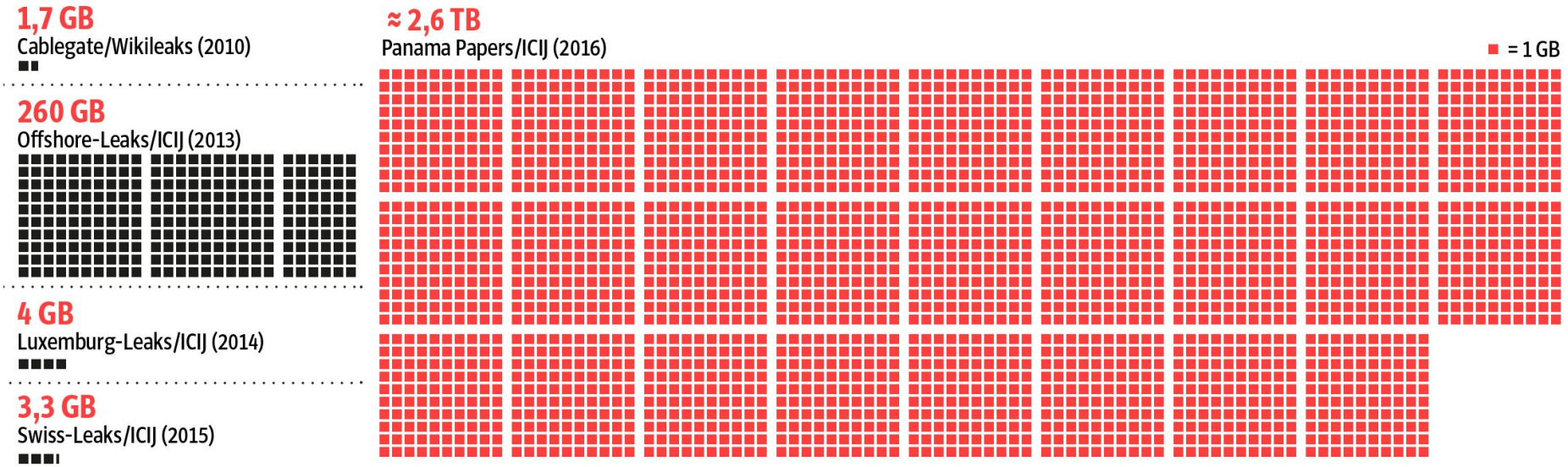
12 frühere und aktuelle Staatsoberhäupter

128 Politiker und Beamte

29 Milliardäre

Die Firma hat alle Vorwürfe abgestritten: **«Weder unterstützt unser Unternehmen gesetzwidrige Handlungen noch fördert es diese gar»**

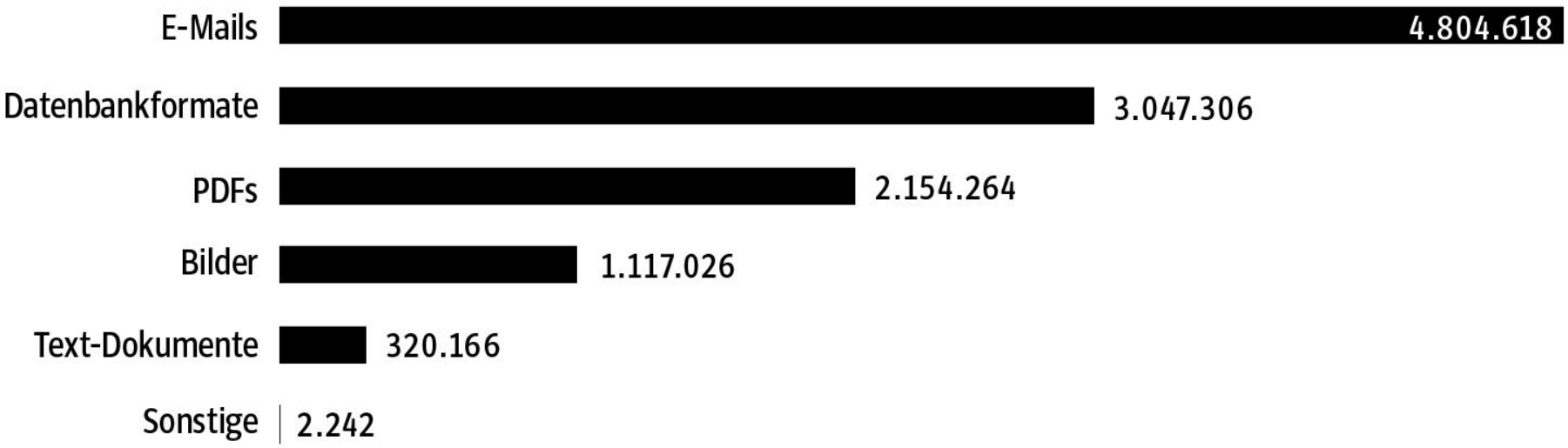
Big Data? Natürlich nicht.



Inside Mossack Fonseca

Wie sich das Leak zusammensetzt

In den 11,5 Millionen Dokumenten sind folgende Dateitypen enthalten



Was wir vorfanden

- Jede Briefkastenfirma hat einen eigenen Ordner
- Emails, Verträge, Urkunden, gescannte Dokumente, Pässe
- Manchmal mehrere Tausend Dokumente in einem Ordner

SHAREHOLDER	CERTIFICATE No.	DATE OF ISSUE	NUMBER OF SHARES	AUTHORIZED CAPITAL
THE BEARER	**2**	**13th February, 2007**	**3.333**	**US$ 10.000,00**

Certificate No. 2 **Number of shares: 3.333**

TICTOM CORP.

Incorporated under the International Companies Act, 1987
of Samoa

AUTHORISED CAPITAL: US$ 10.000,00
Divided into 10.000 shares with a par value of US$ 1,00 each

This is to certify that THE BEARER is entitled to THREE THOUSAND THREE HUNDRED THIRTY THREE (3.333) shares of US$ 1,00 each in the above named company, subject to the Memorandum and Articles of Association of the said company.

IN WITNESS WHEREOF the Company has authorised this certificate to be issued this 13th day of February, 2007.

Director

Director

Director

Director

Herausforderungen für das Team

- Kaum Erfahrung im Umgang mit großen Datenmengen
- Keine leistungsfähige Hardware verfügbar
- Keine geeignete Software zur Analyse vorhanden

Der pragmatische Dreischritt

1. Vorbereiten der Daten

- Die Daten zugänglich machen
  - Die Dokumente wurden auf Hochleistungsrechner geladen
  - Texterkennung (OCR) bei allen technisch möglichen Dokumenten
  - Erstellung eines Suchindex

2. Abgleichen der Daten

- Abgleich von Listen mit interessanten Namen
  - Die Suchsoftware war in der Lage Namenslisten mit den 11.5 Millionen Dokumenten abzugleichen
- Quellen:
  - Personendatenbanken wie munzinger.de
  - händisch erstellte Listen zu bestimmten Skandalen oder Personengruppen

Named Entity Recognition

- Input: Fachliteratur
- Output des Stanford Named Entity Recognizer:

```
JONATHAN MARTIN PERSON  and
PATRICK HEALY   PERSON  JULY 27 , 2016
PHILADELPHIA    LOCATION        -- President
Obama   PERSON  delivered a stirring valedictory address at the Democratic convention Wednesday night , hailing
Hillary Clinton PERSON  as his rightful political heir and the party 's best hope to protect democracy from `` home
Donald J. Trump PERSON  .

                Taking the stage to rapturous roars of `` We love you '' and `` Yes we can , '' Mr.
Obama   PERSON  acknowledged that Democrats were still divided after a bruising nomination fight and that Mrs.
Clinton PERSON  had made `` mistakes . ''

                But he vouched passionately for Mrs.
Clinton PERSON  as a trusted and reliable ally not just for him but for all Americans who need a fighter to improve

                `` She 's been there for us -- even if we have n't always noticed -- and if you 're serious about c
just because she might not align with you on every issue , '' Mr.
Obama   PERSON  said , an explicit appeal to supporters of Senator
Bernie Sanders  PERSON  who continue to resist Mrs.
Clinton PERSON  .
```

- Weiterverarbeitung zu Namenslisten

3. Auswerten der Daten

- Jeder gefundene Name wurde klassisch ausrecherchiert:
  - Was ist die Rolle dieser Person im Firmennetzwerk?
  - Woher kommt das Geld?
  - Wohin geht das Geld?
  - Ist diese Konstruktion legal?

- Tausende Spuren wurden aufgenommen:
  - ”Cross-check” von Beweisen
  - ”Open-source intelligence”
  - Themendossiers aus Pressearchiven
  - Experteninterviews

Paradise Papers

Paradise Papers

- Nicht ein Leak, 21 Leaks
- Unternehmensinterna von
  - Offshore Kanzlei Appleby
  - Treuhandfirma Asiacity Trust
- Unternehmensregister von 19 Jurisdiktionen

Wo kann Text Mining schnell helfen?

Wonach suche ich in 13 Millionen Paradise Papers?

Mit welchen Dokumenten fange ich an?

Welche Themen sind die wichtigsten?

Topic Modelling

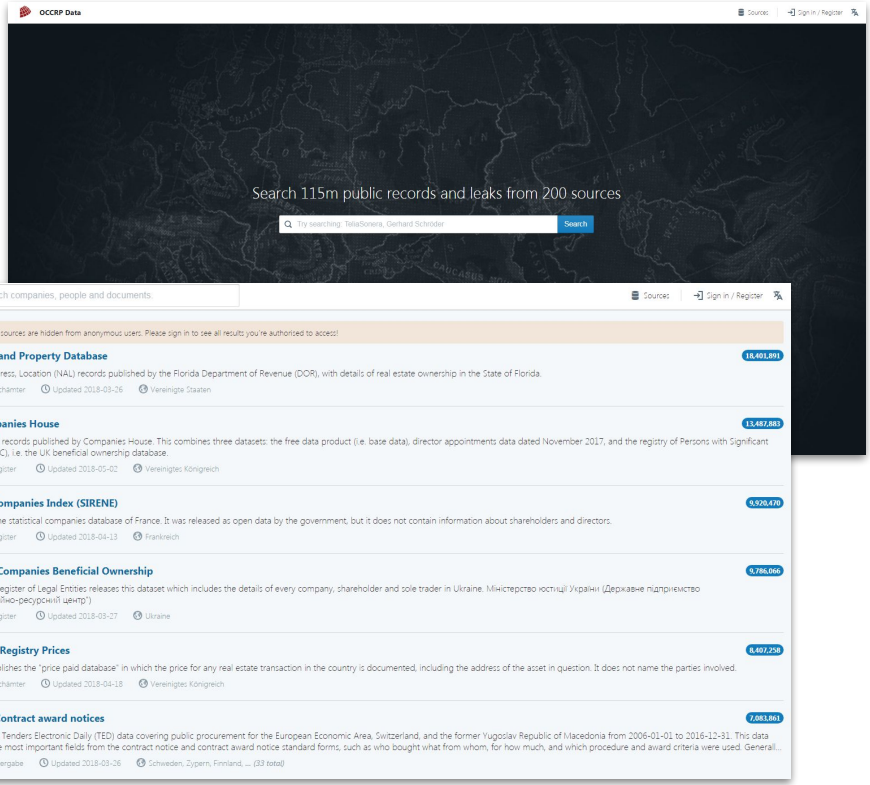
- Identifizierung von übergreifenden Themen in einer Sammlung von Dokumenten
- basierend auf Worthäufigkeiten in einer Document-Term-Matrix
- Unsupervised Machine Learning

Topic Modelling - Notwendige Schritte

1. Dokumente vorbereiten
    a. Texterkennung
    b. Text extrahieren
2. Korpus erstellen
3. Document-Term-Matrix erzeugen
    a. Interpunktion, Ziffern, Großschreibung entfernen
    b. Stopwörter filtern
    c. ggf. Wortstammreduktion anwenden
4. Modell trainieren
5. Sinnvolle Zahl der Topics ermitteln
6. Sinnvolle Darstellung finden

Integration von TDM in Forensiksoftware

- Open Source Software “Aleph”
- https://github.com/alephdata/aleph
- leistet
  - Import unzähliger Dateiformate
  - OCR
  - Sprachenerkennung
  - Entity Extraction
  - Cross-Matching



Datenquellen selbst erfassen

Exklusivität ohne Leaks: Scraping

- Anwendungsfälle für Web-Scraping
  - Information ist an keiner anderen Stelle vorhanden
  - Information ist vorhanden, aber nicht zugänglich

- Anwendungsfall Scraping von Dokumenten
  - Information ist schwer zugänglichem Format vorhanden (z.B. Tabellen in PDFs)

Grenzgänge beim Web-Scraping

- Ist die Nutzung von Proxy-Netzwerken verdecktes Recherchieren?
- Wie limitieren die AGBs von Sozialen Netzwerken das Vorgehen?
- Wie dürfen wir erhobene, teils personenbezogene Daten für Belegzwecke speichern?
- Dürfen wir sie zu Forschungszwecken weitergeben?