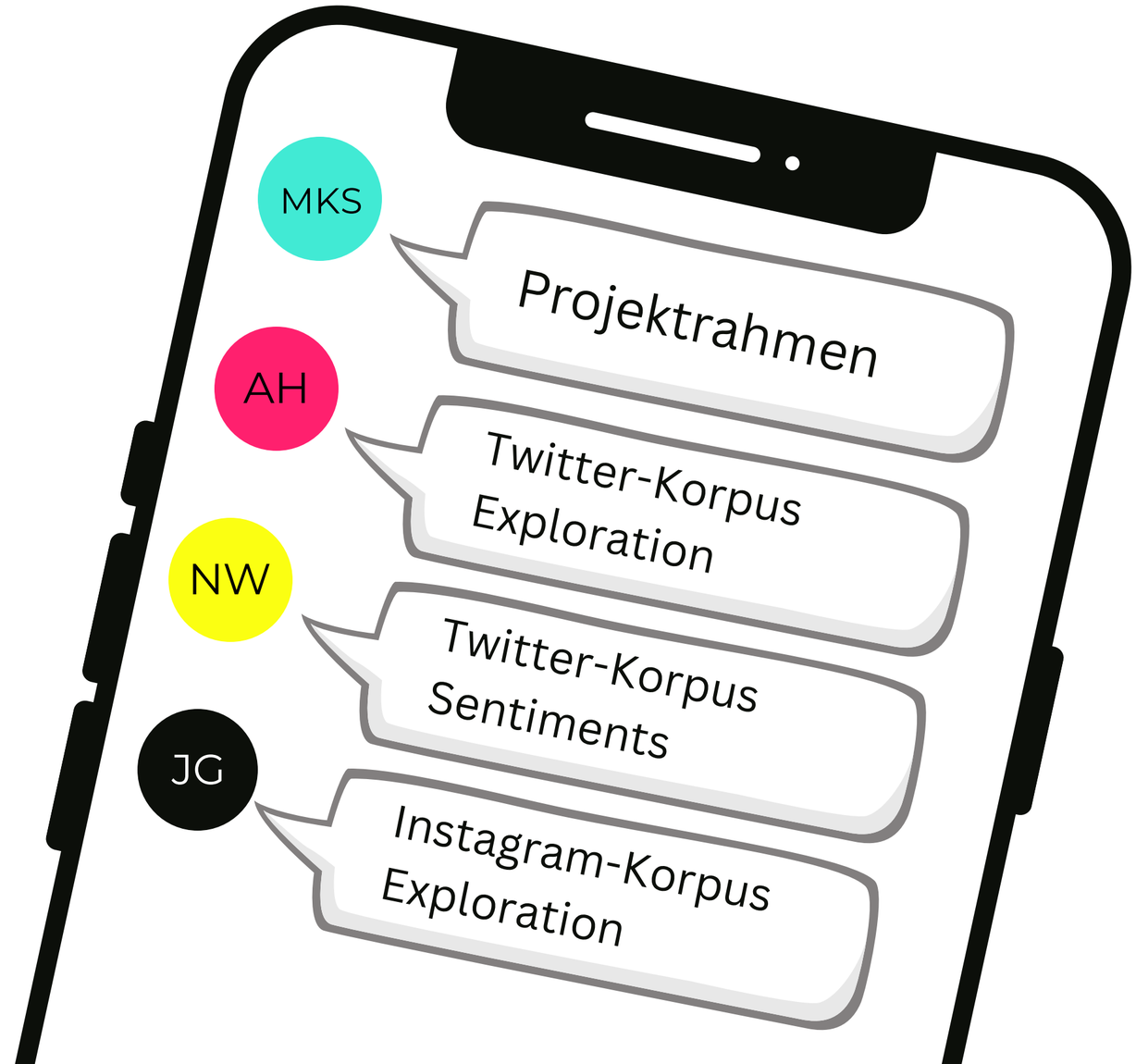
Vortragende: Mareike Schumacher, Annalena Hiergeist, Nikolai Wolle, Johanna Gruenler
Universität Regensburg

Student, Professor, Journalist

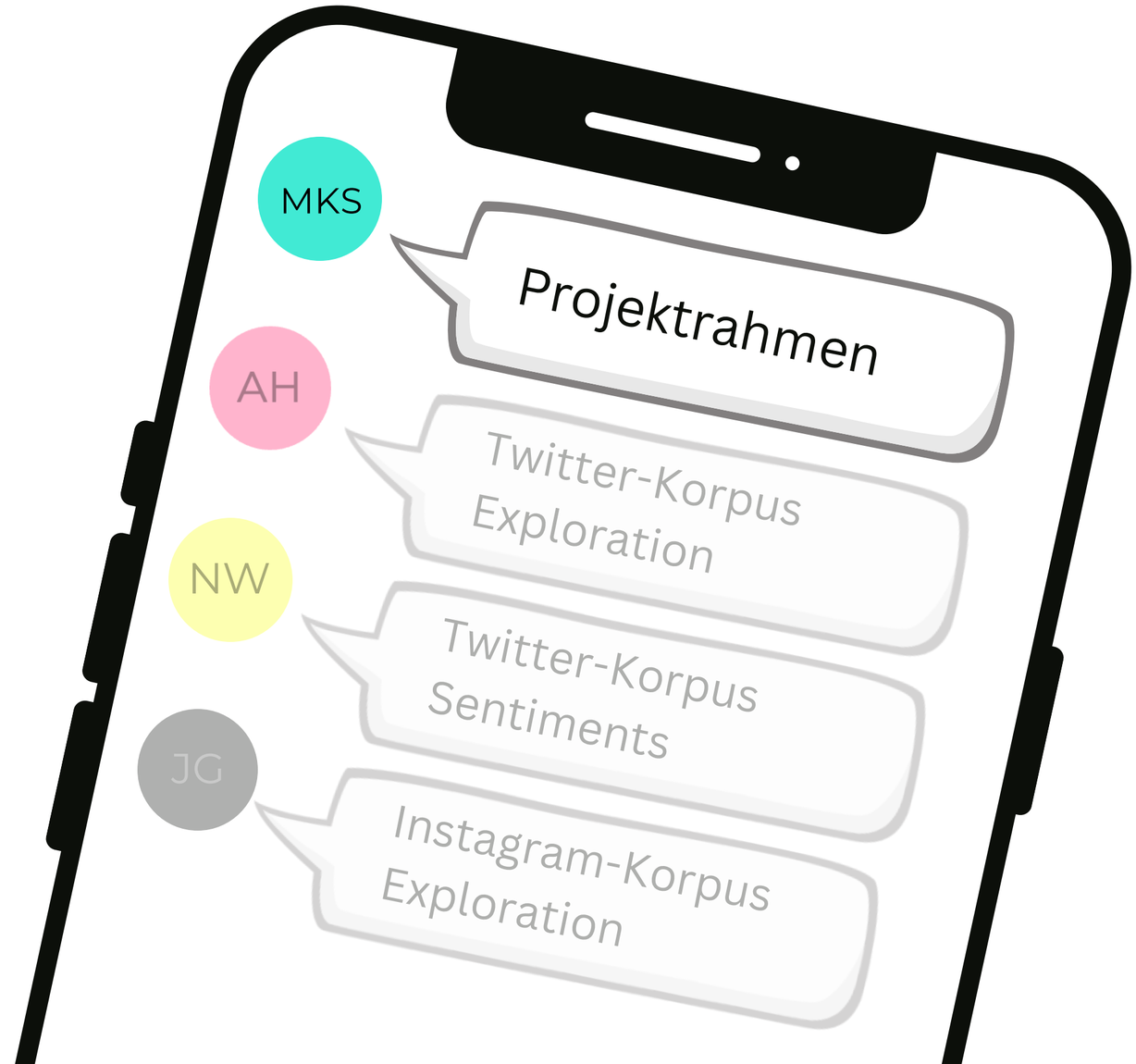
Profilierung von Akteur*innen des Forschungsdiskurses in den sozialen Medien (ein Werkstattbericht)

Mutter, Autorin, Frau

Outline


MKS
Projektrahmen
AH
Twitter-Korpus
Exploration
NW
Twitter-Korpus
Sentiments
JG
Instagram-Korpus
Exploration

Teil I


MKS
Projektrahmen
AH
Twitter-Korpus Exploration
NW
Twitter-Korpus Sentiments
JG
Instagram-Korpus Exploration

WIM-Kooperation

Dachprojekt: Wissenschaftlerinnen in die Medien

CI: Judith Ackermann (FH Potsdam)

Forschung, Baby: qualitative Sozialforschung an Social Media Daten

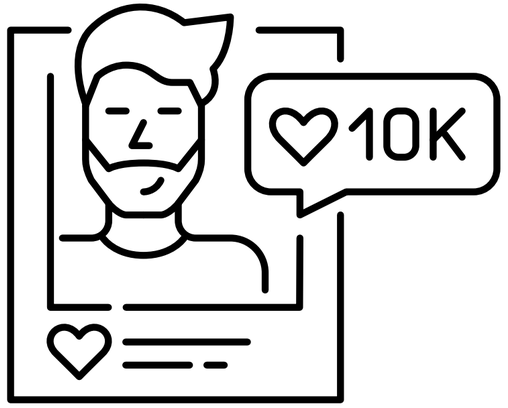
#IchBinWissenschaftlerin: Stärkung von Wissenschaftlerinnen in den (sozialen) Medien durch gezielte Kampagnen auf Instagram und TikTok

DH-Komponente: Wissenschaftler*innen in den Medien

CI: Mareike Schumacher (Uni Regensburg)

Forschung, Baby: Big-Data-Analysen und quantitative Auswertungen von Social-Media-Korpora

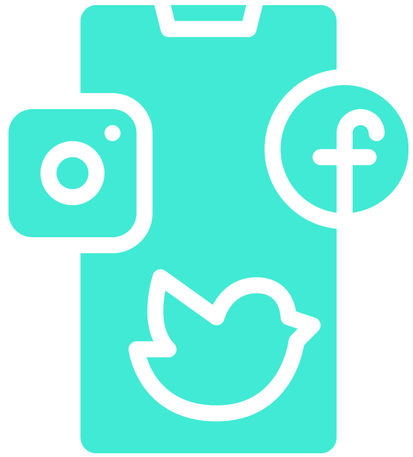
Rahmen: Projektseminar des MA-Studienganges Digital Humanities



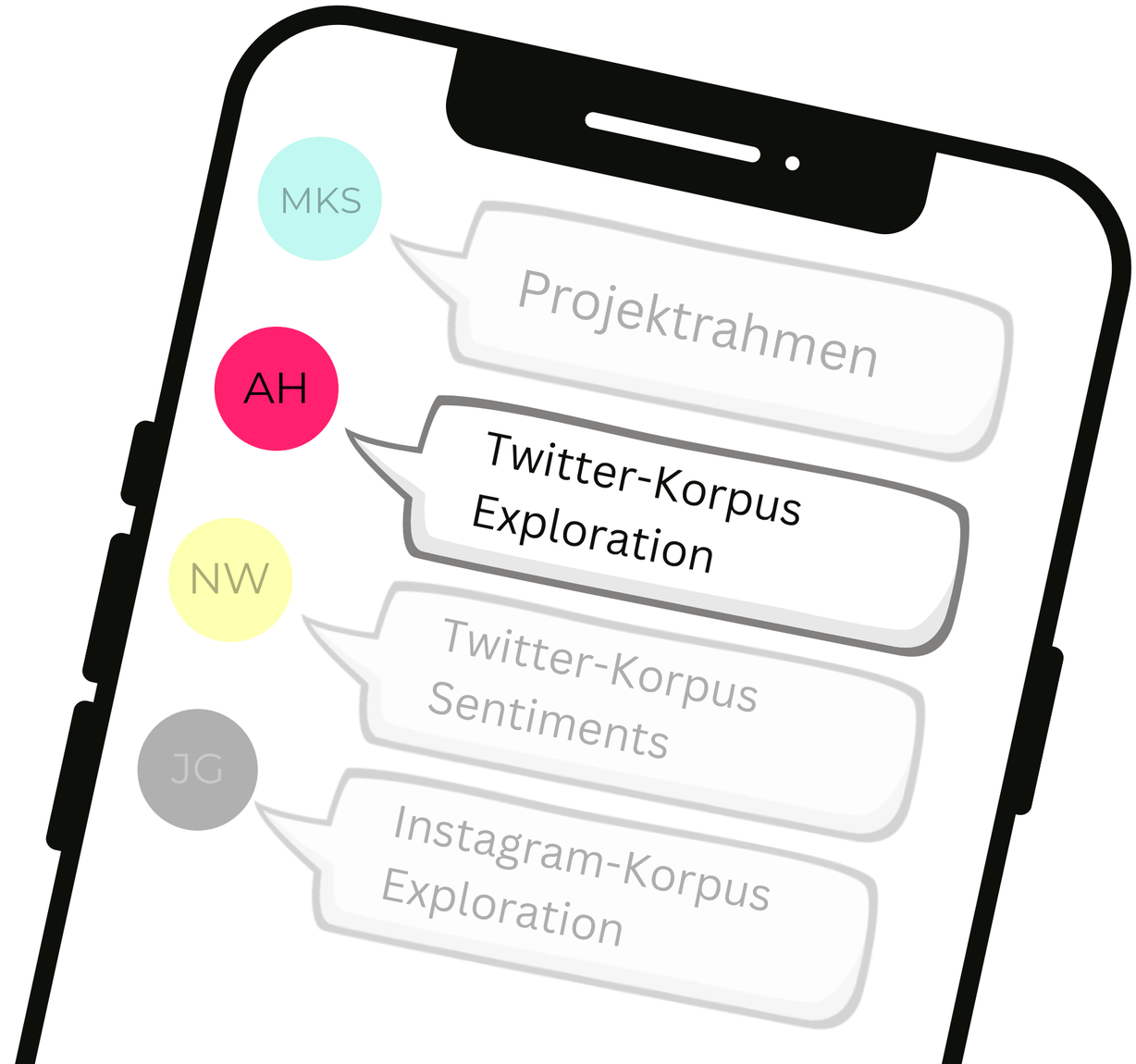
Dachprojekt: Wissenschaftlerinnen in die Medien

Projektseminar: Wissenschaftlerinnen in den Medien

- 2 Korpora
  - Twitter: Scraping (damals) für Zeitraum 14.12.2022-13.2.2023 #Forschung #PhDLife
  - Instagram: Scraping (jetzt) für Zeitraum 14.12.2022-13.2.2023 #Forschung #PhDLife
- Unterschiedliche Textmengen und -formen, unterschiedliche Medien
- Eine Methodenpipeline für Textdaten +Exploration der Bilddaten
  - Preprocessing
    - Problematische characters entfernt / für genderclassification Emojis entfernt
    - Aufteilung nach Sprache (deutsch, englisch)
  - Methoden
    - Wortfrequenzanalyse
    - Genderklassifikation
    - Dependency-Parsing
    - Sentiment Analysis



Teil 2


MKS
Projektrahmen
AH
Twitter-Korpus Exploration
NW
Twitter-Korpus Sentiments
JG
Instagram-Korpus Exploration

Twitter Korpus

Erhebungszeitraum: 14.12.2022 bis 13.02.2023
--> #phdlife #Forschung

	Tweets			Bios
Anzahl Tweets/ Bios	145.690			67.918
	Tweet_text	Retweet_text	Zitat_text	Bios
Anzahl Tweets/ Bios	40.093	94.733	23.683	67.918
Anzahl Wörter	1.400.579	2.515.497	845.866	847.006
Durchschnittliche Anzahl Wörter	9,61	17,26	5,81	12,47
Anzahl Zeichen	9.970.875	20.253.633	5.552.395	6.351.410
Durchschnittliche Anzahl Zeichen	68,44	139,02	38,11	93,52
Anzahl deutsch	81.452			17.848
Anzahl englisch	64.124			31.716

Twitter – Most frequent words

30 häufigste Wörter im Text der Tweets

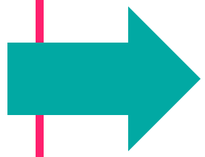
30 häufigste Wörter im Text der Bios

Twitter - Genderklassifikation

StanfordNER Gender-Classifier

von Mareike Schumacher

--> trainiert mit deutsch-sprachiger Literatur aus dem 18.-21. Jahrhundert



Liste Wissenschaftlerinnen

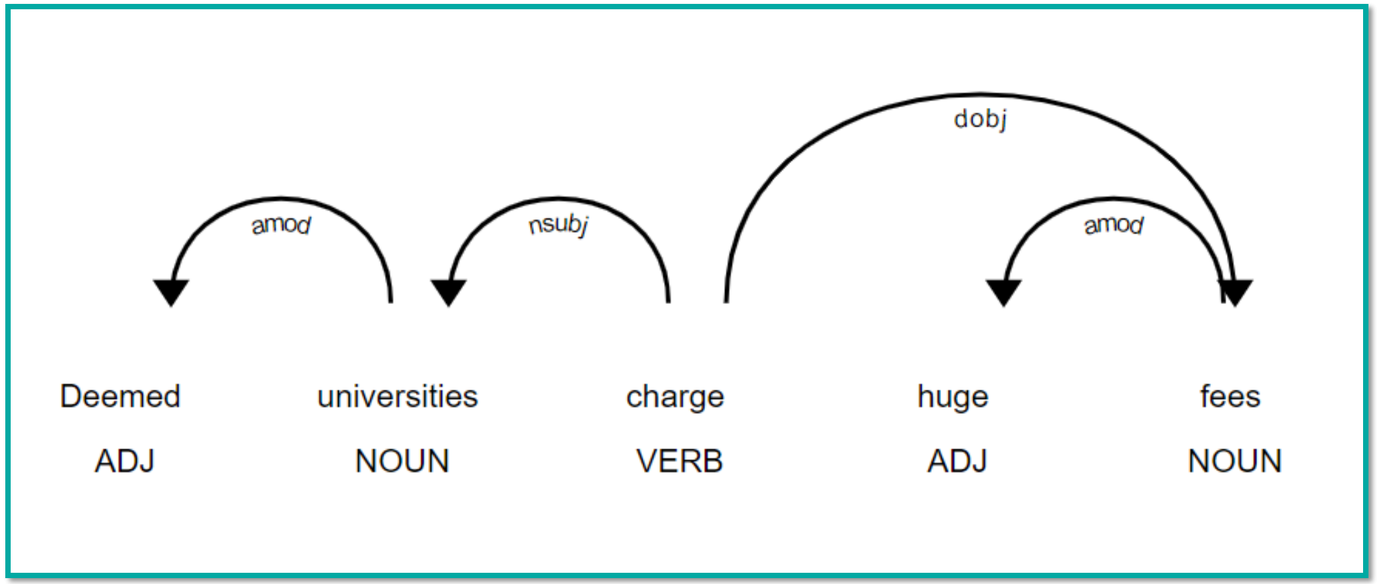
Evolutionsbiologin
Expertin
Fischbiologin
Fischforscherin
Forscherin
Forschungsdirektorin
Forschungsgruppenleiterin
Forschungsreferentin
Frau Prof. Dr.
Frau Professor
Frau Professorin
Gastprofessorin
Gastroenterologin
Gastwissenschaftlerin
Chefchirurgin
Gendermedizinerin
Genetikerin
Geophysikerin
Germanistin
Geschlechterforscherin

Liste Wissenschaftler

Informatiker
Ing.
Ingenieur
Ingenieurswissenschaftler
Institutsleiter
Jungwissenschaftler
Junior Professor
Junior Researcher
Kirchenhistoriker
Konfliktforscher
Kriminologe
Literaturwissenschaftler
Magister
Magister Artium German Philology
Master Student
Nachhaltigkeitswissenschaftler
Naturwissenschaftler
Patentingenieur
Pathologe

Dependency Parsing

Dependency Parsing is the process to analyze the grammatical structure in a sentence and find out related words as well as the type of the relationship between them.

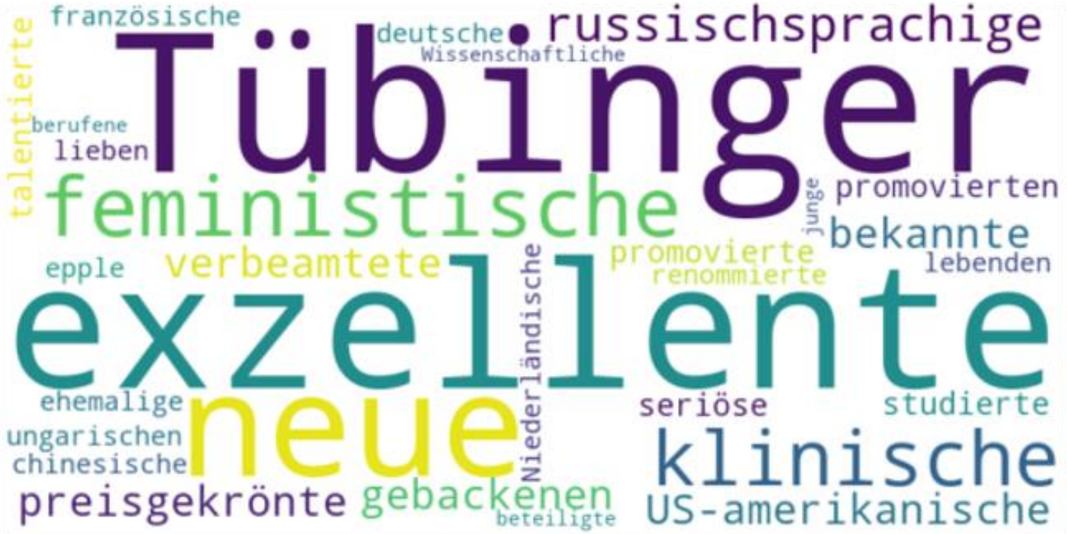


Quelle: https://towardsdatascience.com/natural-language-processing-dependency-parsing-cf094bbbe3f7 (05.12.2023)

Wir interessieren uns dafür, welche Adjektive jeweils mit den Begriffen der Listen der Wissenschaftler:innen vorkommen.

Twitter – Dependency Parsing

Im Text aller Tweets:



Adjektive im Zusammenhang mit Wissenschaftlerinnen

Adjektive im Zusammenhang mit Wissenschaftlern

Twitter – Dependency Parsing

Im Text aller Tweets:

Adjektiv	Häufigkeit
Tübinger	25
exzellente	20
neue	13
feministische	6
klinische	5
russischsprachige	3
verbeamtete	2
gebackenen	2
preisgekrönte	2
bekannte	2

Adjektive im Zusammenhang mit Wissenschaftlerinnen

Adjektiv	Häufigkeit
schwedischen	205
meistzitierten	50
zara	42
einzelnen	40
Schwedischer	29
deutsche	19
rer.	13
Geflohene	13
ukrainische	13
Deutscher	11

Adjektive im Zusammenhang mit Wissenschaftlern

Twitter – Dependency Parsing

Im Text der Tweets (ohne Retweets und zitierte Tweets):

Adjektive im Zusammenhang mit Wissenschaftlerinnen

Adjektive im Zusammenhang mit Wissenschaftlern

Twitter – Dependency Parsing

Im Text der Tweets (ohne Retweets und zitierte Tweets):

Adjektiv	Häufigkeit
neue	4
gebackenen	2
junge	1
ungarische	1
talentierte	1
studierte	1
seriöse	1
renommierte	1
promovierten	1
promovierte	1

Adjektive im Zusammenhang mit Wissenschaftlerinnen

Adjektiv	Häufigkeit
Schwedischer	15
deutsche	7
Leipziger	6
echte	4
praktizierender	4
Geflohene	4
ukrainische	4
junger	3
Dresdner	3
sogenannten	3

Adjektive im Zusammenhang mit Wissenschaftlern

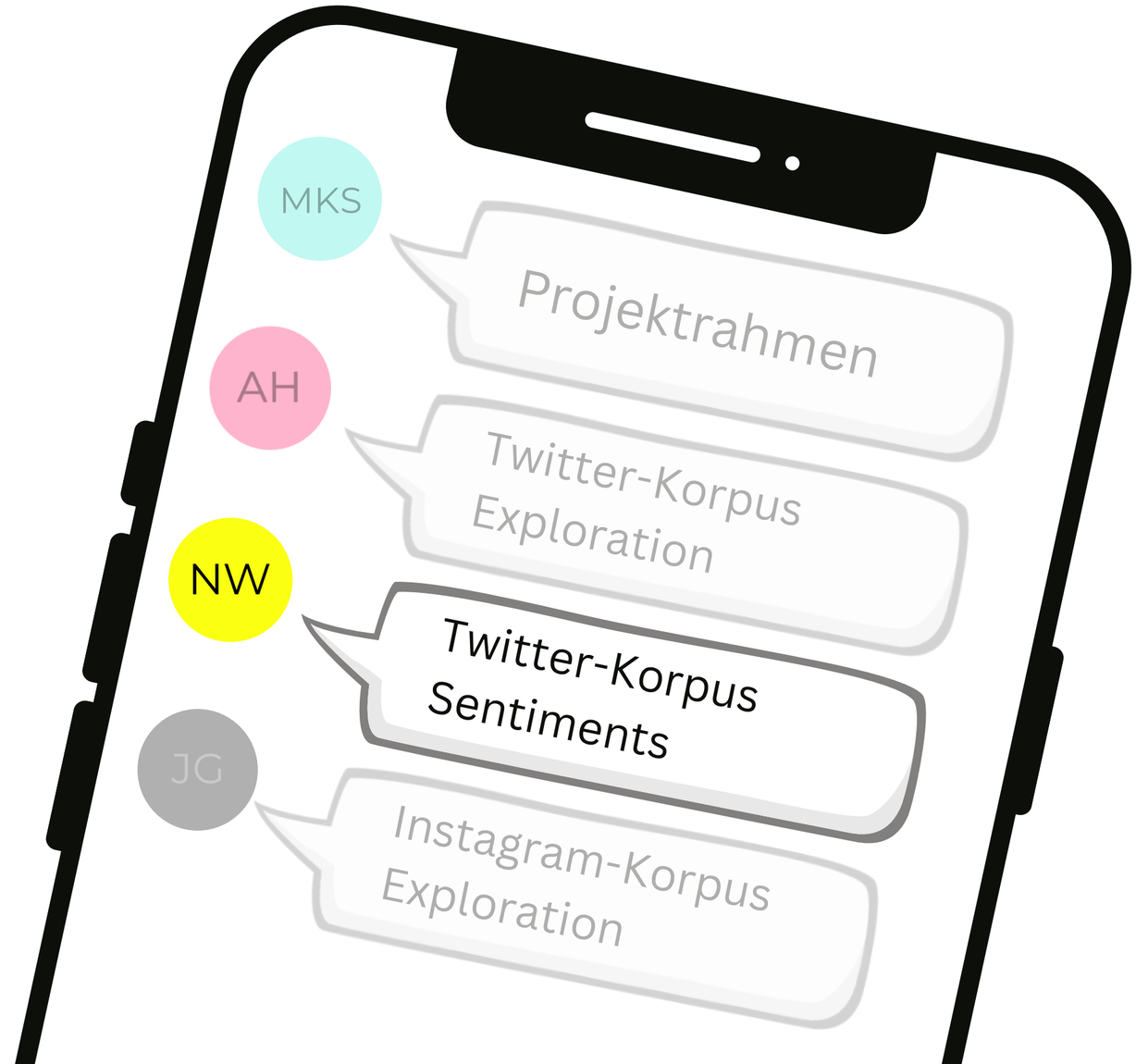
Twitter – Dependency Parsing

Im Text der Bios:

Adjektive im Zusammenhang mit Wissenschaftlerinnen

Adjektive im Zusammenhang mit Wissenschaftlern

Teil 3


MKS
Projektrahmen
AH
Twitter-Korpus
Exploration
NW
Twitter-Korpus
Sentiments
JG
Instagram-Korpus
Exploration

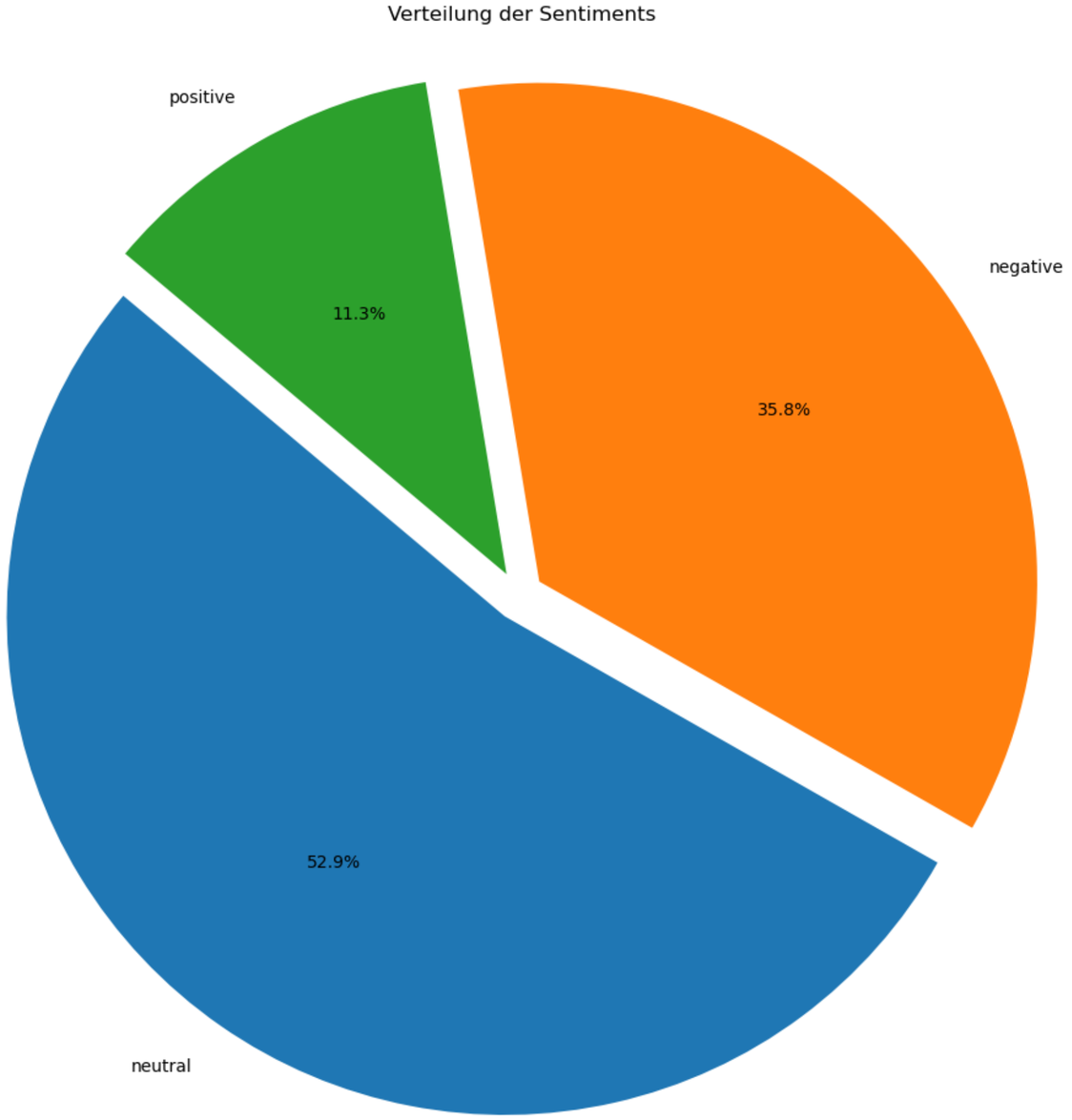
Twitter – Sentiments

Wie emotional sind Tweets zu gesuchten Themen? Wie untersuchen wir das Stimmungsbild?

Schritt 1:

Sentiment Analyse:

- Bereinigung der Texte
- Entfernen von Stopwörtern
- Machine Learning Prozess zur Klassifizierung der Texte
- Berechnete Werte Klassen zuordnen und Labeln der Klassen in negativ, neutral und positiv


Verteilung der Sentiments
positive
11.3%
negative
35.8%
neutral
52.9%

Twitter - Sentiments

	tweet_text	sentiment
1	Es scheint so weit weg, so surreal, aber es fehlt (nicht nur) an ärztlicher Versorgung für MEcfs LongCovid und PostVac Betroffene! Weil es an Aufklärung, Forschung und Unterstützung fehlt!Handeln Sie _Lauterbach und . JETZT!MEAwarenesshour	negative
4	@_donalphonso Müsste halt viel in Forschung gesteckt werden.Aber why not?Nur in DE wollen wir lieber Degrowth, weil wir so eine romantische Vorstellung d deindustrielisierten Landes haben.Man stelle sich vor, 80% sind reine Selbstversorger mit kleinem Acker, Rest ist • Politiker o beim ÖRR	negative
5	Seit einiger Zeit bringt Forschung und Wissenschaft nur Unheil, Krankheit, Verschmutzung und einigen wenigen unvorstellbaren Reichtum.Wenn Forschung und Wissenschaft wieder in alle Richtungen erlaubt wird könnte es wieder Nutzen bringen.	negative
7	Habe letzte Woche gelesen, dass Forscher jetzt festgestellt haben, dass das Einatmen von Alkoholdämpfen eine ähnlich entkrampfende Wirkung auf asthma-krampfende Lungen hat, wie hochpotente sofortinhalatoren.Jetzt frag ich mich, ob in der Forschung da keine Asthmatiker arbeiten.	negative
11	Hoffentlich wandern sie nicht aus, da in vielen Bereichen kein Geld zur Verfügung gestellt wird. Forschung zu Atomenergie sind schon lange in D gekappt. Wir hinken damit wissenschaftlich hinterher. Entwicklungen werden im Ausland gemacht.	negative
13	Zu lange ist es her, dass ich unsere Freundin nicht mehr gesehen habe, nicht mehr Klavier spielen und singen hörte, nicht mehr mit ihr sprach und Geburtstag feierte wir brauchen Aufklärung, Forschung und Versorgung. Es muss unbedingt endlich was vorwärts gehen!!MEawarenesshour	negative

Twitter - Sentiments

	tweet_text	sentiment
0	KleinschnitzRausschmiss sofort !	neutral
2	Hier in der Familie hat jeder Erfahrung in der Forschung:)) Wir sind allerdings auch nicht bei der FDP.	neutral
3	_i Wenn sich Politiker zur Forschung äußern ist sachdienlich wenn man sich vorher über die Vita informiert. Dann kann man den Inhalt vieler Beiträge besser einordnen.	neutral
8	_m_k Deshalb ist die Bildung und Forschung so wichtig.Es sollte doch chemisch möglich sein, Atome so zu spalten, dass es bei der Spaltung neben der freigesetzten Energie keinerlei radioaktive Rückstände gibt bzw. einen Stoff zu finden, der nicht radioaktiv ist. 😉	neutral
9	_SM Kernspaltung ist nicht sonderlich kompliziert; es gab sogar vor langer Zeit einen natürlichen Reaktor in Afrika. Fusion hingegen ist ein sehr dickes Brett:	neutral
10	Die Politikerin schätzt, dass wir in 10 Jahren den ersten dt. Kernfusionsreaktor sehen werden.Die Forschenden aus dem Projekt, dem die Fusion gelungen war, nennen 2035 als das frühstmögliche Jahr einer selbsttragenden Versuchsfusion.	neutral
12	__boehm Sie wissen aber schon, dass die Hersteller auf Ergebnisse intensiver Sars-CoV-1-Forschung zurückgreifen konnten und schon Jahre lang an mRNA- und Vertorimpfstoffen geforscht wurde?	neutral

Twitter - Sentiments

	tweet_text	sentiment
6	Ja, 10 Jahre sind ambitioniert, aber mit viel Fleiß und mächtig Unterstützung der Forschung machbar.Das würde viele Probleme lösen.Übrigens ist das keineswegs das Aus für andere Energiequellen.Also bitte kein Neid!	positive
25	Forschung ist geil. Die Ankündigung eines Fusionsreaktors innerhalb von zehn Jahren ist Lack gesoffen lächerlich. Wir scheitern ja schon an Bahnhöfen und Flughäfen.	positive
32	Wir haben in Deutschland die geniale Möglichkeit, unser Problem mit schwankender Stromversorgung durch Forschung und Technologie im Bereich Speichermedien zu lösen. Wahrscheinlich auch schneller als mit Fusionsreaktoren.erneuerbare Nachhaltigkeit	positive
37	Auch wieder schön zu sehen das der voraussichtliche Einsatz ab 2028 " geplant" ist. Also nach deutscher Zeit 2030-35. Bis dahin sind die Maschinen 10 Jahre alt, ein Lichtjahr in Forschung und Entwicklung. Was soll das alles? Milliarden für dann überholte Technik.	positive
44	Wenn die Aufträge so gut und schnell kommen, ist das ein sehr gutes Zeichen. Ich habe mir auch eine Existenz als academic Freelancer erarbeitet (aber nicht direkt in der Forschung) und das hat ein paar jahre gedauert.	positive
47	_derBlues Unser ganzes modernes Leben, einschl. der Möglichkeit doofe Kommentare online abgeben zu können verdanken wir der militärischen Forschung.	positive
68	Interessante Forschung, und das eröffnet ganz neue Möglichkeiten bei der Polizeikontrolle: „Herr Wachtmeister, ich bin Asthmatiker, ich habe nur inhaliert.“ 😂	positive

Twitter – Sentiments

Einschränkungen:

- Ironie, Sarkasmus, Zynismus
- Englischsprachige Modelle oftmals genauer
- Balance bei Texten mit negativen und positiven Sätzen → werden einige Male als neutral gewertet

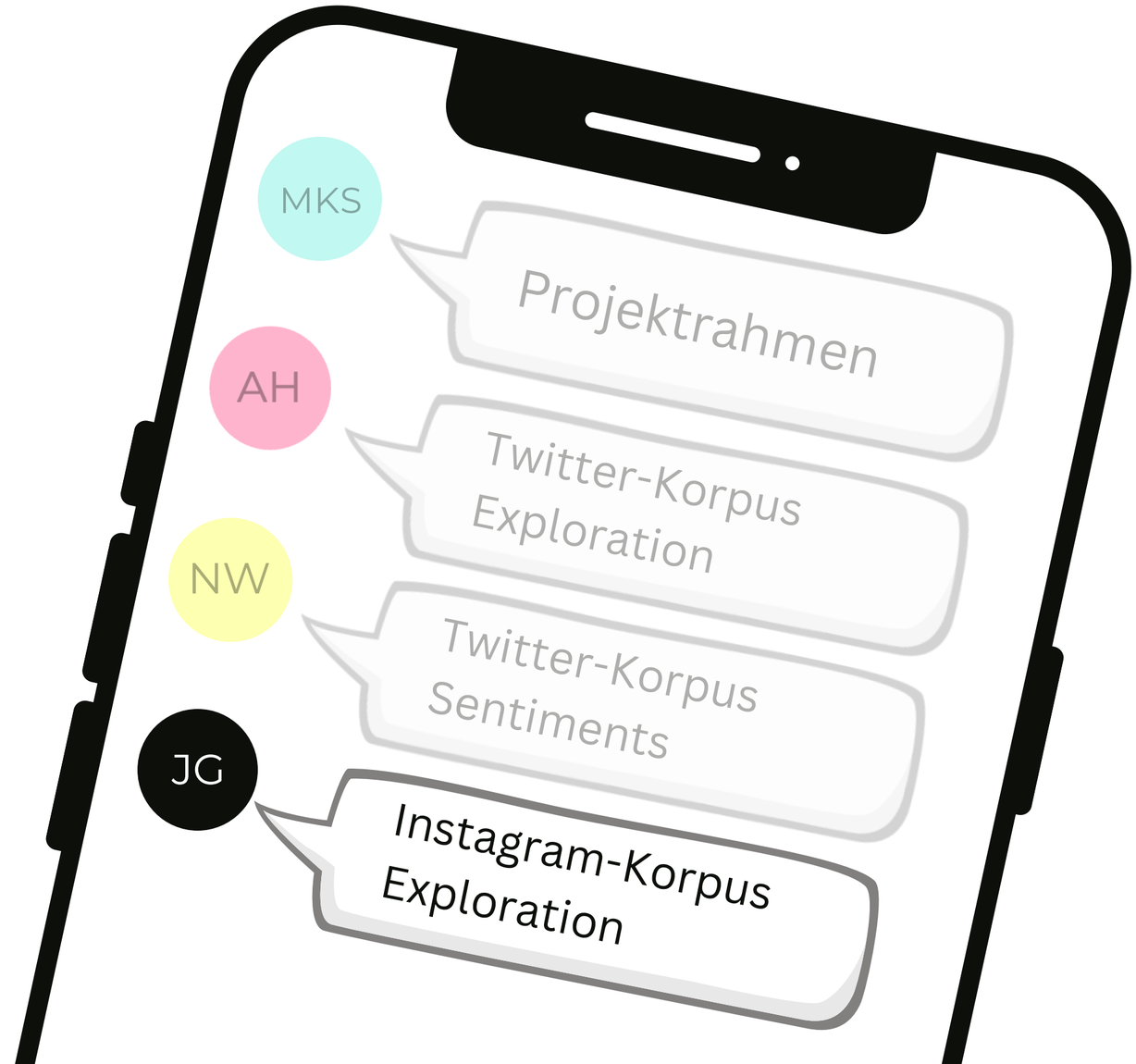
Möglichkeiten:

- Wie stellen sich Wissenschaftler:Innen dar?
- Ist der Grundton tendenziell negativ oder positiv?
- Gibt es genderspezifische Unterschiede bzw. Auffälligkeiten?
- Sind Tweets mit häufig genutzten Keywords negativ, neutral oder positiv?

Ausblick:

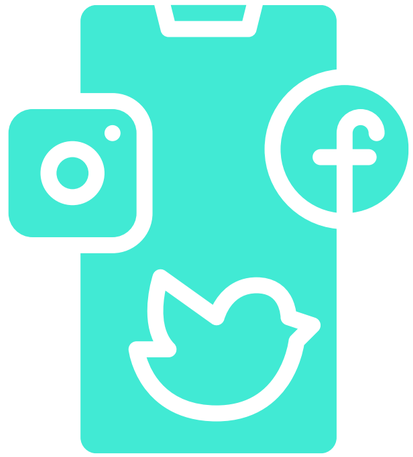
- Feintuning Sentiments
- Emotion Detection (bisher keine geeigneten Modelle für Deutsch, Eigenentwicklung sehr arbeitsintensiv)

Teil 4


MKS
Projektrahmen
AH
Twitter-Korpus
Exploration
NW
Twitter-Korpus
Sentiments
JG
Instagram-Korpus
Exploration

Instagram Daten

- Zeitrahmen: 14.12.22 - 13.02.23 (wie Twitter Daten)
- Geteilt in Englische (736) und Deutsche Posts (1375)
- Keine Reels; bei Karussell nur Bild 1
- Tool: CrowdTangle



CrowdTangle

#forschung, #phdlife

Share Save

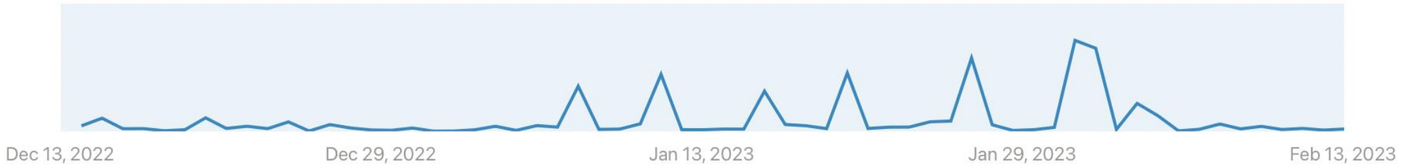
Instagram | Dec 13, 2022 - Feb 13, 2023 | Post Type | Accounts | German only | Branded Content | Verified Status | Lists | Meme Search

Due to a change in Instagram policy, all videos under 15 minutes are now classified as reels and currently not supported in CrowdTangle and the CrowdTangle API. All Instagram video posts that are shorter than 15 minutes and were created after July 31st, 2022 will not be available in search results, live displays or email digests. At this time, we are exploring how to address this issue, and will share more details when we have them.

Searches for hashtags, handles, and cashtags will also match their non-tag counterpart in image text.

INTERACTIONS 1,443,288

POSTS 1,376

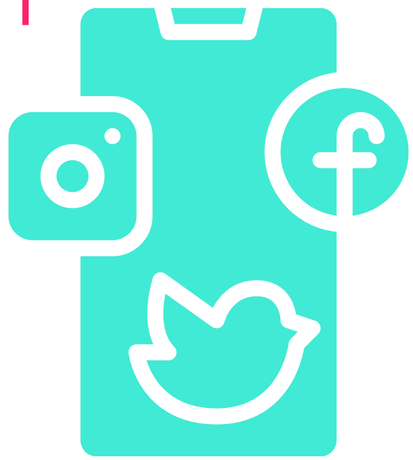


Showing **25** of **1,376** public posts from Dec 13, 2022 11:59 PM to Feb 13, 2023 11:59 PM

SORT BY Total Interactions

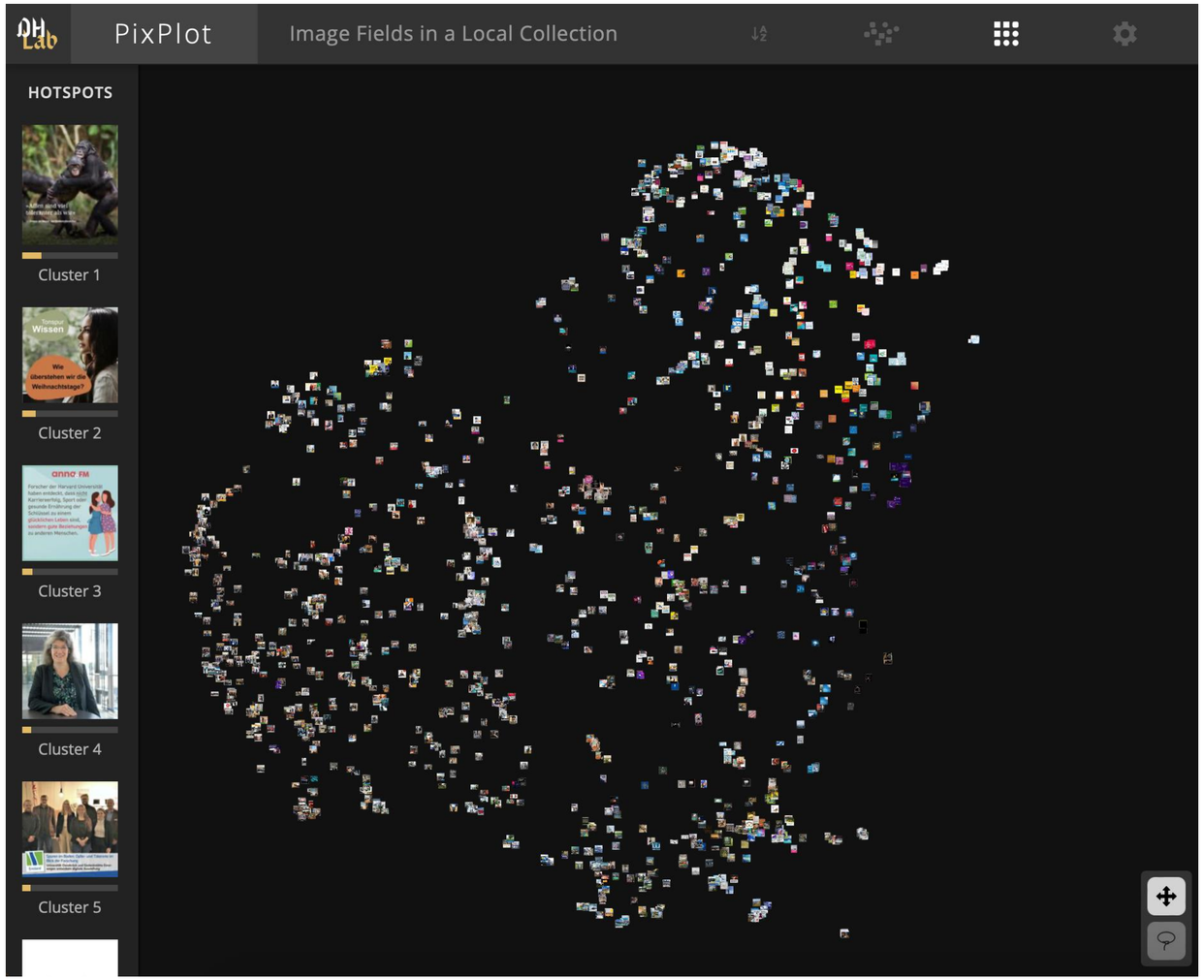
Instagram Bilddaten

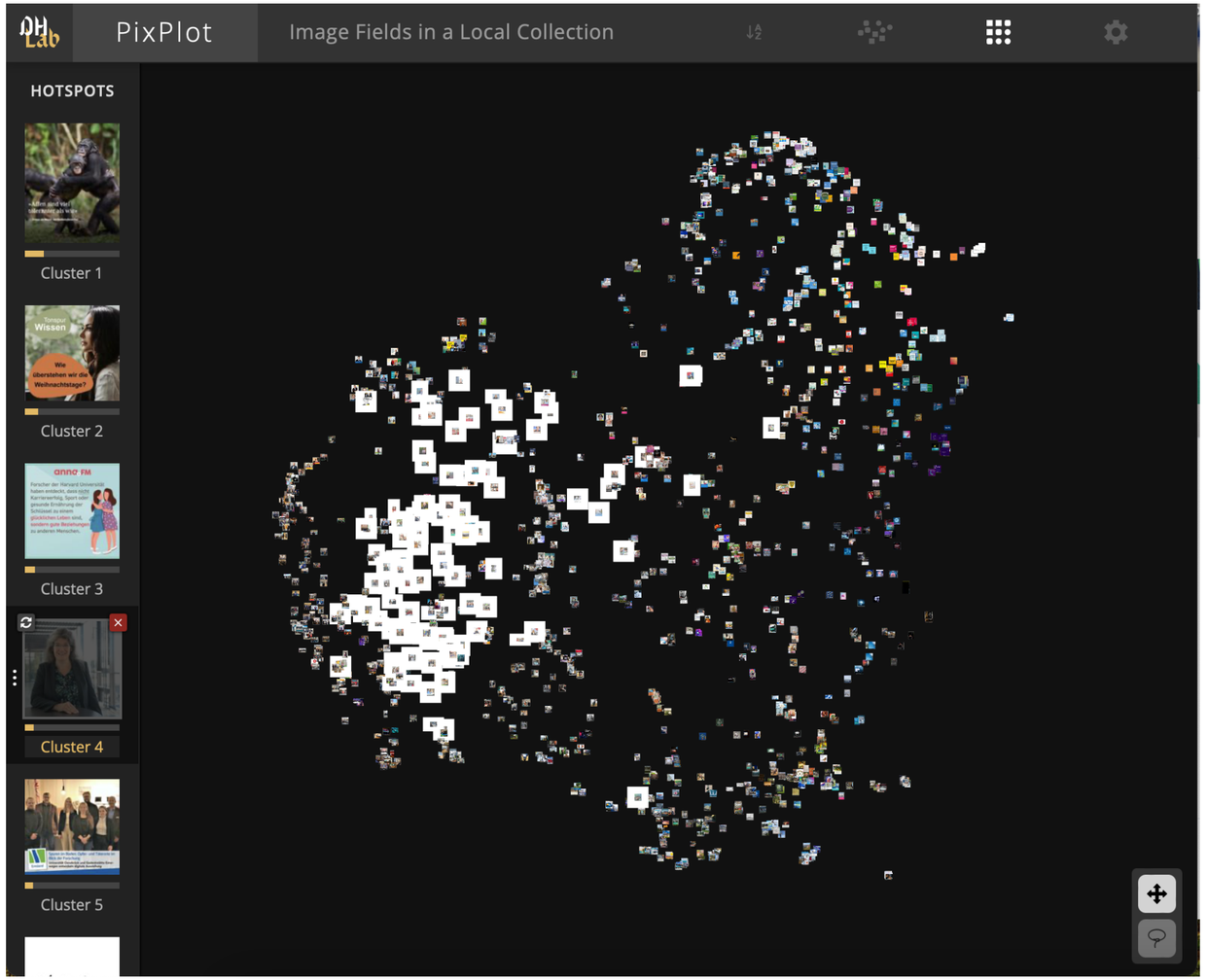
- Pro Post *ein* Bild (auch bei Video und Karussell)
- Clustering mit *YaleDHLab's* ***Pixplot***
- HDBSCAN bzw. KMeans Algorithm
- Gruppieren der Bilder in Kategorien
- Ziel: Mit Metadaten kombinieren: z.B. Clustern nach Gender



(https://github.com/YaleDHLab/pix-plot)

PixPlot


PixPlot
Image Fields in a Local Collection
HOTSPOTS
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5


PixPlot
Image Fields in a Local Collection
HOTSPOTS
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

Instagram – Textdaten

Description

Image Text

Erhebungszeitraum:
14.12.2022 bis 13.02.2023

--> Entsprechend Twitter-Daten
--> #phdlife #Forschung

	deutsch	englisch
Posts	1.370	731

Instagram – deutsche Textdaten

	Posts	
Anzahl	1.370	
	Description	Image Text
Anzahl	1.369	429
Anzahl Wörter	143.770	10.522
Durchschnittl. Anzahl Wörter	104,9	7,7
Anzahl Zeichen	1.136.294	75.489
Durchschnittl. Anzahl Zeichen	829,4	55,1

Instagram - Most frequent words

50 häufigste Wörter in Description

30 häufigste Wörter in Image Text

Instagram – Most frequent words

Wort	Häufigkeit
forschung	1627
wissenschaft	587
mehr	414
link	408
bio	390
unserer	312
dr	241
science	218
neue	213
of	202

Häufigste Wörter in Description

Wort	Häufigkeit
forschung	141
bildung	41
wissenschaft	29
mehr	27
uhr	25
neue	23
dr	23
2023	23
wissen	22
future	22

Häufigste Wörter in Image Text

Fragen

- CrowdTangle: Automatisiertes Scrollen
- Disambiguierung Maskulinum / generisches Maskulinum
- Wie kommen wir noch tiefer in die Daten hinein?
- Domänenspezifischere Genderklassifikation?
- Weitere Bilddatenanalyse
- Geeignetes Modell für Emotion Detection?