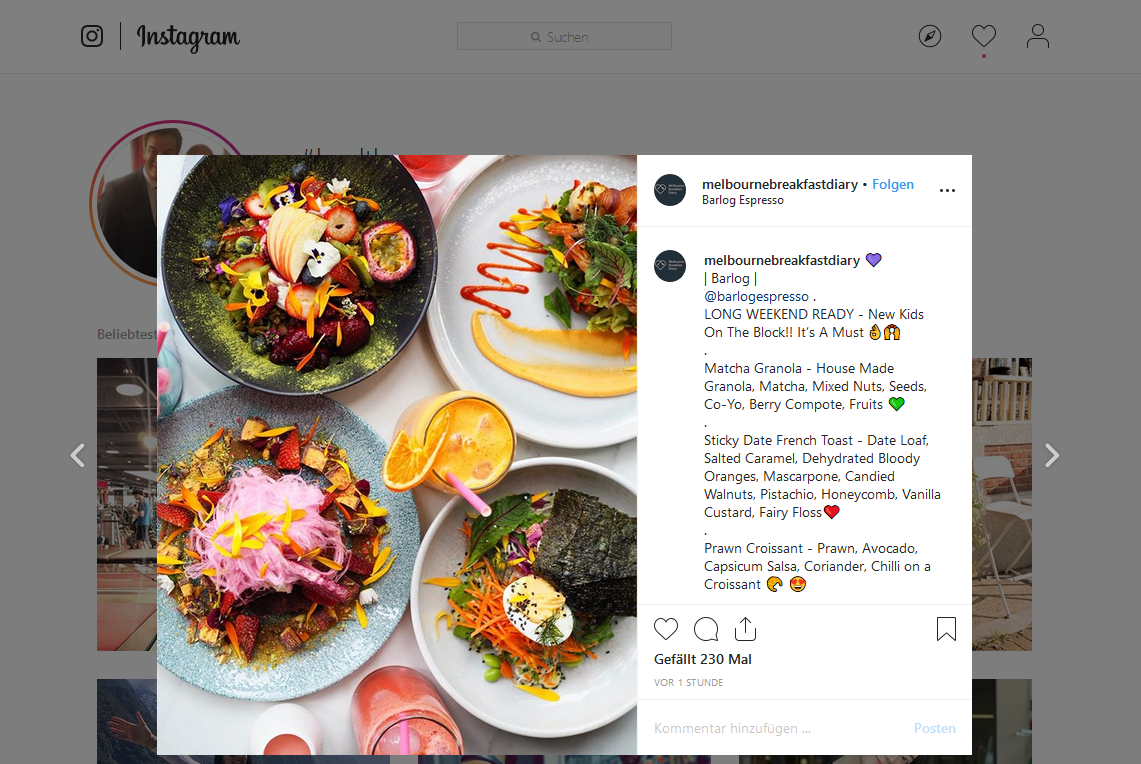
Leeching Instagram

Mit Hilfe eines einfachen Programms wurde Instagram-Content heruntergeladen, indem der Text mittels copy&paste direkt aus dem Browser kopiert und das Bild als Screenshot gespeichert wurde.

Das Programm startet mit einem dargestellten Bild einer Gallerie in Instagram und klickt alle 2 Sekunden auf den ‚weiter‘-Pfeil rechts des Bildes, kopiert allen Text in eine Textdatei und den Screenshot in ein PNG-Bild.

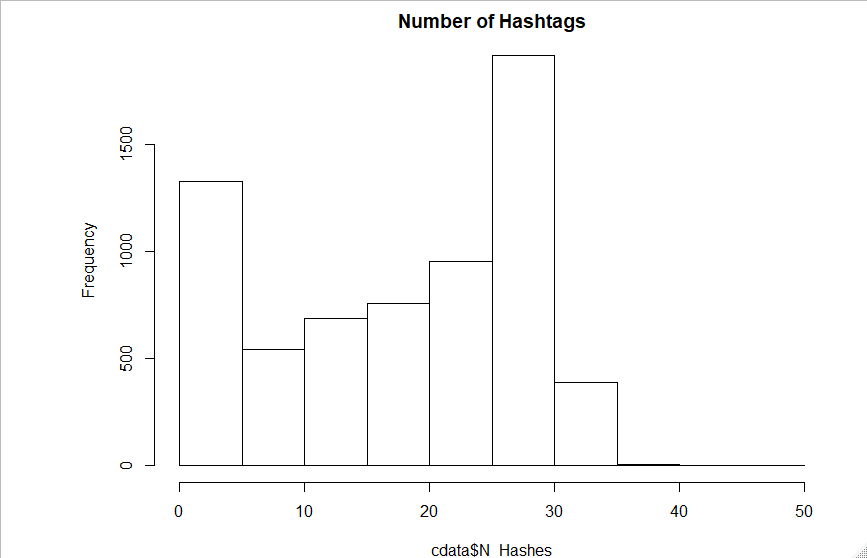


Der kopierte Bildausschnitt enthält das Bild und oben und unten gegebenenfalls einen grauen Rand, da die Höhe der Bilder variieren kann (die Breite ist durch die Website starr). Der Text enthält sowohl den Text des Beitrags als auch den Text aller Kommentare. Die Emojs sind ebenfalls gespeichert (UTF-8-Codierung).

# Probelauf

In einem ersten Probelauf wurde die Gallerie ‚Beliebteste Beiträge‘ zu den Hashtags #health (N=2000) und #gesund (N=4500) heruntergeladen.

Die Anzahl Hashtags bewegt sich pro Post zwischen 0 und 30. Mehr sind selten.



Die geringe Anzahl von Worten pro Dokument (Hashtags pro Post) machen Text-Clustering schwierig. Eine LDA wird bei diesen Daten kein robustes Ergebnis liefern. Besser wäre eine Contingency-Analyse der beliebtesten Hashtags. Hier läuft man allerdings Gefahr, dass zu viele Verbindungen zwischen den Tags bestehen und man viel Rechenleistung für die Clusterbildung benötigt.

Insgesamt wurden in diesen 6500 Posts 25‘473 Hashtags gefunden. Von diesen traten 16‘000 nur einmal auf und sind damit für eine Analyse nicht interessant.

Die beliebtesten Hashtags kommen indes sehr häufig vor und sind mit vielen anderen Tags gemeinsam in Posts:



***Note: Occurrence: Anzahl Posts, in denen dieser Hashtag aufgetreten ist / Co-Occurring: Hashtags, mit welchen dieser zusammen aufgetreten ist***

# Daten

Die Sammlung generiert ein bereinigtes Korpus, das für jeden Beitrag den Namen des Users, die Anzahl Hashtags im Post und die Liste aller Hashtags im Post und in den Kommentaren getrennt beinhaltet.

Der Rest des Texts ist entfernt, um wirklich nur eine Hashtag-Anayse zu machen. Die Texte sind aber gespeichert und können bei Bedarf nachgeschlagen werden (Dateiname im Korpus).

Zudem generiert die Sammlung einen Ordner mit Bildern mit identischer Nummerierung. Die Bilder sind in PNG-Format, was evtl. nicht optimal ist, wenn man Speicherplatz sparen möchte. Die Möglichkeit, sie in JPEG zu komprimieren, besteht aber.

Die Bilder befinden sich im Moment auf meinem Switch-Drive:

<https://drive.switch.ch/index.php/s/I33ShijkJFcUfez>

# Mögliches Problem beim Herunterladen:

Das Vorgehen, bei dem einfach eine Gallerie durchgeklickt wird, ist instabil. Passiert während der Sammlung irgendwas (jemand stösst an die Maus, Instagram kann gerade nicht geladen werden, ein Bild wird breiter dargestellt...), bricht die Sammlung ab und kann nicht an derselben Stelle wieder aufgenommen werden. Da nicht mit eindeutigen Post-Identifiern gearbeitet wird und sich die Gallerien dynamisch füllen und bei jede Aufruf verändern dürften, kann eine Datensammlung nicht reproduziert werden.

Dies ist im Probedurchlauf beim Hashtag #health passiert, wo nach ca. 2030 Bildern die Gallerie geschlossen wurde (nicht nachvollziehbar wieso). Zwar ist es möglich, das letzte Bild zu sehen, das noch gefunden wurde, aber man kann die Gallerie nicht mehr ab diesem Bild neu starten.

Das Gleiche würde auch beim Manuellen Sammeln passieren, sobald die Gallerie geschlossen wird.

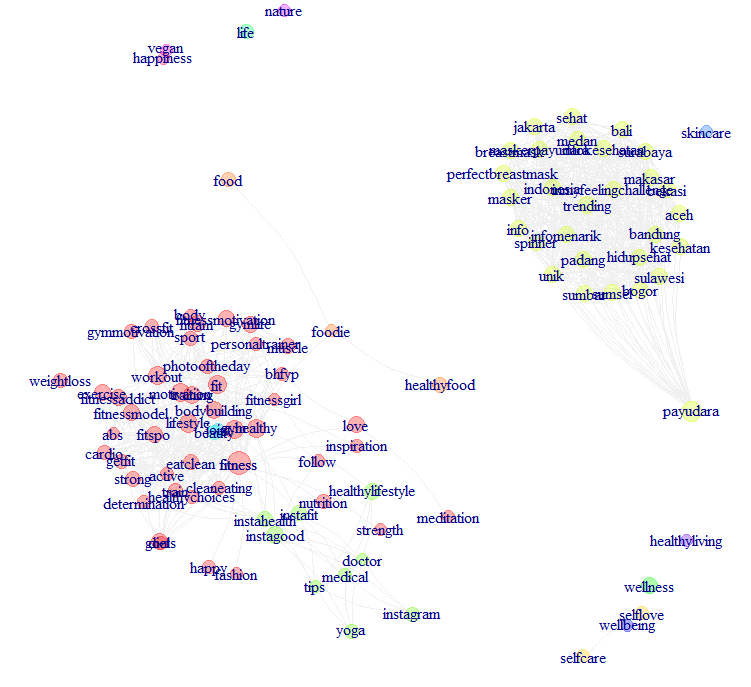
Man könnte diesem Problem begegnen, indem man immer die Gallerie mit ‚Neueste‘ verwendet. So kann man jeden Tag oder jede Woche die neuen Posts sammeln. Sobald man Duplikate mit der letzten Sammlung hat, kann man abbrechen (Duplikatanalyse ist relativ einfach).

Dass man alle mehrere Millionen Bilder auf diese Weise finden könnte, ist aber illusorisch.

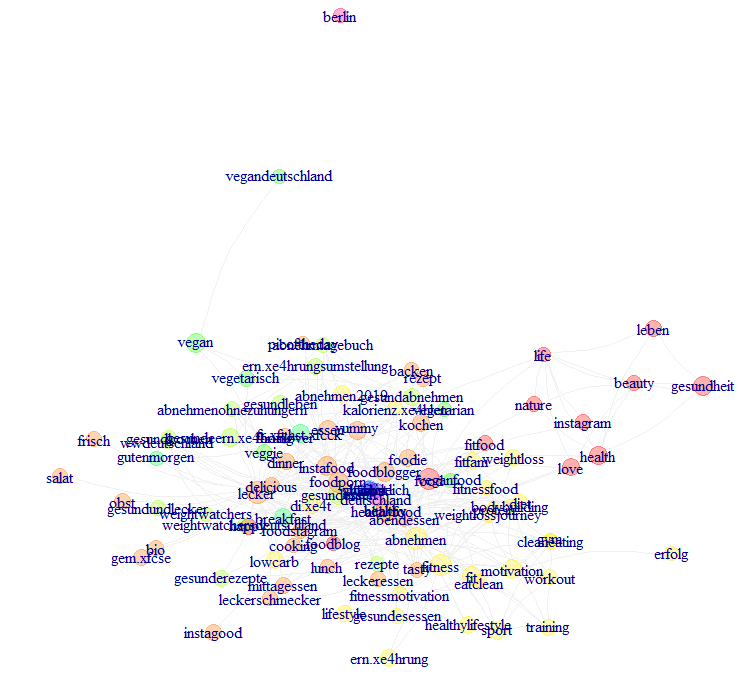
# Netzwerkanalyse

Für eine erste Analyse wurde das gemeinsame Auftreten der Hashtags quantifizert und in eine Netzwerkanalyse eingegeben. Das gemeinsame Auftreten zweier Tags ist die höhere bedingte Wahrscheinlichkeit, dass einer der Tags auftritt, wenn auch der andere auftritt. Das heisst, alle Tags haben eine Wahrscheinlichkeit von 1.0, mit Health aufzutreten. Deswegen wird dieser zentrale Knoten entfernt.

Alle bedingten Wahrscheinlichkeiten unter 0.3 wurden vernachlässigt, um einen interpretierbaren Plot zu erhalten. Eine Fast&Greedy Clusteranalyse des Netzwerks brachte beim Englischen 3 Cluster mit einigen Ausreissern. Eines der Cluster ist fitness und Ernährung, eines ist Medizin und Yoga und eines ist Payudara (Brustpflege aus Malaysia). Dazu gibt es einige Ausreisser, die zwar auch häufig mit #health vorkommen aber nicht häufig (mit anderen der 100 häufigsten Tags).



Im deutschen Ergebnis wurden alle Kanten mit einem Gewicht unter 0.2 (also weniger als 20% bedingte Wahrscheinlichkeit) gelöscht. Das Resultat ist etwas fragmentierter als bei der englischen Auswertung mit mehreren grossen Clustern: Es gibt ein Vegan/Vegetarisch, ein Leben/Beauty, ein Sport/Training, ein Abnehmen, und ein Rezept/Kochen Cluster. Daneben auch einige Ausreisser.



# Clustern

Unglücklicherweise sind die Netzwerke zu dicht, um sie sinnvoll zu clustern. Es ist eine recht überschaubare Zahl von Hashtags und sie sind alle mit allen verknüpft weil eine Reihe von Standard-Hashtags hinter jeden Post geklemmt werden.

LDA fällt flach, weil die Texte dennoch zu kurz sind und praktisch nur aus Stopwords (in den meisten Posts wiederkehrende Elemente) bestehen. Nur wenig Varianz würde eine LDA informieren und die kann gut zufällig sein. Eine robuste Lösung ist nicht zu erwarten.

Eine Cluster-Analyse über das dichotom gemessene Auftreten von Hashtags dürfte schwierig werden weil auch da zu viele immer gemeinsam auftreten.

Eine Möglichkeit wäre, die Hashtags zu gewichten und einem Tag, der zu Beginn einer Liste angeführt wird (also vom Verfasser als besonders wichtig wahrgenommen wird) mehr Gewicht zuzuschreiben als einem, der ganz am Ende noch als Nachgedanke und für die Reichweite hinzugefügt wird.

Dies habe ich getan, und eine Term-Document-Matrix erstellt, in der jeder Tag ein Gewicht hat. Der erste hat Gewicht 1.0, der letzte eine 0. Dazwischen sind sie linear bewertet. Dadurch entsteht eine TDM, die stark Zero-Inflated ist. Jedes Dokument und jeder Term ist schief verteilt.

Die TDM wurde mit dem HECANE-Verfahren geclustert, um störende Elemente automatisch auszuschliessen. Folgende 11 Cluster wurden als homogene Untercluster gefunden:

## Cluster #1 (radius=1.142190521980848): Weight Wachers

-------------

#gesundleben (N=152.56379092643175)

#gesundeern\xe4hrung (N=161.477721890576)

#ern\xe4hrungsumstellung (N=83.18133605595344)

#kalorienz\xe4hlen (N=61.06661927573356)

#yazio (N=36.10490703758692)

#abnehmblog (N=32.75419298077541)

#derspeckmussweg (N=30.930659902421834)

#fitwerden (N=34.26322082468179)

#abgerechnetwirdamstrand (N=26.04847777546888)

#weightwatchersdeutschland (N=53.4086604813008)

#weightwatchers (N=61.770045544339354)

#wwdeutschland (N=61.61521775604267)

#wwm\xe4dels (N=50.43745073805155)

#weightloss (N=138.51266408466756)

#weightlossjourney (N=51.39876439625895)

#abnehmen2019 (N=163.46901322730542)

#abnehmenohnezuhungern (N=64.69092623245537)

#abnehmtagebuch (N=88.46119644257335)

#gesundessen (N=465.8576396269303)

#gesundabnehmen (N=161.33309996896423)

#abnehmen (N=600.3602100590872)

#di\xe4t (N=182.7525828905818)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #2 (radius=1.1932398827719128): Essen

-------------

#fitfood (N=69.21790358721985)

#foodstagram (N=53.76844454037494)

#salad (N=52.64709469629079)

#vegetables (N=47.979707927703295)

#healthylifestyle (N=115.00690047183228)

#foodlover (N=73.12478360500846)

#abendessen (N=151.44867152151377)

#dinner (N=89.00247102526784)

#tasty (N=62.88891554564222)

#delicious (N=83.77560966897934)

#healthy (N=616.063227186012)

#healthyfood (N=458.31308025213474)

#foodie (N=115.40386430201498)

#yummy (N=145.68188156494648)

#lecker (N=597.144272103912)

#essen (N=306.4556927306526)

#instafood (N=220.85335966313693)

#food (N=454.9380226100147)

#foodporn (N=290.2564429299275)

#foodblogger (N=223.65730263812915)

#cleaneating (N=111.66480611290844)

#fooddiary (N=31.223741076857742)

#eatclean (N=70.17064618418725)

#foodpics (N=24.299088304271837)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #3 (radius=1.1058623124409084): Fitness

-------------

#body (N=29.378984380456732)

#fitfam (N=108.36461474853338)

#training (N=98.6594755329454)

#bodybuilding (N=39.92766481256829)

#krafttraining (N=27.973562019297795)

#muskelaufbau (N=35.09971281251886)

#muskeln (N=31.433626891997296)

#fit (N=284.35555246663574)

#sportmotivation (N=27.126336266484437)

#fitnessmotivation (N=75.11031023637827)

#fitnessgirl (N=34.414770454165385)

#motivation (N=203.8794624440612)

#fitness (N=415.8597741524464)

#sport (N=190.31641607883554)

#gym (N=86.77152963418378)

#workout (N=77.0417454084034)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #4 (radius=1.033028040443669): Vegan

-------------

#vegandeutschland (N=54.470226908731604)

#veganessen (N=33.223598530995645)

#plantbased (N=43.88663185107485)

#vegan (N=293.7508043131707)

#veganfood (N=55.06087494177907)

#veggie (N=68.16176681682069)

#vegetarisch (N=106.28100107998073)

#vegetarian (N=45.91346241742656)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #5 (radius=1.0686851244462694): Kochen

-------------

#gesundkochen (N=58.008821671038724)

#rezept (N=88.9659304719467)

#kochen (N=155.73807965161163)

#kochenmitliebe (N=31.30369776968733)

#kochenmachtspass (N=27.280176422158036)

#gutenappetit (N=28.4684616865764)

#leckeressen (N=72.99130749782945)

#leckerschmecker (N=65.88672570020691)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #6 (radius=1.0526449711944788): Frühstück

-------------

#m\xfcsli (N=27.50264285431403)

#haferflocken (N=26.190488247603763)

#fr\xfchst\xfcck (N=248.9305368405504)

#breakfast (N=165.60128352845905)

#gutenmorgen (N=67.66161075197584)

#goodmorning (N=33.336223072318134)

#porridge (N=36.46550972182725)

#healthybreakfast (N=34.91524784121999)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #7 (radius=1.1014906702361953): Obst

-------------

#garten (N=49.21156405898799)

#apfel (N=39.80014593435883)

#vitamine (N=56.16033285327908)

#smoothie (N=30.28762360494229)

#obst (N=102.95821511194305)

#fruit (N=33.42279489782873)

#fruits (N=34.00392803674711)

#fr\xfcchte (N=24.585804026727278)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #8 (radius=0.9720080627262431): Kürbissuppe

-------------

#suppe (N=27.946630619680608)

#k\xfcrbis (N=52.349052784607096)

#herbst (N=98.14606140408044)

#autumn (N=27.022155819751035)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #9 (radius=1.1144738081295777): Bio

-------------

#regional (N=28.604334357951654)

#nachhaltig (N=31.701607847803093)

#nachhaltigkeit (N=26.409663130195362)

#bio (N=92.8308045071912)

#organic (N=29.387944364799562)

#yoga (N=26.938315711581037)

#lifestyle (N=90.33869298023598)

#tee (N=31.230092808022263)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #10 (radius=1.0213421671278549): Deutschland

-------------

#deutschland (N=55.18937557014268)

#germany (N=36.62927240637416)

#berlin (N=36.958060569294204)

#urlaub (N=26.709415521162537)

#hamburg (N=33.51751870188536)

#m\xfcnchen (N=25.79287746452197)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

## Cluster #11 (radius=1.2497554420461154): Wasser

-------------

#gesundheit (N=245.34909228321945)

#leben (N=92.69917155753612)

#health (N=155.03451139229958)

#beauty (N=77.1022865179693)

#water (N=25.08915676615476)

#nature (N=87.12973489338346)

#life (N=61.790407903523494)

#wasser (N=55.72577840655803)

#\xf6sterreich (N=47.34280902943306)

#smile (N=56.592277423200336)

#umwelt (N=52.85928777314087)

#europe (N=32.41533689809553)

#environment (N=42.004926108374406)

#wasserexperten (N=35.9562397372742)

#qualit\xe4tswasser (N=24.51724137931034)

#love (N=90.26623239808859)

#happy (N=66.9062288656656)

Prototypische Posts:

|  |  |
| --- | --- |
|  |  |

Eine Korrelationsanalyse der Clusterzentren (über den Cosinus des Zwischenwinkels) zeigt, dass gerade Cluster 2 eine hohe Überlappung mit den meisten anderen Unterclustern hat. Es ist als ‘Essen’ auch ein Cluster, das viel mit Kochen, Vegan, Obst, etc. zu tun hat.

Positiv am Ergebnis ist, dass es dennoch einzelne Untercluster für einzelne Ernährungs-Typen gegeben hat. Es gibt nicht einfach ein ‘Essen’ und ein ‘Fitness’-Cluster, sondern eine feinere Auflösung.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Weight Watchers | Essen | Fitness | Vegan | Kochen | Frühstück | Obst | Kürbissuppe | Bio | Deutschland | Wasser |
| Weight Watchers | 1 | 0.262 | 0.2 | 0.128 | 0.103 | 0.114 | 0.042 | 0.051 | 0.032 | 0.022 | 0.029 |
| Essen | 0.262 | 1 | 0.193 | 0.263 | 0.251 | 0.218 | 0.134 | 0.123 | 0.095 | 0.087 | 0.046 |
| Fitness | 0.2 | 0.193 | 1 | 0.04 | 0.037 | 0.075 | 0.036 | 0.024 | 0.092 | 0.093 | 0.054 |
| Vegan | 0.128 | 0.263 | 0.04 | 1 | 0.126 | 0.09 | 0.09 | 0.06 | 0.129 | 0.039 | 0.018 |
| Kochen | 0.103 | 0.251 | 0.037 | 0.126 | 1 | 0.067 | 0.05 | 0.102 | 0.039 | 0.018 | 0.01 |
| Frühstück | 0.114 | 0.218 | 0.075 | 0.09 | 0.067 | 1 | 0.109 | 0.024 | 0.072 | 0.03 | 0.017 |
| Obst | 0.042 | 0.134 | 0.036 | 0.09 | 0.05 | 0.109 | 1 | 0.09 | 0.067 | 0.024 | 0.019 |
| Kürbissuppe | 0.051 | 0.123 | 0.024 | 0.06 | 0.102 | 0.024 | 0.09 | 1 | 0.057 | 0.022 | 0.018 |
| Bio | 0.032 | 0.095 | 0.092 | 0.129 | 0.039 | 0.072 | 0.067 | 0.057 | 1 | 0.087 | 0.039 |
| Deutschland | 0.022 | 0.087 | 0.093 | 0.039 | 0.018 | 0.03 | 0.024 | 0.022 | 0.087 | 1 | 0.024 |
| Wasser | 0.029 | 0.046 | 0.054 | 0.018 | 0.01 | 0.017 | 0.019 | 0.018 | 0.039 | 0.024 | 1 |