04.12.2020

Wikipedia - ein Blick in die Zahlen

Zusammenfassung

Wikipedia bietet enorme Mengen an Wissen an. Durch das offene Prinzip der Plattform entstehen ebenfalls Unmengen an Metadaten. In diese werfen wir hier einen Blick.

DATA NEWS SOCIAL RADIO NETWORK MONEY LIKE ONLINE SHARE PEOPLE CONTENT INTERNET INFORMATION COMMUNITY MOBILE SEARCH FLICKR MEDIA WEB GLOBAL FRIENDS POPULAR WORDPRESS COMMUNICATION YOUTUBE GOOGLE+ FACEBOOK LINKEDIN PINTEREST

Inhaltsverzeichnis

1 Einleitung

Wikipedia gehört zu den meist besuchten Internetseiten. Der deutschsprachige Ableger ist mit rund 2,5 Millionen Seiten die drittgrößte Version der Wikipedia [2]. Damit dürfte kaum ein:e Nutzer:in des Internets um diese Enzyklopädie herum kommen. Sie bittet einen enormen Umfang an Wissen und Informationen.

Die Wikipedia als Palttform verfolgt einen offenen Ansatz bei dem jede:r Nutzer:in Informationen beitragen darf. Die Offenheit erstreckt sich auch auf die angebotene API, diese bietet sehr viele Schnittstellen und Zugriff auf quasi den gesamten Bestand der Wikipedia. Diese Möglichkeit wird in dieser Arbeit genutzt um einen zufälligen Sub-Datensatz aus der Wikipedia zu extrahieren und einen Blick in Zahlen und Statistiken zu werfen. Interessante Fragestellungen, etwa nach der Verknüpfung der Nutzer:innen, stellen sich, aber auch Informationen über eine:n “Durchschnittsnutzer:in” zeigen wie die Wikipedia funktioniert.

2 Stand der Forschung

Grundsätzlich gibt es wenig Forschungsaktivität rund um die größe der Wikipedia und deren Nutzungs-Statistiken, dies lässt sich wohl damit begründen, dass jegliche Informationen über die eigene API zugänglich sind [2].

McMahon et al. beschreiben die Abhängigkeit zwischen Wikipedia und Google, zu deren Geschäftsbereich die weltgrößte Suchmaschine gehört [3]. Diese Verknüpfung zeigt den Stellenwert der Wikipedia auf. Sie ist für viele Nutzer:innen erste Anlaufstelle im Internet auf der Suche nach Informationen. So werden Auszüge aus der Wikipedia auch direkt in den Suchergebnissen bei Google dargestellt.

In einer Arbeit von Truong wird eine soziale Netzwerkanalyse der gesamten Wikipedia vorgenommen [6]. Zeigt sich bei der Netzwerksortierung nach den bearbeiteten Sprachen noch

eine Gruppenbildung, so ist diese bei Betrachtung der Artikel nicht mehr gegeben.

Es lassen sich zahlreiche semantische Analysen der Wikipedia finden, diese werden hier jedoch mangels Bezug zur Arbeit außenvor gelassen.

3 Methodik

3.1 Datenerfassung

Die in dieser Arbeit verwendeten Daten werden direkt von der deutschen Wikipedia bezogen. Diese stellt hierfür eine frei zugängliche API zur Verfügung. Die Bibliothek “Wikipedia” für Python wählt einen zufälligen Artikel aus der Wikipedia aus [4]. Dessen 500 letzte Revisionen, sofern verfügbar, werden dann direkt über die JSON-API der Wikipedia abgefragt.

In einer MySQL-Datenbank werden diese Informationen dann gespeichert. Hierzu gehören alle Revisionen, die Seiten und die damit verbundenen Kategorien, sowie die Nutzer:innen die die Revisionen vorgenommen haben.

Der gewonnene Datensatz beinhaltet 2.525 Seiten mit etwa 450.000 Revisionen von 26.000 Nutzer:innen. Diese Seiten sind mit knapp 9.000 Kategorien verknüpft.

3.2 Aufbereitung und Visualisierung

Für die Aufbereitung der Daten kommen Python-Skripte zum Einsatz, diese bereiten die Daten direkt aus der Datenbank auf und visualisieren diese über die “matplotlib” Bibliothek [5]. Für die Visualisierung von Netzwerken kommt darüber hinaus die Graphen-Software “Gephi” zum Einsatz [1].

4 Auswertung

4.1 Durchschnittswerte

Die gesammelten Revisionen beginnen Anfang 2002, dies passt zur Geschichte der Wikipedia, die 2001 begann [7]. Die Zahl der Artikel, oder Seiten, steigt nach den eigenen Angaben der deutschen Wikipedia seitdem linear an. Dieser grob lineare Trend findet sich auch in der Historie der gesammelten Revisionen, wo jedoch von Mitte 2003 bis 2007 ein etwas steilerer Anstieg zu sehen ist (Abb. 2). Abweichungen zum Artikelwachstum ergeben sich durch die lediglich ausschnittsweise Betrachtung und der Differenz von Artikeln und Revisionen. Mit einem R^2 Wert von 0,87 bestätigt sich der annähernd lineare Verlauf.

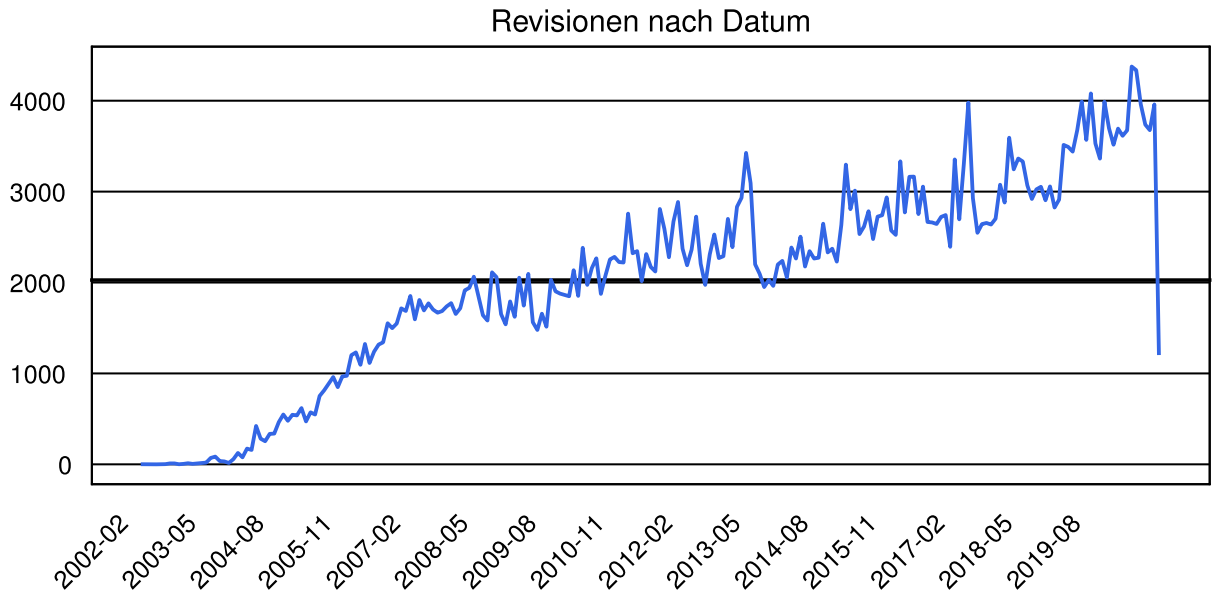

Abbildung 1: Revisionen nach Datum

Im Durchschnitt beinhaltet der Datensatz 176,7 Revisionen pro Seite. Jede:r Nutzer:in hat 17,2 mal Seiten bearbeitet. Jede Seite ist im Schnitt mit 5,7 Kategorien versehen worden und beinhaltet 224,6 Links und Bilder.

In jedem Monat wurden 2027 Revisionen angefertigt, diese haben im Schnitt 335,1 Bytes

(Stellen) verändert. 119471 Revisionen (26,2 Prozent) haben Daten aus den Seiten genommen, 305139 (66,9 Prozent) haben Daten hinzugefügt. 6,9 Prozent haben im gleichem Maße Daten hinzugefügt und entfernt (Abb. 2).

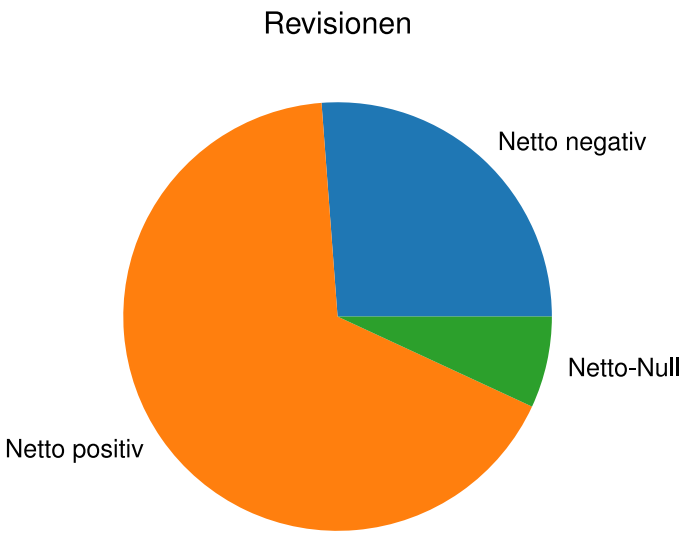

Abbildung 2: Revisionen nach Änderung

4.2 Weitere Werte

Pro Kategorie sind im Datensatz durchschnittlich 4,5 Revisionen vorhanden. Die Werte entstammen einem großen Wertebereich von 2 bis 458 Revisionen pro Kategorie. Die 15 am meisten vertretene Kategorien sind in Abbildung 3 dargestellt. Darunter sind fünf interne Kategorien der Wikipedia, die der Verwaltung dienen. Spitzenreiter ist die Kategorie “Mann”, 18 Prozent der Seiten ist diese Kategorie zugeordnet, gefolgt von “Deutscher”, das hier als generisches Maskulinum verwendet wird. “Frau” folgt auf dem fünften Platz, entsprechend 4 Prozent der Seiten, dies ist jedoch unter einem Viertel an Revisionen im Bezug zur Kategorie “Mann”. Zumindest der hier vorliegende Datensatz ist also deutlich von Personen, insbesondere Männern, geprägt.

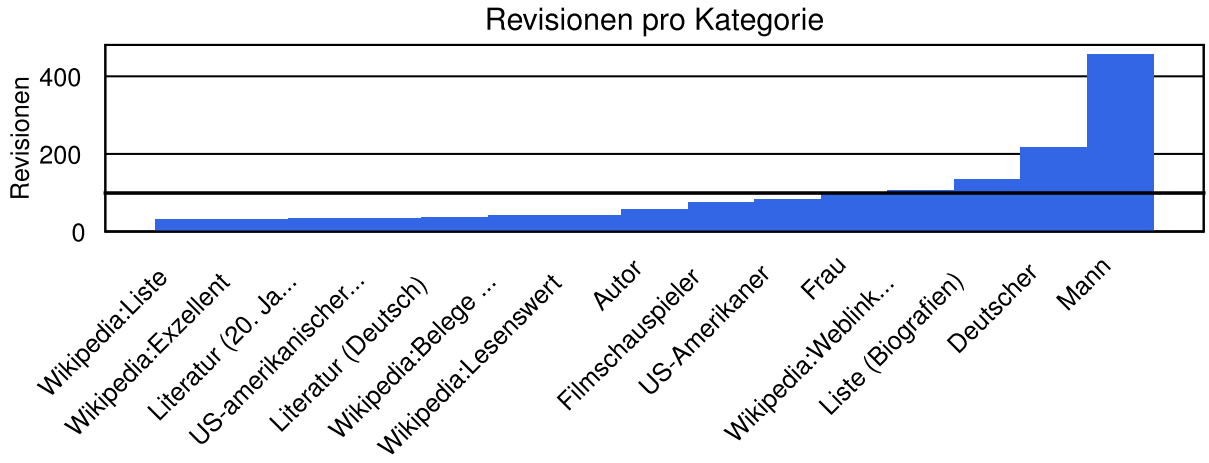

Abbildung 3: Revisionen nach Kategorie

Abbildung 4 zeigt, dass es viele Nutzer:innen mit nur wenigen Editierungen, also Revisionen, gibt. Erst durch die doppelt-logarithmische Darstellung lassen sich Details erkennen. Bis zu etwa einer Revisionszahl in der Größenordnung von 1000 nimmt die Zahl der Nutzer:innen exponentiell ab. Darüber hinaus finden sich nur noch einzelne Nutzer:innen. Ein großer Teil der Wikipedia-Inhalte stammt somit von einer großen Zahl verschiedener Nutzer:innen.

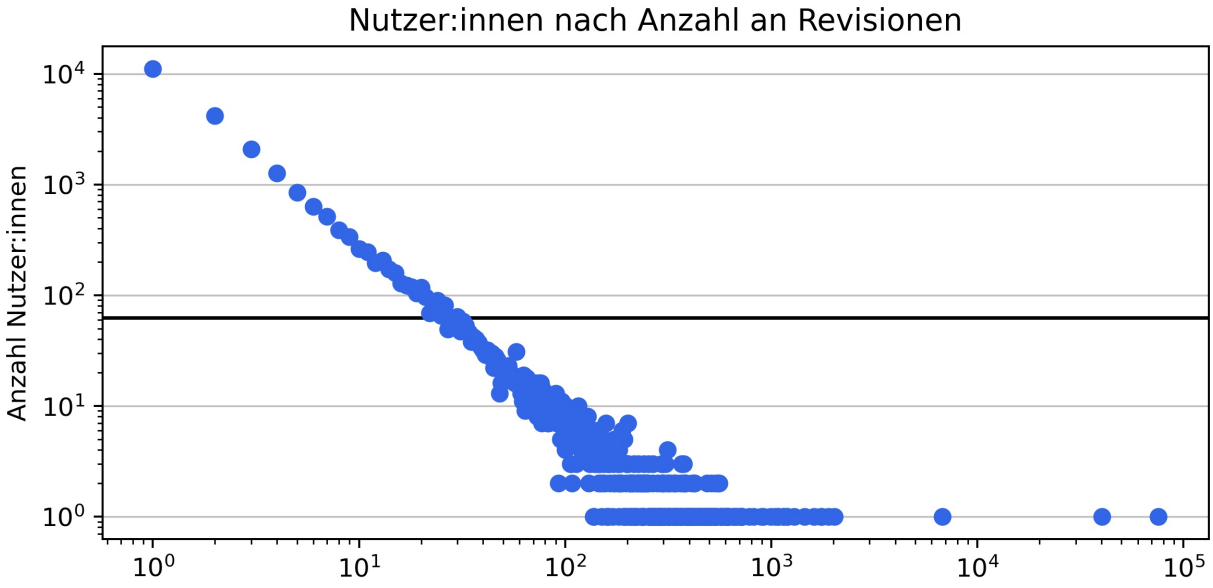

Abbildung 4: Revisionen nach Nutzer:in

4.3 Netzwerk-Analyse

Für die Netzwerkanalyse müssen die Daten in einen Bezug zueinander gebracht werden. Hierfür wird die Zahl an Revisionen an einem Artikel die Nutzer:innen gemeinsam haben herangezogen. Editierte Nutzer:in A den Artikel Y 5 mal und Nutzer:in B denselben Artikel 8 mal, so wird eine Verknüpfung mit der Stärke 5 (das Minimum beider Revisionsanzahlen) zwischen ihnen angelegt. Zusätzlich wird eine farbliche Sortierung der Nutzer:innen in eine Kategorie vorgenommen. Diese Kategorie entstammt der Seite die am meisten editiert wurde, aus den Kategorien der Seite wird die meist genutzte Kategorie extrahiert. Um einen besseren Überblick über die Daten zu erhalten und um die Bedienung von Gephi mit der vorhandenen Rechenleistung noch flüssig gestalten zu können wird der Datensatz zuvor reduziert. Lediglich Nutzer:innen mit mehr wie zwei Revisionen werden aufgenommen, Kanten (gemeinsame Revisionen) müssen größer wie fünf sein. Somit hat der Datensatz etwa 10500 Nutzer:innen als Knoten und 53000 Kanten.

Eine Layout-Optimierung des Netzwerks mit dem Force-Atlas-2-Algorithmus und eine nicht-lineare Größenverteilung der Knoten ist in Abbilung 5 zu sehen. In Grau sind hier Nutzer:innen dargestellt denen keine Kategorie zugeordnet werden konnte. Die beiden großen Punkte in der Mitte des Netzwerks gehören zu einer IP-Adresse und zum “APPERbot”. Letzterer übernimmt automatisiert Korrekturen in Artikeln, erstere lässt sich vermutlich ebenfalls als Bot deklarieren aufgrund der hohen Anzahl an Revisionen und der starken Verknüpfung. Diese beiden Nutzer:innen befinden sich in der Mitte des Netzwerks, da sie durch ihren automatisierten Ansatz keine thematischen Abgrenzungen in ihren Revisionen aufweisen.

An den Rändern des Netzwerks lassen sich mehrere kleine, farbliche Cluster erkennen. Es handelt sich dabei um Nutzer:innen die vorwiegend in der gleichen Kategorie Editierungen vornehmen. Trotz farblicher Ähnlichkeit kommen keine Cluster mit gleicher Kategorie mehrfach vor, die Ähnlichkeit entsteht durch die Limitierung des Farbenspektrums bei den vorhandenen 900 Kategorien. Die Cluster sind jedoch immer wieder durchsetzt von Nutzer:innen anderer Kategorien. Das Netzwerk weist grundstzlich eine geringe Cluster-Bildung auf.

Der Mittlere Grad im Netzwerk beträgt 10,2. Ein:e Nutzer:in ist also im Schnitt mit 10,2 weiteren Nutzer:innen verknüpft. Die durchschnittliche Kantenlänge beträgt 48 Prozent des Netzwerkdurchmessers, es existieren somit nur wenige Verbindungen zu den unmittelbaren Nachbarknoten.

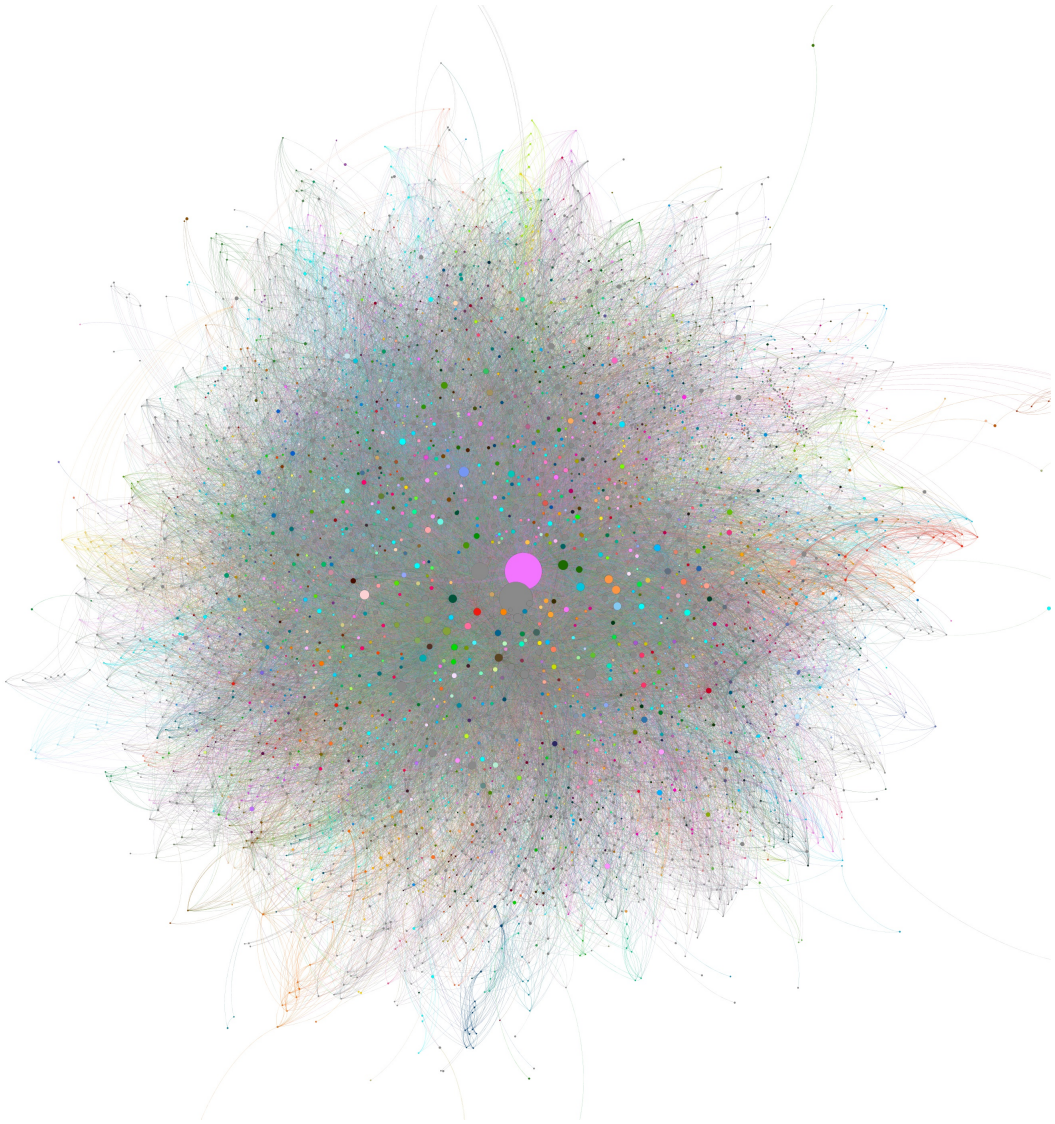

Abbildung 5: Netzwerk-Graph (ForceAtlas 2, konvergiert)

Bricht man den ForceAtlas-2-Algorithmus ab, deutlich bevor das Netzwerk konvergiert, gerade nachdem es sich entfaltet hat, so lässt sich der Aufbau des Netzwerks noch besser erkennen (Abb. 6). Die beiden sehr großen Knoten sind noch nicht ins Zentrum des Netzwerks gewandert, einzelne Knoten bilden sich an den Rändern heraus. Diese sind stark mit den Knoten im Zentrum verknüpft, aber lediglich über die jeweils gleiche Kategorie. Es handelt sich bei

den Knoten am Rand des Netzwerks als um Nutzer:innen die zwar, je nach Größe, durchaus einige Editierungen vorweisen können, jedoch hauptsächlich in einer Kategorie verbleiben.

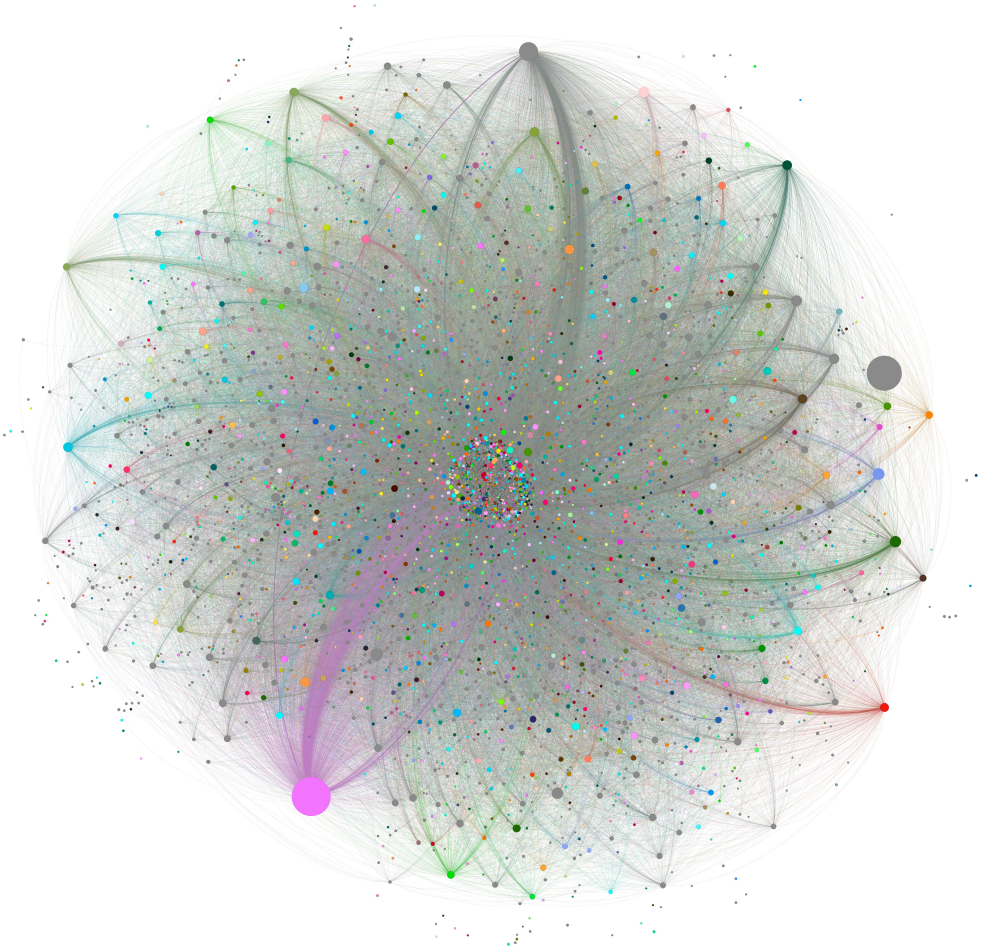

Abbildung 6: Netzwerk-Graph (ForceAtlas 2, unterbrochen)

5 Fazit

Die deutsche Wikipedia kann auf eine große Anzahl an Nutzer:innen zurückgreifen. In diesem zufällig gewählten Datensatz sind weisen Nutzer:innen vorwiegend geringe Revisionszahlen

auf. Der überwiegende Teil der Revisionen fügt neue Inhalte zu den Seiten hinzu. Diese Seiten sind vowiegend über männliche, deutsche Personen.

Die Analyse des berechneten Netzwerks bestätigt diesen Eindruck, die Nutzer:innen mit den meisten Editierungen sind Bots. Das Netzwerk weist keine starke Clusterbildung auf. Die Nutzer:innen sind untereinander recht lose verbunden und auch über einzelne Kategorien heraus verknüpft. Ein geringer Anteil der Nutzer:innen ist jedoch hauptsächlich in einer einzelnen Kategorie unterwegs, diese bilden sich am Rand des Netzwerks heraus.

Literatur

[1] *Gephi.org*. https://gephi.org/.

[2] *Wikipedia Statistik*. https://de.wikipedia.org/wiki/Wikipedia:Statistik.

[3] B. H. CONNOR MCMAHON, ISAAC JOHNSON, *The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies*. https://brenthecht.com/publications/icwsm17_googlewikipedia.pdf, 2017.

[4] JGOLDSMITH, *Wikipedia - Pypi*. https://pypi.org/project/wikipedia/.

[5] MATPLOTLIB, *Matplotlib - Pypi*. https://pypi.org/project/matplotlib/.

[6] K. T. TRUONG, *Soziale Netzwerke in Wikipedia*. https://elib.uni-stuttgart.de/bitstream/11682/9764/1/Soziale%20Netzwerke%20in%20Wikipedia.pdf, Jan. 2018.

[7] WIKIPEDIA, *Deutschsprachige Wikipedia*. https://de.wikipedia.org/wiki/Deutschsprachige_Wikipedia.