Woche 12 Anwendungen: Topic Modelling (Teil 2)

Skript

Erarbeitet von
Dr. Jacqueline Klusik-Eckert im Interview mit M. Ed. Stefan Reiners-Selbach

Inhalt

Lernziele

- Wissen über den Einsatz von Clustering als Methoden in den Geisteswissenschaften erhalten
- Unterschiedliche methodische Ansätze für den Einsatz von Clustering kennenlernen
- Kennenlernen eines möglichen Szenarios für den Einsatz von Clustering in den Geisteswissenschaften

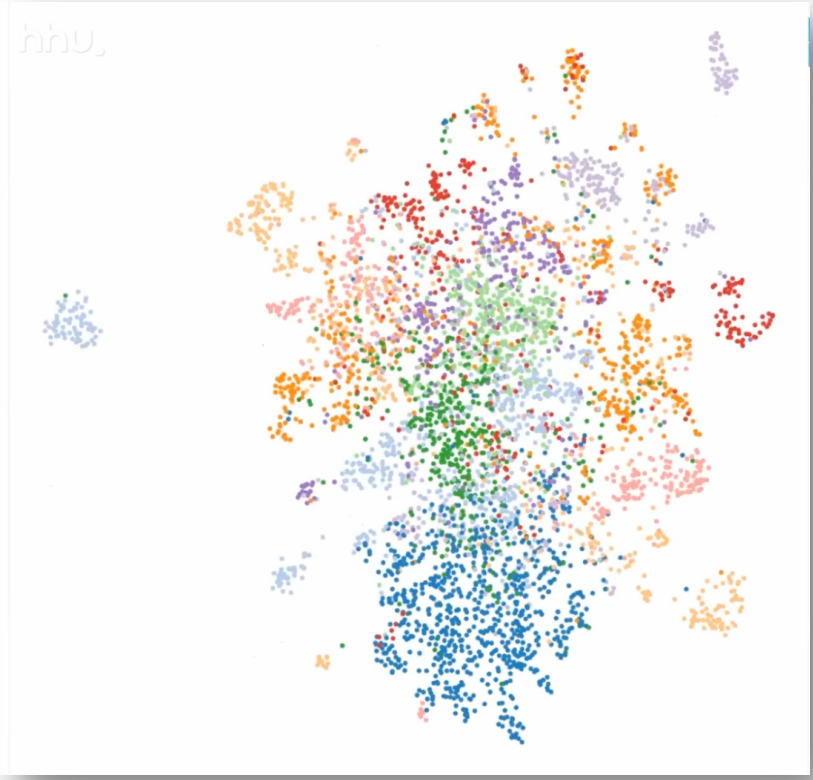
Interview

Jacqueline:
Du hast mir auch noch ein zweites Beispiel versprochen. Her damit. Oh, ein ganz schön bunter Haufen. Was für Daten sehen wir denn da jetzt eigentlich?

Stefan:
Das ist die Zeitschrift für Völkerpsychologie und Sprachwissenschaft, die von 1860 bis 1890 erschienen ist. Hinter dem Begriff Völkerpsychologie, der klingt ja ganz schrecklich, verbirgt sich eigentlich so etwas wie die erste Kulturwissenschaft. Der Begriff Volk wurde da nur

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

reingebracht, weil es eben für die Theoriemacher des Ganzen die prototypische soziale Gemeinschaft ist. Es ist sowas wie Kulturwissenschaften gepaart mit sowas wie einer Vorform der Soziologie, aber der Anspruch ist, die Gesamtheit der Kultur abzubilden. Das heißt also, alles soll dazugehören und das erste Mal nicht nur die Hochkultur, sondern eben auch die Alltagskultur und sowas, was damals eben die Volkskultur genannt wurde, und dabei geht es darum, die Gesetzmäßigkeiten in der Kultur zu entdecken.



Was ihr jetzt hier seht, es ist genau derselbe Vectorizer drüber gelaufen, genau dieselbe Dimensionsreduktion. Das heißt also, es ist wieder jeder einzelne Punkt als ein Text zu lesen. Jeder Punkt steht für einen Text. Die Farben hier stammen aber von Top2vec, das ist ein thematischer Clustering-Algorithmus. Der bildet nicht nur Cluster, die ich visualisieren kann anhand der Farben, sondern der liefert auch noch Wortlisten, die diese Cluster beschreiben sollen.

Jacqueline:
So jetzt hast du diese ganzen Cluster der Themen. Was lässt sich für dich aus der Visualisierung ablesen?

Stefan:
Ich kann an die Visualisierung rangehen und erstmal explorieren, das heißt schauen, was gibt es überhaupt für Themen und wie kombinieren sich vielleicht auch bestimmte Themen miteinander. Auf der anderen Seite kann ich auch Fragen stellen, die ich aus dem programmatischen Text in dieser Zeitschrift generiere, was zum Beispiel eigentlich Bestandteil dieser Wissenschaft sein soll, worum es alles gehen soll und welche Themen dazu gehören sollen. Überprüfen, gibt es diese Themen, kommen die vor und eben vielleicht

welchen Anteil machen sie aus und wie sind sie mit anderen Themenkomplexen verknüpft. Das heißt also, ich kann hier Muster erkennen, ich kann hier Ordnung in das Chaos dieser Zeitschrift, in diesen Datenhaufen bringen. Auch wenn das jetzt ein bunter Datenhaufen ist, der noch mal chaotisch aussieht. Aber gerade diese Farben, gerade diese Visualisierung, geben ja auch eine Struktur, die ich mir nochmal genauer anschauen kann, wo ich gerade mit diesem Clustering-Algorithmus Themen entdecken kann, worum es in dieser Zeitschrift geht.

Ich nehme jetzt einmal ein ganz kurzes Beispiel, hier oben diese Outlier. Wir können uns diese Wortlisten angucken. Das ist hier unter dem Titel Topic Keys Top2Vec: statistisch, Bevölkerung, Nation, Stammbaum, Nationalität. OK, komischer Themenkomplex. Aber hier sieht man also einmal, dass dieser Text „Zur Moralstatistik“ nicht zum gesamten Korpus zu passen scheint. Das ist ganz spannend, weil wir nämlich in der Mitte diesen hellgrünen Bereich haben. Das ist der Themenkomplex, der hier genannt wird: spekulativ, Erkenntnis, Gesetzmäßigkeit, Historiker, Geschehen. Wenn man ein bisschen in die Texte reinschaut und schaut, wie diese Begriffe da zusammenpassen, dann sieht man, dass das eigentlich quasi der Themenkomplex ist, der das Programm der gesamten Zeitschrift bestimmt. Worum soll es hier gehen? Es geht nämlich darum, Gesetzmäßigkeiten zu finden, in der Kultur beispielsweise, und Erkenntnisse zu generieren. Jetzt ist es total spannend zu sehen, dass es einige Texte gibt, einige Cluster, die sich bilden, also auch räumliche Cluster, die sich bilden, die total wenig mit diesen programmatischen Texten zu tun haben.

Jacqueline:
Also das Clustering hilft dir jetzt, diese Muster und Gesetzmäßigkeiten zu erkennen in einer Zeitschrift, die über 30 Jahre hinweg erschienen ist?

Stefan:
Also ich habe quasi den Überblick über diese Zeitschrift, den ich von Hand gar nicht hätte erreichen können, weil es einfach viel zu viele Texte sind, als dass ich sie hätte von Hand lesen können. Jetzt mithilfe dieser Analyse kann ich eben diesen Überblick gewinnen und quasi die Struktur hinter der gesamten Zeitschrift sehen. Und das ist auch gerade aus meinem Fachhintergrund spannend, weil die Theorie, das Programm vielleicht etwas völlig anderes ist als dann die Praxis, die hier umgesetzt wird. Also quasi, was soll da drin sein, worum soll es gehen? Wenn sich das stark davon unterscheidet, was tatsächlich in dieser Wissenschaft gemacht wird, dann ist das auf jeden Fall eine Erkenntnis, die total spannend ist und wo man nochmal genauer nachschauen muss, wie es denn wirklich aussieht.

Jacqueline:
Also von wie vielen Texten sprechen wir denn da jetzt eigentlich, die du mit dem Clustering hier sortiert und für dich strukturiert hast?

Stefan:
Also, das sind 482 Texte insgesamt in dieser Zeitschrift, die ich aber für diese Analyse noch einmal auseinandergenommen habe, die ich noch mal zu kleineren Chunks zerlegt habe. Denn manche Texte in dieser Zeitschrift sind 80 Seiten lang und behandeln dann in einem Text mehrere Themen und ich möchte aber mit dieser Visualisierung, mit dieser Analyse,

mit meinem Clustering eine größere Schärfe, eine größere Tiefe erreichen und genauer in die Texte reinschauen. Das habe ich eben dadurch erreicht, dass ich die Texte, diese langen Texte, in kleinere Segmente zerlegt habe und dann eben hier visualisiert habe.

Jacqueline:
Naja, jetzt kennst du diese Zeitschrift ja aufgrund deiner Forschung schon sehr gut. Gibt es denn nach der Analyse etwas, das dich überrascht hat? Man hat ja immer gewisse Erwartungen oder vorgefertigte Eindrücke über Zeitschriften. Hat sich da etwas aufgrund dieser Visualisierung bei dir geändert?

Stefan:
Ja, absolut. Also es ist für mich einfach überraschend gewesen … Um das ganz konkret zu machen: Sprachwissenschaften sollen eben ein Teil des Ganzen sein. Aber wieso diese Sprachwissenschaften die absolute Mehrheit thematisch an Texten ausmachen, das ist schon ziemlich überraschend. Das passt aber wiederum eigentlich doch ganz da rein, denn es sollen quasi in dieser Zeitschrift Gesetzmäßigkeiten in der Kultur entdeckt werden. Ich glaube, und ich glaube, das kann man ganz gut behaupten, dass die Sprachwissenschaften hier für diesen Gesamtkomplex der Zeitschrift als Vorbild dienen, weil die Grammatik in der Sprache so etwas wie eine Gesetzmäßigkeit der Kultur ist. Das heißt also, es ist total spannend zu sehen, wie Sprachwissenschaft hier eingebunden wird, weil es eben diese Vorbildfunktion hat. Aber auf der anderen Seite Themen, die eigentlich dazugehören sollten, ziemlich weitab von den restlichen Themen der Zeitschrift geclustert werden und ziemlich weitab sich dann eben auch in der Visualisierung finden, wo dann die Frage ist: Wieso ist da diese Lücke, also wieso passt das scheinbar nicht zu den anderen Texten, wieso passt das da scheinbar nicht so rein?

Jacqueline:
Wie es bei Kultur immer so ist. Es gibt einfach ein unglaublich breites Spektrum. Und anscheinend hat die Zeitschrift es geschafft, das irgendwie abzudecken.

Stefan:
Ja, es ist wirklich verrückt, weil den Begriff der Kulturwissenschaften gibt es zu diesem Zeitpunkt noch nicht. Deswegen verbirgt sich das alles hinter diesem sperrigen Begriff Völkerpsychologie. Aber der Anspruch war eben, die Gesamtheit der Kultur und hier auch das erste Mal sowas wie Alltagskultur, mit abzubilden. Ich glaube, das ist ein bisschen auch das, worum es dann in der Forschung geht. Das ist ein bisschen zu viel auf einmal gewesen, was man sich da vorgenommen hat, dass man quasi die gesamte Breite der Kultur und eben das erste Mal auch nicht nur die Hochkultur behandeln will, sondern eben alles. Das zeigt sich wiederum ja auch in dieser Visualisierung. Aber ich glaube, umso spannender ist das, weil diese Wissenschaft sich so bunt aufstellt, das Ganze dann eben auch so zu entdecken und zu schauen: Ok, welche Themen gibt es da, was kommt da alles drin vor?

Jacqueline:
Vielen Dank für diese tollen Beispiele, Stefan. Jetzt haben wir nicht nur einen Einblick bekommen, für was man Clustering alles benutzen kann, wie es als Methode in den Geistes- und digitalen Geisteswissenschaften eingesetzt werden kann. Wir haben auch noch etwas über die ganz frühen Anfänge der Kulturwissenschaft gelernt.

Stefan:
Ja, ich freue mich, dass ich das Beispiel mitbringen konnte und euch einmal ein bisschen in meine Arbeit einführen konnte. Ich glaube, es ist total spannend, sich in den Geisteswissenschaften auch mit solchen digitalen Methoden auseinanderzusetzen. Es gibt immer mehr digitale Texte und es gibt immer mehr digitale Methoden, warum sollte man sich nicht darauf stürzen.

Jacqueline:
Ja, warum sollte man sich nicht darauf stürzen? Danke dir.

Stefan:
Gerne.

Quellen

Quelle [1] Zeitschrift für Völkerpsychologie und Sprachwissenschaft. (1860.). https://opacplus.bsb-muenchen.de/search?id

Weitere verwendete Literatur

Angelov, Dimo (2020): Top2Vec: Distributed Representations of Topics, arXiv:2008.09470v1

Bokeh Development Team (2018). Bokeh: Python library for interactive visualization, http://www.bokeh.pydata.org

McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

Noichl, Maximilian (2019): Modeling the structure of recent philosophy, Synthese, https://doi.org/10.1007/s11229-019-02390-8

Pence, C. H. (2022). Testing and discovery: Responding to challenges to digital philosophy of science. Metaphilosophy, 53, 238–253. https://doi.org/10.1111/meta.12549

Lean, O. M., Rivelli, L., & Pence, C. H. (2023). Digital Literature Analysis for Empirical Philosophy of Science'. British Journal for the Philosophy of Science, 74, https://doi.org/10.1086/715049

Disclaimer

Transkript zu dem Video „Woche 12 Anwendungen: Topic Modelling (Teil 2)“, Dr. Jacqueline Klusik-Eckert, M. Ed. Stefan Reiners-Selbach.