Explorative Datenanalyse der Videos deutscher Hochschulen auf YouTube

Fallstudie im Sommersemester 2024

Inhaltsverzeichnis

Einführung

Fast alle deutschen Hochschulen nutzen YouTube als Plattform zur Kommunikation ihrer Inhalte. Die Ziele sind dabei vielfältig und reichen von Wissenschaftskommunikation bis hin zu Studierendenmarketing. In dieser Fallstudie untersucht ihr die YouTube-Videos von deutschen Hochschulen hinsichtlich ihres Inhalts und adressiert in einer explorativen Datenanalyse eine Reihe spannender Fragen. Ziel der Analyse ist es, die Nutzung von YouTube durch Hochschulen besser zu verstehen, Muster zu erkennen und mögliche Unterschiede zwischen den Hochschulen zu identifizieren.

Aufgabe 1: Datensammlung

In eurer Analyse sollt ihr die Videos der größten niedersächsischen Hochschulen mit denen der erfolgreichsten deutschen privaten Hochschulen vergleichen. Als Filterkriterium soll eine Mindestanzahl Studierende von 10.000 oder mehr verwendet werden.

Erstellt dazu zunächst eine Liste der YouTube-Videos für jede der insgesamt 10 Hochschulen, auf die das Kriterium zutrifft. Ladet anschließend die Tonspur der Videos basierend auf dieser Liste herunter und transkribiert sie mithilfe des Whisper-Modells auf dem HPC-Cluster

der Hochschule Osnabrück. Speichert die Transkripte in separaten Textdateien. Überführt anschließend alle Transkripte in eine große CSV-Datei mit den beiden Spalten `yt_id` und `transcript`.

Laut Wikipedia sind die in der Tabelle aufgeführten Hochschulen die größten in Niedersachsen, gemessen an der Studierendenzahl. Die Hochschule Braunschweig wurde trotz mehr als 10.000 Studierenden nicht berücksichtigt.

Hochschule	Anzahl Studierende	YouTube Channel ID
Universität Hannover	28.925	UCC0Eg0d1gGkpChelGo4-b4g
Universität Göttingen	28.614	UCzg-z2TL0Ks4Efz5o0z7AxQ
TU Braunschweig	17.709	UC8X4NAyIUr9Q12hVUOoqyhQ
Universität Oldenburg	15.635	UCbVtIjTIrNYqozDxIJPjHHw
Universität Osnabrück	13.640	UCN_aFLAXQEKZIa96PXxbi0Q
Hochschule Osnabrück	13.620	UCblmQHzIyzVHnKHwu0nISrw

Die erfolgreichsten privaten Hochschulen in ganz Deutschland mit mehr als 10.000 Studierenden sind laut Wikipedia:

Hochschule	Anzahl Studierende	YouTube Channel ID
International University	85.483	UCR4ZvrahpeSe1TBrmObSkSw
FOM	58.836	UCdxxZI5F0u0l_wY8Bimsnow
Hochschule Fresenius	17.045	UCq9E5LlsAWjzK5RkvhWUZKQ
Hamburger Fern-Hochschule	13.629	UCSoSuBd4s-xJWHxJfqQCuvQ

> **Startet bald!**
>
> Beginnt frühzeitig mit diesem Aufgabenteil. Wie ihr wisst, kann es bei der YouTube API zu Engpässen kommen. Gleichzeitig dauert das Transkribieren einer Vielzahl von Videos einige Zeit.

> **Seid fair und schont Ressourcen**
>
> Nutzt für diesen Aufgabenteil eine Jupyter-Umgebung mit 20 GB GPU-Speicher. Mehr benötigt ihr für die Transkription mit dem Whisper-Modell nicht.

Aufgabe 2: Datenvorbereitung

Im zweiten Schritt sollt ihr die Transkripte mit den im Seminar kennengelernten Methoden weiterverarbeiten und zusätzliche Metainformationen extrahieren. In dieser Fallstudie sollt ihr

insbesondere mit den Sprachmodellen experimentieren, die ihr im Seminar kennengelernt habt. Findet geeignete Prompts, um nützliche Informationen aus den Transkripten **in Bezug auf die in Aufgabe 3 gestellten Fragen** zu extrahieren. Geht dabei iterativ vor und testet eure Ideen für Prompts zunächst an einer überschaubaren Anzahl von Texten, bevor ihr den gesamten Korpus verarbeitet. So könnt ihr Zeit sparen und euren Prompt zügig optimieren. Ein guter Prompt braucht einige (Fehl-)Versuche, um die gewünschten Informationen zuverlässig zu extrahieren.

Wenn ihr glaubt, einen passenden Prompt gefunden zu haben, wendet diesen auf alle Transkripte an und speichert die Ergebnisse in einer CSV-Datei. Die CSV-Datei soll neben dem Ergebnis des Sprachmodells auch die Spalte `yt_id` enthalten, um die gewonnenen Daten später einfach in R zusammenführen zu können.

Aufgabe 3: Explorative Datenanalyse

Im letzten Teil der Fallstudie sollt ihr auf Grundlage der YouTube-Metadaten aus der API sowie den transkribierten Texten aus den Videos die unten stehenden Fragen explorativ bearbeiten. Wie in Aufgabe 2 beschrieben, könnt ihr die Sprachmodelle verwenden, die ihr im Seminar kennengelernt habt, um weitere Informationen aus den Texten zu extrahieren. Die folgenden Fragen sollt ihr beleuchten:

Explorative Analyse der Metadaten

Die folgenden Fragen beziehen sich auf die Metadaten der YouTube-Videos. Für die Bearbeitung dieser Fragen benötigt ihr keine Transkripte.

1. Wie viele Videos hat jede Hochschule auf YouTube veröffentlicht?
2. Wie ist die Verteilung der Längen der Videos für jede Hochschule? Gibt es Unterschiede oder Auffälligkeiten zwischen den Hochschulen?
3. Seit wann sind die Hochschulen auf YouTube aktiv? Wie intensiv ist die Nutzung im Zeitverlauf?
4. In welchem Zeitabstand posten die Hochschulen neue Videos? Gibt es hier Muster zu erkennen?
5. Die Videos welcher Hochschulen werden am meisten angesehen? Vergleicht die Hochschulen diesbezüglich miteinander.
6. Werden kürzere Videos häufiger angesehen als längere?
7. In welcher Sprache veröffentlichen die Hochschulen ihre Videos? Wie sehen die Hochschulen im Vergleich aus?

8. Was sind die häufigsten Begriffe, die in den Titeln der Videos verwendet werden? Unterscheiden sich die Hochschulen hier?

Explorative Analysen unter Hinzunahme der Transkripte

9. Welche Themen werden in den Videos der Hochschulen adressiert?

    a. Findet geeignete Prompts, um Themen aus den Transkripten der Videos zu extrahieren! Überprüft die Ergebnisse der Prompts bevor ihr sie auf die gesamten Videos anwendet.

    b. Prüft, inwieweit ihr das Ergebnis der Sprachmodelle anschließend weiter verfeinern müsst, sei es mit einem weiteren Prompt oder mit den Möglichkeiten der Texttransformation in R, um es in eurer Analyse zu verwenden.

    c. Formuliert mindestens drei Fragen an die neu gewonnen Daten und führt geeignete Analysen zur Beantwortung durch. Bezieht dabei auch die Metadaten der Videos mit ein.

    Tipp: Neben dem Transkript könnt ihr dem Sprachmodell mit dem Prompt auch noch andere Informationen zu einem Video mitgeben.

10. Überlegt euch eine eigene Fragestellung, die ihr anhand der vorliegenden Transkripte und unter Zuhilfenahme eines Sprachmodells bearbeiten könnt. Formuliert Hypothesen für eure Fragestellung. Entwerft geeignete Prompts und wendet sie auf alle Transkripte an. Führt anschließend für eure Fragestellung und Hypothesen eine explorative Datenanalyse unter Einbezug der neu gewonnen Informationen durch!

Hinweise

Beachtet bitte unbedingt die folgenden Hinweise zur Fallstudie.

Bearbeitung

- Ihr sollt jede Frage unter Zuhilfenahme von mindestens einer geeigneten Visualisierung beantworten. Eure Aufgabe ist es, eine geeignete Visualisierungsform für die jeweilige Frage und Antwort zu finden.
- Die explorative Datenanalyse soll mit R und dem RStudio durchgeführt werden. Für sämtliche Datentransformationen und Visualisuerungen sollt ihr Funktionen aus dem Tidyverse verwenden.

- Notwendige Vorverarbeitungsschritte der Daten finden auf dem HPC-Cluster statt und werden mit Python durchgeführt. Das betrifft insbesondere die Transkription der Videos mit dem Whisper-Modell sowie die Anwendung der Open Source Sprachmodelle.
- Eine explorative Datenanalyse lebt von eurer Kreativität und Ausdauer. Gebt euch nicht mit der ersten Analyse zufrieden, die ihr erstellt. Das gilt insbesondere, aber nicht nur, für offene Fragestellungen. Probiert mehrere Ansätze aus. Wenn ihr ein interessantes Muster entdeckt, geht dem tiefer auf den Grund. Obwohl alle Teams die gleichen Aufgaben haben, wird es viele verschiedene Lösungswege geben.
- Wenn ihr bei einer Aufgabe nicht sicher seid, wie sie gemeint ist, oder Informationen fehlen, trefft sinnvolle Annahmen und dokumentiert diese als Kommentare in eurem R-Skript!
- Lesbarer Code ist besser als schwer verständlicher Code (verwendet den Pipe-Operator, brecht lange Ausdrücke in mehrere Zeilen, rückt den Code ein und fügt dort Kommentare hinzu, wo es angebracht ist).
- Reproduzierbarkeit ist in der Wissenschaft von großer Bedeutung. Stellt sicher, dass eure R-Skripte ohne zusätzlichen Aufwand auf einem anderen Computer ausgeführt werden können.
- Es ist in Ordnung, ab und an Code zu kopieren und sich von anderen inspirieren zu lassen, solange der kopierte Code verstanden und erklärt werden kann und die Quelle als Kommentar angegeben wird. Dies schließt auch große Sprachmodelle wie ChatGPT ein.
- *Make it work, make nice!* Erst die Funktion, dann die Ästhetik! Verliert euch nicht in Details wie Achsenformatierung oder Farbpaletten, bevor ihr die korrekte Datenbasis erstellt und die geeignete Visualisierungsform gewählt habt. Eine unpassende Visualisierung, selbst wenn sie optisch ansprechend ist, wird euch nicht helfen. Denkt umgekehrt!

Abgabe

- Führt den Code eurer Datenanalyse für die Abgabe in *ein* R-Skript zusammen. Verwendet Kommentare, um die Fragen, zu dem der jeweilige Code gehört, in eurem Skript zu kennzeichnen. Das Skript muss von oben nach unten fehlerfrei ausführbar sein.
- Erstellt für die Abgabe ein neues R-Projekt. Legt sämtliche benötigte Daten, die ihr in eurer Analyse verwendet, in einem Unterordner `data/` ab. Ladet die Daten in eurem Skript relativ zum Pfad eures Projekts.
- Erstellt einen weiteren Ordner `python/`, in den ihr alle verwendeten Jypyter-Notebooks für die Vorverarbeitung der Daten ablegt.

- Erzeugt für die Abgabe eurer Fallstudienergebnisse ein ZIP-Archive eures Projektordners. Die Abgabe erfolgt durch *ein* Teammitglied über den bereitgestellten Abgabeordner in ILIAS.

Hier ein *Beispiel* für die abzugebende Projektstruktur (die Dateien im Ordner `data/` können selbstverständlich variieren):

```
case_study/
--- data/
------ transcripts.csv
------ youtube_videos.csv
------ ...
--- python/
------ whisper.ipynb
------ phi3.ipynb
------ ...
--- analysis.R
--- case_study.RProj
```

Kurzvorstellung der Ergebnisse

- Im Prüfungsgespräch sollt ihr eure Ergebnisse für die eigene Fragestellung aus Aufgabe 3 (Punkt 10) vorstellen.
- Bereitet für die Vorstellung eine kurze Präsentation vor. Die Präsentation muss folgende Folien beinhalten:
  - **Fragestellung (1 Folie)**: Darstellung und Erläuterung eurer Fragestellung inklusive Begründung, warum ihr diese gewählt habt.
  - **Hypothesen (1 Folie)**: Eure Hypothesen zur Fragestellung, die ihr *vor* der Durchführung der explorativen Datenanalyse aufgestellt habt.
  - **Prompt Design (1-2 Folien)**: Die Prompts, mit denen ihr gearbeitet habt. Beschreibt auch verbal, wie ihr zu den Prompts gelangt seid.
  - **Ergebnisse (max. 3 Folien)**: Eure wichtigsten Analyseergebnisse in visueller Form sowie jeweils einem kurzen Satz zur Interpretation.
- Darüber hinaus sind keine weiteren Folien erlaubt.
- Der Zeitrahmen für die Präsentation darf 20 Minuten nicht überschreiten, anschließend sind 10 Minuten für Fragen vorgesehen.
- Jedes Gruppenmitglied sollte ungefähr den gleichen Sprechanteil haben.

- Gebt eure Präsentation im PDF-Format zusammen mit eurer ZIP-Datei über ILIAS ab.

Viel Spaß und Erfolg bei der Bearbeitung der Fallstudie!