Bericht der Gruppe E

Identifikation von Nutzergruppen - Reddit \rplace

Christian Funk, Sebastian Heuer, Greta Wetzel, Nicole Derr, Jannik Littmann
Martin-Luther-Universitï¿½t Halle-Wittenberg
christian.funk@student.uni-halle.de,sebastian.heuer@student.uni-halle.de,greta.wetzel@student.uni-halle.de,
nicole.derr@student.uni-halle.de,jannik.littmann@student.uni-halle.de

Figure 1: Some large rubber duck.

ABSTRACT

Some general hints on what to mention in an abstract: What are the questions you address? Why are they interesting? What approaches did you use? What answers did you find?

As for how to structure the abstract: Give some motivation and context on the general topic you address (1–2 sentences). Then state the specific questions you address (1–2 sentences) and describe how you approach them (2–3 sentences). Finally, results and some conclusion (1–3 sentences).

KEYWORDS

ï¿½bung „Big Data Analytics“, Sommersemester 20XX

1 INTRODUCTION

Jedes Jahr zum 1. April führt Reddit ein soziales Experiment durch. Im Jahr 2017 wurde dabei zum ersten Mal r/place vollkommen unangekÃ¼ndigt duchgefÃ¼hrt. FÃ¼nf Jahre spÃ¤ter wurde das gleiche Experiment wiederholt, jedoch mit VorankÃ¼ndigung. Bei r/place können Nutzer alle 5 Minuten einen Pixel auf einem Canvas setzen. In den Jahren 2017 und 2022 war der Canvas 1000 x 1000 Pixel bzw. 2000 x 2000 Pixel groß. Das Experiment dauerte 4 Tage.

Nach erster Verwirrung der Nutzer zu Beginn des Experiments von 2017, bildeten sich einzelne Communities, welche gemeinsam Projekte verfolgten. Es entstanden erste Artworks und kreative Explosionen bis hin zu KÃ¤mpfen um bestimmte Regionen auf dem Canvas.

Vor allem bei der Wiederholung 2022 kamen vermehrt Scripts fÃ¼r Bots zum Einsatz, was vor allem der VorankÃ¼ndigung geschuldet war, sodass Nutzer die MÃ¶glichkeit bekamen sich gezielt vorzubereiten. Anders als 2017, wurde die Wiederholung 2022 von Moderatoren begleitet, welche unerwÃ¼nschte Artworks zensierten.

Da sich viele Communities vor allem 2022 zusammenschlossen, wurde die Frage, "ob es mÃ¶glich ist, Nutzer anhand von Ã¤hnlicher Verhaltensweisen automatisch Nutzergruppen zuzuordnen?" zum Ziel dieses Projekts.

Weiterhin wurde auch betrachtet, ob Bots und Moderatoren identifizierbar sind und ob entsprechende Statistiken zu diesen entwickelt und ausgewertet werden kÃ¶nnen.

2 DATA

Der offizielle Datensatz von r/place aus der Wiederholung 2022 beinhaltet 72 Millionen Pixel, von 6 Millionen Nutzern und besitzt eine GrÃ¶Ãe von ca. 22GB. Dieser ist in einer CSV-Datei strukturiert, welche den Zeitpunkt des gesetzten Pixels (timestamp), eine gehashte Nutzeridentifikation ($user_id$), die verwendete Farbe ($pixel_color$) und die x- und y-Koordinaten auf dem Canvas (coordinates) beinhaltet. Eine Besonderheit stellen die Moderatoren im Datensatz dar. Diese besitzen in der Spalte coordinates nicht nur eine x- und y-Koordinate, wie alle anderen Nutzer, sondern zwei x- und y-Koordinaten. Das erste Koordinatenpaar bildet dabei die obere linke Ecke des Zensur-Quadrates eines Moderators und das zweite Koordinatenpaar die rechte untere Ecke.

Zu finden sind die Daten online unter: https://www.reddit.com/r/place/comments/txvk2d/rplace_datasets_april_fools_2022/ und https://www.reddit.com/r/redditdata/comments/6640ru/place_datasets_april_fools_2017/ fÃ¼r 2017.

Der Datensatz der ersten DurchfÃ¼hrung 2017 ist mit einer GrÃ¶Ãe von ca. 1GB um einiges kleiner als der von 2022. Dieser hat eine Ã¤hnliche Gliederung wie der von 2022, jedoch wurden hier die x- und y-Koordinaten direkt getrennt aufgelistet statt unter einer gemeinsamen Spalte, wie in 2022. Zudem ist hier die Restriktion zu beachten, dass Nutzer nur zwischen 16 Farben wÃ¤hlen konnten und diese im Datensatz mit Integer-Values zwischen 0 bis 15 codiert wurden. Unter dem angegebenen Link fÃ¼r den Datensatz von 2017 findet sich eine entsprechende Tabelle, die wiedergibt, welcher Integer fÃ¼r welchen Farb-Hashwert steht.

Nachdem die DatensÃ¤tze runtergeladen waren, mussten diese zunÃ¤chst minimal bereinigt bzw. angepasst werden. Im Datensatz fÃ¼r 2017 mussten die Integer-Values der Farben in den jeweiligen Hashcode umgewandelt werden und fÃ¼r jeweils beide DatensÃ¤tze wurden die Werte der timestamp-Spalte in nutzbare Werte umgewandelt werden. DafÃ¼r wurde das UTC-Timestampformat zunÃ¤chst in das Unix-Epoch-Seconds-Format umgewandelt und anschlieÃend der kleinste Wert von allen timestamps abgezogen. Damit waren die timestamps so normalisiert, dass der erste Eintrag mit einem timestamp gleich null beginnt.

Eine weitere Bereinigung der Daten beinhaltet, dass Moderatoren und "normale" Nutzer gespalten wurden, sodass diese in separaten DataFrames betrachtet werden kÃ¶nnen. Im Datensatz von 2022 wurde damit die Spalte coordinates jeweils in die x- und y-Koordinaten fÃ¼r "normale" Nutzer aufgespalten und fÃ¼r Moderatoren in x1- , y1- , x2- und y2-Koordinaten. Eine solche Spaltung musste fÃ¼r den 2017-Datensatz entsprechend nicht durchgefÃ¼hrt werden. - Beispieldaten

3 IDENTIFIKATION VON NUTZERGRUPPEN

Fragestellung 1: Ist es mÃ¶glich anhand von ähnlichen Verhaltensweisen Nutzer ihren Nutzergruppen zuzuordnen?

26.04.

Betrachtung verschiedener $User_IDs$, die mehr als x-Mal zur selben Zeit und recht nah beieinander (Koordinaten) Pixel platziert/verändert haben

ZusÃ¤tzlich anhand der gewählten Farben der User bzw. Nutzergruppen betrachten, ob zwei oder mehr Gruppen miteinander

konkurrierten
03.05.
Definition Nutzergruppe:
- Nutzer ist zu verschiedenen Zeitintervallen aktiv
- Nutzer kann mehrere Interessen vertreten
- In jedem aktiven Zeitintervall vertritt ein Nutzer pro Raumzeitgebiet aber nur genau ein Interesse
- Überschneiden sich Interessen mehrerer Nutzer räumlich und zeitlich, so stehen diese im Konflikt zueinander
- Beispiel Deutschland - Frankreich Flaggen
10.05.
- Aggregation Nutzerdaten
- Raumzeitmetrik (SVM)
- 3-dimensionalität der Daten, x-, y-Koordinate, Zeit
- Zusammenfassen von Raumzeiten (Bounding Boxes)
- Infrastrktur?
17.05.
- Bilder Bounding Boxes und zusammenhängende Pixel (Zusammenhangskomponenten) mit Beschreibung

14.06. - Methodik weiteres Vorgehen

Abschluss
Was ist unser Abschluss, Ergebnis?

4 BOTS UND MODERATOREN

Statistische Auswertungen des Datensatzes:
In diesem Abschnitt soll nun auf die Auswertung hinsichtlich der Bots und Moderatoren eingegangen werden.
Ist es möglich Bots und Moderatoren anhand des Datensatzes zu identifizieren?
Können wir dazu entsprechende Statistiken entwickeln und auswerten?
17.05.
Vermutungen:
- viele Nutzer setzen nur wenige Pixel, grosse Häufigkeit bei kleiner Pixelzahl
- Bots setzen konstant viele Pixel, leicht erhöhte H"aufigkeit bei grosser Pixelzahl
- Testdatensatz umfasst eine Stunde, maximal 12 Pixel pro Nutzer möglich
Vorgehen:
- Nach Nutzern aggregierte Pixeldaten nochmal nach Pixelanzahl aggregieren
- Grafik
30.05.
Statistiken:
- meiste umkämpfter Pixel: **Grafik** ev. in 3-Dimensionen
- meiste verwendete Farben: **Grafik**
21.06.
- Filterkriterien Bots: ca. 14.000 Bots im Testdatensatz
- Filter Moderatoren, Welche Farben wurden verwendet?
28.06.
- Grafik Bots 2022
05.07.
- Präzisierung Filter Bots
- Bearbeitung Datensatz 2017, **Grafik Bots 2017**
- Grafik Mods
- Fragen zur Rekonstruktion

4.1 Moderatoren

Hier Input von Nicole! - Die Struktur der CSV-Datei erleichterte die Identifikation von Moderatoren im Datensatz 2022
- Der Datensatz von 2017 blieb dabei unbetrachtet, da hier unklar ist, ob Moderatoren aktiv waren und wenn ja, ob diese mit in den Datensatz aufgenommen wurden
- Wie bereits eingangs erwÃ¤hnt, besitzen Moderatoren in der Spalte coordinates vier Werte, welche die x- und y-Koordinate der linken oberen Ecke und der rechten unteren Ecke des Zensur-Quadrates darstellen
- Anhand dieser Struktur wurde fÃ¼r die Moderatoren ein von den "normalen" Nutzern getrenntes DataFrame generiert, sodass keine Identifikation in engeren Sinne durchgefÃ¼hrt werden musste
- ZunÃ¤chst legte sich der Fokus auf einfach auszuwertende Statistiken:
- WÃ¤hrend des gesamten Experiments 2022 waren 16 Moderatoren aktiv, von denen jeder im Durchschnitt ca. 6262 Pixel zensierte und insgesamt 100.197 Pixel zensiert wurden.
- Im Vergleich dazu wurden 159.024.375 Pixel von "normalen" Nutzer gesetzt.
- Teilt man nun die Anzahl der Zensuren durch die Anzahl an gesetzten Pixeln, so erhÃ¤lt man mit 0,063% das VerhÃ¤ltnis , in welchem zensiert wurde.
- Aus diesem prozentualen Anteil lÃ¤sst sich schlieÃen, dass Moderatoren kaum Pixel zensierten wÃ¤hrend des gesamten Verlaufes.
- Das bedeutet wiederum , dass entweder kaum unerwÃ¼nschte Artworks von Nutzern gezeichnet wurden oder Artworks so schnell Ã¼berzeichnet wurden, dass die Moderatoren kaum aktiv werden mussten.
- HIER BIDL EINFÃGEN MIT DEM CANVAS WO DIE RAHMEN DER ZENSUR-QUADRATE ZU SEHEN SIND

- Die Abbildung Nr.? zeigt ein Canvas auf dem die Rahmen der Zensur-Quadrate der Moderatoren nachgezeichnet wurden.
- Damit erhÃ¤lt man einen Ãberblick darÃ¼ber, in welchen Regionen des gesamten Canvas Moderatoren Ã¼berhaupt aktiv werden mussten
- Die Abbildung verdeutlich noch einmal, dass Moderatoren kaum aktiv werden mussten und das nur in der unteren Region zensiert werden musste
- Da das Canvas wÃ¤hrend des gesamten Verlaufes 2022 zwei mal verdoppelt wurde und der untere Teil dabei als letztes hinzukam, lÃ¤sst sich daraus schlieÃen, dass Moderatoren erst zu einem spÃ¤teren Zeitpunkt des Experimentes aktiv werden mussten

- Eine aufwendigere Statistik fÃ¼r die Moderatoren, die betrachtete werden sollte:
- Wie viele Pixel wurden von den Bots bzw. "normalen" Nutzern im VerhÃ¤ltnis zensiert? LÃ¤sst sich daraus ableiten, dass eher von Bots oder von Nutzern gesetzte Pixel zensiert werden mussten?
- Um dies zu beantworten, musste zunÃ¤chst eine Funktion geschrieben werden, welche rÃ¼ckwirkend die Pixel identifiziert, die mit dem jeweiligen Zensur-Quadrat eines Moderators zensiert wurden.
- Beschreibung wie die Funktion vorgeht:
- Vor dem Funktionsaufruf werden zunÃ¤chst alle Pixel bestimmt, die von den Zensuren der Moderatoren betroffen sind
- Dies dient als einer der Inputs der Funktion
- Den zweiten Input bietet ein DataFrame mit allen gesetzten

Table 1: Some example table.

Some entries	Some numbers
Entry A	400 million
Entry B	300 million
Entry C	200 million

Pixeln im Verlaufe des Experimentes
- DafÃ¼r wird innerhalb der Funktion jeder Pixel schrittweise betrachtet, der von einer Zensur betroffen ist
- Mittels einer For-Schleife wird dann von dem Zeitpunkt der Zensur aus runtergezÃ¤hlt, bis der Zeitpunkt 0 erreicht wurde oder der zensierte Pixel gefunden wurde
- wurde ein zensierter Pixel identifiziert, wird dieser in ein neues DataFrame geschrieben, welches alle zensierten Pixel mit ihren Daten beinhaltet
- Am Ende liefert die Funktion das entsprechende DataFrame mit allen zensierten Pixel und deren Daten, wie UserID, x- und y-Koordinate, etc.
- Ein Zwischenergebnis dieser Funktion ist, dass die zensierten Pixel nachgezeichnet werden kÃ¶nnen.
- Eine Erwartung was, dass die zensierten Artworks evtl "interessante" Themen darstellen, wie z.B. politisch Themen die unerwÃ¼nscht waren
- Jedoch wurde diese Erwartung nicht erfÃ¼llt, da es sich bei den zensierten Pixeln nur um unangemessene, kleinere zeichnungen handelte, die in diesem Rahmen auch nicht weiter betrachtet werden sollen.
- Aufgrund von Problemen mit der Infrastruktur, war es nicht mÃ¶glich den gesamten Datensatz aufeinmal auszuwerten, sodass die Frage "Wie viele Pixel wurden von den Bots bzw. "normalen" Nutzern im VerhÃ¤ltnis zensiert? LÃ¤sst sich daraus ableiten, dass eher von Bots oder von Nutzern gesetzte Pixel zensiert werden mussten?" entsprechend nicht beantwortete werden konnte
- (Ein Problem ist, dass die DatensÃ¤tze voneinander fÃ¼r diese Auswertung zu stark zusammenhÃ¤ngen, als das man diese hÃ¤tte einzeln auswerten und zu einem Gesamtergebnis zusammentragen kÃ¶nnen) -> entsprechend im Fazit weiter drauf eingehen bzw. sagen dass es einen mÃ¶glichen Ausblick bieten kÃ¶nnte!
- damit bleibt diese Frage weiterhin offen

Vergleiche
- Statistiken 2022 und 2017 im Vergleich - Grafiken
- Vergleich Arbeit und Menge Bots
- Vergleich der Zensur von Mods

5 EVALUATION

Some evaluation section if appropriate. You might want to refer to some table with results in this section (e.g., to Table **??**).

6 CONCLUSION

The introduction in less detail. Summarize story in retrospective, give outlook on possible next steps. Semi-technical ...