

Bericht der Gruppe E

Identifikation von Nutzergruppen - Reddit \rplace

Christian Funk, Sebastian Heuer, Greta Wetzel, Nicole Derr, Jannik Littmann

Martin-Luther-Universität Halle-Wittenberg

erika.musterfrau@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

Übung „Big Data Analytics“, Sommersemester 20XX

1 INTRODUCTION

Reddit führt jedes Jahr zum 1. April ein soziales Experiment durch. 2017 war dies zum ersten Mal \ rplace. Bei \ rplace können Nutzer alle 5 Minuten einen Pixel auf einem Canvas verändern. 2017 war der Canvas 1000 x 1000 Pixel groß. Das Experiment dauerte 4 Tage.

Nach erste Verwirrung der Nutzer zu Beginn fanden sich Communities zusammen, welche gemeinsam Projekte verfolgten. Es entstanden erste Artworks und kreative Explosions bis hin zu Kämpfen um bestimmte Regionen auf dem Canvas.

- 2022 Wiederholung des Experiments
- Nutzung von Scripts und Bots
- Moderation des Events (Mods)
- Bilder Canvas 2017 und 2022

2 DATA

Der Datensatz von \ rplace aus 2022 beinhaltet 72 Millionen Pixel, von 6 Millionen Nutzern. So wurden in den 4 Tagen des Experiments jede Stunde 2,5 Millionen Pixel gesetzt.

Der Datensatz ist in einer CSV-Datei strukturiert, welche den

Zeitpunkt einer Aktion (timestamp), eine verschlüsselte bzw. gehashte Nutzeridentifikation ($user_id$), die gesetzte Pixelfarbe ($pixel_color$) und die x- und y-Koordinaten auf dem Canvas ($coordinates$) beinhaltet.

Zu finden waren die Daten online unter: ??? (Link als Quelle verknüpfen)

- Infos zu Daten 2017
- Unterschiede Datensätze 2017-2022 26.04.
- wie haben wir die Daten verarbeitet
- ev. Besonderheiten zu Mods
- ev. Beispieldaten

Description of the datasets you used. You might want to mention: how acquired, how post-processed / cleaned, some basic characteristics, some examples from the data, etc.

3 IDENTIFIKATION DER NUTZERGRUPPEN

Fragestellung 1: Können anhand von ähnlichen Verhalten der Nutzer Nutzergruppen identifiziert werden?

26.04.

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

Zusätzlich 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-dimensionaleität der Daten, x-, y-Koordinate, Zeit
- Zusammenfassen von Raumzeiten (Bounding Boxes)
- Infrastruktur?

17.05.

- Bilder Bounding Boxes und zusammenhängende Pixel (Zusammenhangskomponenten) mit Beschreibung

14.06. - Methodik weiteres Vorgehen

Abschluss

Was ist unser Abschluss, Ergebnis?

Table 1: Some example table.

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

4 BOTS UND MODS

Statistische Auswertungen des Datensatzes:

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 Häufigkeit 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

Abschluss

- 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 1).

6 CONCLUSION

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