Projektseminar – erste Projektidee

Karoline Jüttner

1) Welche Themen im Bereich der Informationswissenschaft und Sprachtechnologie interessieren mich?

Informetrie, Datenverarbeitung und -visualisierung, Social Media, IR, Gamification, …

2) Welche Programmiersprachen und Tools liegen mir bzw. welche würde ich gerne ausprobieren?

Python in Verbindung mit einem Visualisierungstool

3) Wie könnte ich mich in Form eines Projektes mit dem gewählten Thema auseinandersetzen?

Erste Idee: Ich habe einen Datensatz zur Google Play App von 2015 gefunden, mit ziemlich detaillierten Informationen zu Installationen, Abstürzen und Bewertungen und das bezogen auf u.a. Länder, Sprachen, App-Version und Geräte. Außerdem gibt es einen zweiten Datensatz von 2016, mit dem man durch einen Vergleich die Veränderungen untersuchen könnte. Das ganze würde ich dann mit Python einlesen, weiterverarbeiten und mithilfe von Visualisierungstools darstellen und eine Datei anlegen, in der die Ergebnisse und das Vorgehen vorgestellt werden.

4) Ist dies in der Bearbeitungszeit (ca. 4-8 Wochen) realisierbar?

Den kompletten Datensatz abzuarbeiten würde in der Zeit vermutlich nicht klappen, ich würde mich dann auf ein paar Aspekte davon beschränken und diese lieber ausführlich untersuchen und vergleichen.

Detaillierter Projektplan

Die Idee:

Es sollen fünf verschiedene CSV-Dateien, jeweils für die Jahre 2015 und 2016, eingelesen und verarbeitet werden, die sich mit der Google Play App beschäftigen. Dort werden Installationen, Länder und Sprachen, Abstürze, Geräte und App-Versionen sowie Bewertungen und Länder miteinander in Bezug gesetzt. Mithilfe von einigen Listen, die mit Funktionen bearbeitet werden, sollen neue Erkenntnisse aus diesen Datensätzen gewonnen werden. Diese sollen dann kurz erklärt und mit Diagrammen veranschaulicht werden.

Update:

Ich habe einen Datensatz gefunden, mit dem Titel "Google Play statistics 201501". Zunächst dachte ich, es handle sich um die Google Play App selbst, allerdings sind die Zahlen dafür zu niedrig. Beispielsweise ist der höchste Eintrag für Deutschland in der Spalte "total user installs" nur 115835. Daher interpretiere ich den Datensatz so, dass es sich um irgendeine App aus dem Google Play Store handelt, deren Name nicht veröffentlich wird. Außerdem überschneidet sich der Datensatz von 2016 nur in zwei von mir untersuchten Aspekten, nämlich der Zusammenhang zwischen Installationen und Ländern und Installationen und Sprachen. Weil dadurch einige geplante Vergleiche wegfallen, möchte ich zusätzlich anhand von einem der Installationsdatensätze weitere Informationen berechnen, wie beispielsweise "Wie hoch ist der Anteil der Nutzer, der die App behält?", "Wie hoch ist die Deinstallationsquote?" oder "Wie hoch ist die Differenz zwischen täglichen Installationen und Deinstallationen?". Das kann dann auch zwischen 2015 und 2016 verglichen werden. Außerdem, falls die Auswertung Visualisierungen enthalten sollte, die nicht nur aus einfachen Bildern bestehen (beispielsweise wenn mit d3 gearbeitet wird), muss das Auswertungsdokument diese Dateien auch darstellen können. Die Auswertung würde ich dann z.b. mit Sphinx machen und die Diagramme dort einbinden.

Was genau soll untersucht werden?

- In welchen Ländern wird die App am häufigsten installiert, in welchen am seltensten?
- In welchen Sprachen wird die App am häufigsten installiert, in welchen am seltensten?
- Welche App-Versionen verursachen die meisten Abstürze, welche die wenigsten?
- Auf welchen Geräten stürzt die App am häufigsten ab, auf welchen am seltensten?
- Bewertungen bezogen auf Länder
- Alles (soweit möglich) jeweils für 2015 und 2016 und dann vergleichen (was hat sich verändert?), zusätzlich:
    - Summe der Installationen pro Jahr vergleichen
    - Durchschnittsbewertung von 2015

    - Summe der Abstürze von 2015
- Zusätzlich:
    - Wie viel Prozent der Nutzer, die die App installiert haben, behalten sie auch?
    - Wie hoch ist die Deinstallationsquote?
    - Wie hoch ist die Differenz zwischen Installationen und Deinstallationen?
    - Beides für 2015 und 2016 berechnen und vergleichen

Erste Schritte:

- Die CSV-Dateien einlesen
- Für jede Datei eine Liste erstellen, in der alle Reihen abgespeichert werden, um alle Daten zu erfassen
- Für alle Spalten, mit denen gearbeitet werden soll, eine extra Liste erstellen
- Falls nötig weitere Listen definieren und für alle das erste Element (die erste Reihe) löschen, das sind die Spaltenüberschriften
- Als Strings eingelesene Zahlen in Integers umwandeln, damit mit ihnen gerechnet werden kann

In welchen Ländern wird die App am häufigsten installiert, in welchen am seltensten?

- Liste erstellen, die nur die Spalte total_installs enthält
- Liste erstellen, die die Länderabkürzungen enthält
- Funktion schreiben, die diese beiden Listen einliest, die fünf höchsten Werte der Installation heraussucht und direkt das Land mit abspeichert, damit es trotz neuer Sortierung dem Wert zugeordnet werden kann und eine neue Liste mit dieser Zuordnung ausgibt
- Das gleiche noch einmal, nur dass statt des höchsten Wertes der niedrigste herausgesucht wird (hier eventuell größeres Spektrum betrachten, falls viele Länder mit 0 Installationen dabei sind)
- Zur Veranschaulichung wird ein Diagramm erstellt, das die Länder je nach Installationshäufigkeit farbig darstellt. Dazu wird die Liste benötigt, die alle

Installationen enthält (falls nötig absteigend sortiert) und für das Diagramm eingelesen wird. Unsortiert sind die Länder leicht zuzuordnen, andernfalls wird die Funktion von oben (abgewandelt) genutzt. Wenn es funktioniert, würde ich die Visualisierung gerne mit einem Länderdiagramm von d3 machen.

- Jeweils für 2015 und 2016

<u>In welchen Sprachen wird die App am häufigsten installiert, in welchen am seltensten?</u>

- Ähnliches Vorgehen wie bei vorheriger Frage, nur dass statt den Länderabkürzungen die Sprachen den Werten zugeordnet werden
- Hier wird ein Histogramm für die 5 Sprachen in denen die App am häufigsten installiert wird erstellt und eines für die 5(?) in denen es am seltensten installiert wird
- Jeweils für 2015 und 2016

<u>Welche App-Versionen verursachen die meisten Abstürze, welche die wenigsten?</u>

- Hier sollen die App-Versionen aufsteigend sortiert werden und mithilfe der bereits erklärten Funktion die Abstürze zugeordnet werden
- Alle Werte können dann in einem Histogramm mit Verlaufskurve (seaborn) dargestellt werden, um zu zeigen, wie sich die Versionen auf die Abstürze auswirken. Wie sieht die Kurve aus? Stürzen neuere Versionen vielleicht öfter oder seltener ab als ältere? Oder gibt es gar keinen Zusammenhang?

<u>Auf welchen Geräten stürzt die App am häufigsten ab, auf welchen am seltensten?</u>

- Es werden die 5 höchsten und die 5(?) niedrigsten Absturz-Werte mithilfe der Funktion von oben gesucht und dem entsprechenden Gerät zugeordnet
- Ein Histogramm soll darstellen, welche 5 Geräte die meisten, und ein weiteres, welche Geräte die wenigsten Abstürze verzeichnen

<u>Bewertungen bezogen auf Länder</u>

- Ein weiteres Länderdiagramm, das farbig darstellen soll, wie die App in welchem Land bewertet wurde. Auch hier soll, wenn es funktioniert, ein Länderdiagramm von d3 benutzt werden.
- Dazu werden die einzelnen Bewertungen in Bezug zu der Länderbezeichnung eingelesen

<u>Summe der Installationen pro Jahr vergleichen</u>

- Mithilfe einer Funktion, die die Summe einer Liste berechnet, werden die Installationswerte jeweils von 2015 und von 2016 addiert
- In einem horizontalen Balkendiagramm (pygal) sollen beide Werte eingelesen werden und zeigen, wie groß der Unterschied in den beiden Jahren war

Durchschnittsbewertung 2015

- Mithilfe einer Funktion, die den Durchschnitt der Werte in einer Liste berechnet, werden die Durchschnittsbewertung von 2015 berechnet

Summe der Abstürze 2015

- Mithilfe einer Funktion, die die Summe einer Liste berechnet, werden die Abstürze von 2015 addiert

Wie viel Prozent der Nutzer, die die App installiert haben, behalten sie auch?

- Es werden die "total user installs" und die "current user installs" jedes Landes addiert, anschließend werden die beiden Werte in eine Funktion eingelesen, die den Prozentwert der "current user installs" berechnet.

Wie hoch ist die Deinstallationsquote?

- 100 minus den Wert, der vorher berechnet wurde. Das soll nur eine Zusatzinformation sein.
- Dadurch kann aber eine Art Kreisdiagramm erstellt werden, das den Anteil von Benutzern, die die App vehalten und Nutzern, die die App deinstallieren verdeutlichen soll
- Das soll für beide Jahre gemacht und dann verglichen werden

Wie hoch ist die Differenz zwischen Installationen und Deinstallationen?

- Es werden alle "daily user installs" und alle "daily user uninstalls" summiert. Das Ergebnis resultiert aus einem Bewertungszeitraum von 4 Tagen. Dann können beide Werte im Vergleich visuaisiert werden
- Beide Werte können dann ebenfalls in eine Funktion eingelesen werden, die die Prozentsätze berechnet, die dann ebenfalls in einem Kreisdiagramm im direkten Vergleich der Anteile visualisiert werden können
- Auch das kann jeweils für 2015 und 2016 gemacht und dann verglichen werden

Fazit

Zum Schluss sollen die Erkenntnisse noch einmal kurz zusammengefasst werden.

Bibliotheken, die auf jeden Fall benötigt werden:

- csv
- matplotlib
- numpy

Bibliotheken, die eventuell benutzt werden:

- bokeh
- seaborn
- pygal
- plotly
- geoplotlib
- pandas.DataFrame
- d3

Links die bei der Visualisierung helfen sollen:

Verschiedene Visualisierungstools mit Beispielen und Code:

http://pbpython.com/visualization-tools-1.html

https://www.dataquest.io/blog/python-data-visualization-libraries/

10 verschiedene Visualisierungs-Bibliotheken mit Code:

https://blog.modeanalytics.com/python-data-visualization-libraries/