

# Projektseminar I4 „Angewandte Informationswissenschaft“

Raphael Katschke

## TWEET-DATENANALYSE AUF HÄUFIGKEITEN, SCHLAGWORT- UND HASHTAGVERTEILUNG ZUM THEMA: WORLD OF WARCRAFT – LEGION RELEASE ZUM 30.08.2016

WOW-HYPE. WELCHE SCHLAGWORTE, HASHTAGS UND ERWÄHNUNGEN ZUM WOW-LEGION RELEASE WERDEN IN TWITTER AM HÄUFIGSTEN VERWENDET UND WELCHE KOMMEN BESONDERS HÄUFIG ZUSAMMEN VOR? LASSEN SICH TENDENZEN HERAUSFILTERN?

### • Bereits vorhanden (Aus Zusammenarbeit mit Thorsten Brückner):

- Twitter-Crawler in Python (Tweepy)
  - ◆ API-Access-Token
  - ◆ Daten extrahieren
  - ◆ Daten in eine .csv-Datei schreiben
- Daten aus einer .csv-Datei einlesen

### • To-Do:

#### ❖ Schritt 1:

- Crawler anpassen:
  - ◆ Key-Attributes anpassen, benötigt werden folgende Felder:
    - ID um Doppelverarbeitung zu verhindern
    - CREATED\_AT für eine Timeline
    - TEXT zur weiteren Bearbeitung d. Hashtags, Erwähnungen und Schlagworte
  - ◆ In 2 versch. Varianten ausführen:
    - Die letzten 3000 Tweets von gewissen Hashtags am 29.08.16
    - Stream von gewissen Hashtags am 30.08.16

#### ❖ Schritt 2:

- Daten verteilen
  - ◆ Doppelspeicherungen entfernen
    - Anhand des ID Feldes
    - ID in eine Liste speichern und beim Bearbeiten die Liste checken
  - ◆ Daten in entsprechende Listen und Dictionaries speichern

- TEXT in eine Liste für spätere Bearbeitung durch NLTK & Co
- CREATED\_AT in ein Dictionary für eine Timeline Häufigkeitsverteilung

### ❖ Schritt 3:

- Stoppwortliste
  - ◆ Auf die Liste mit den Tweettexten anwenden
- Schlagworte, Erwähnungen und Hashtags filtern und separieren
  - ◆ Schlagworte:
    - Top Worte der Häufigkeitsverteilung und ggf. manuell hinzugefügte Spielrelevante Worte.
- Alle Begriffe der Tweettexte nach der Stoppwortliste in Permutationslisten eintragen
  - ◆ Für spätere Clusteranalyse

### ❖ Schritt 4:

- Analyse:
  - ◆ Häufigkeitsverteilung (Top 25? Top 50?)
    - Schlagworte
    - Hashtags
    - Erwähnungen
  - ◆ Tendenzen herausarbeiten.
    - Informationell?
      - Supportanfragen, Informationsaustausch
        - Anhand gewisser Spielbezogener Schlagworte oder Hashtags
          - ◆ #Hunter, #Priest, Balance, OP
    - Emotionsaustausch?
      - **Minimum: Intellektuell Tendenzen der Top Schlagworte herausarbeiten**
      - **Optional: Sentimentanalyse anhand vorhandener tools und ggf. erweitern**
      - Positiv, Neutral, Negativ
        - Problematisch:



- ◆ Sarkasmus
- ◆ Für richtige Sentimentanalyse fehlt wahrsch. die Zeit
- Auch hier sind gewisse Begriffe nützlich:
  - ◆ Nerf / Buff / Overpowered (OP) / Underpowered (UP)
  - ◆ Sind oft mit Emotionen verbunden
- Wörterbuch schreiben
- Kontakt mit Entwicklern?
  - Anzahl d. Tweets in Relation zur Anzahl d. Erwähnungen stellen
    - Wie häufig wurde der Kontakt zu den Entwicklern gesucht?
- ◆ Clusterbildung – Itertools lib?
  - **Minimum: Top-Terme durch die Texte jagen und vorhandene Begriffe in Relation zu den Termen stellen – Clustergrafiken nach Größenordnung erstellen.**
  - **Optional: Alle vorhandenen Schlagworte/ Hashtags/ Erwähnungen in Verbindung miteinander bringen**
  - Welche Begriffe stehen mit den Top Schlagworten/ Hashtags/ Erwähnungen in verschiedenen Tweets in Verbindung?
  - Top-Terme durch die Listen jagen und Begriffe in Verbindung zählen

#### ❖ Schritt 5:

- Grafische Darstellung der Ergebnisse
  - ◆ Matplotlib? Plotly? Networkx?
    - Häufigkeitsverteilung von Schlagworten, Hashtags und Erwähnungen
    - Clusterbildung → Zusammenhängende Begriffe zu den Top-Termen
    - Tendenzen
    - Timeline
- Über PHP die Ergebnisse auf Homepage anzeigen lassen
  - ◆ Falls genug Zeit übrig ist

