Favourite Counts of Tweets of Donald Trump

Felix Herron

TU Berlin

15.5.2020

Table of Contents

Introduction

- Donald Trump ist US-Amerikanischer (bzw. mein) Präsident #notMyPresident
- Seit Kontoeröffnung beinah 50.000 mal getweeted
- Einer der aller größten Twitterpersönlichkeiten [1]
- Ein geiles Thema für die Bachelorarbeit

[1] https://www.brandwatch.com/blog/most-twitter-followers/

Introduction

- Donald Trump ist US-Amerikanischer (bzw. mein) Präsident #notMyPresident
- Seit Kontoeröffnung beinah 50.000 mal getweeted
- Einer der aller größten Twitterpersönlichkeiten [1]
- Ein geiles Thema für die Bachelorarbeit

Ziele

1. Fav-count untersuchen/vorhersagen (heutiges Thema)
2. Mit topics verknüpfen (noch in Bearbeitung)
3. Feedback von euch sammeln :)

[1] https://www.brandwatch.com/blog/most-twitter-followers/

Table of Contents

Tables and Figures

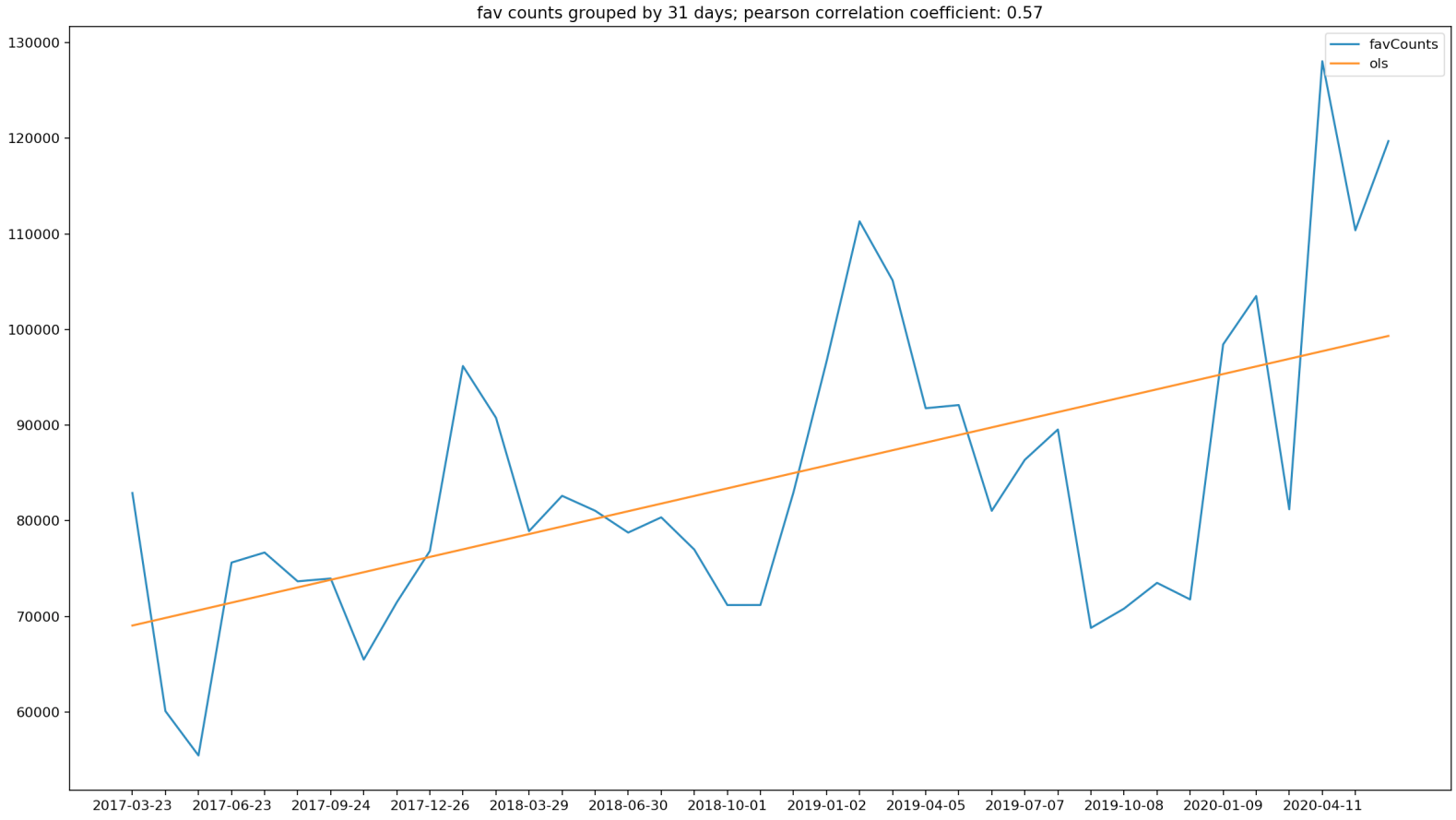
1. Tweets heruntergeladen, in eine DB geladen
2. Tweets gecleanet:
   - “RT”, links, “#”, u.a. entfernen
   - lemmas behalten

Tables and Figures

1. Tweets heruntergeladen, in eine DB geladen
2. Tweets gecleanet:
   - “RT”, links, “#”, u.a. entfernen
   - lemmas behalten
3. Mit verschiedenen Paramtern beschriftet:
   - isRT, isReply, isDeleted, mediaType, publishTime ...
   - n-Grams (1-4)

FavCount x Zeit

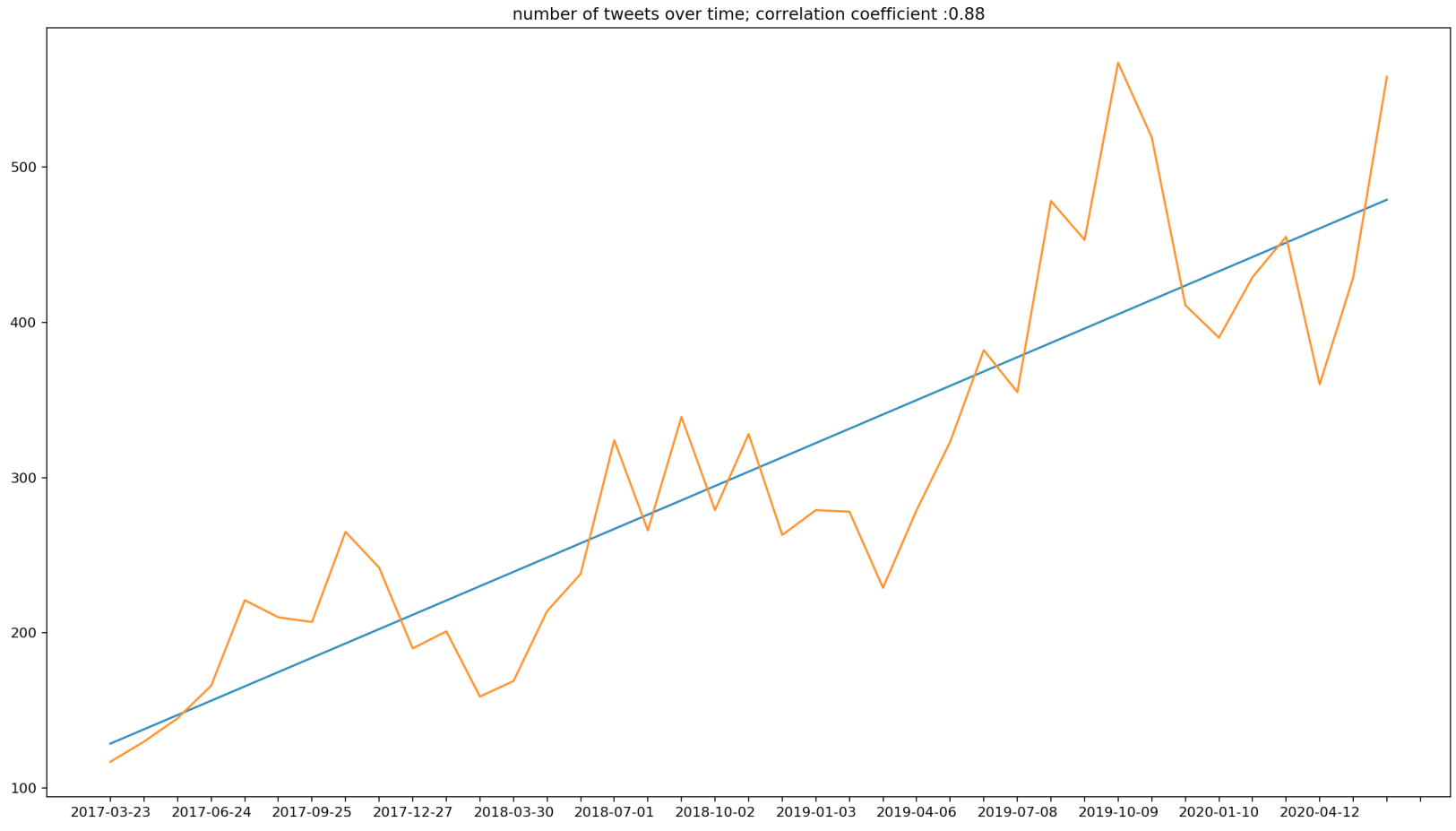
Die FavCounts von Trump steigen mit der Zeit



=> Bei der Vorhersage muss das in Betrachtung kommen... mehr dazu später

TweetCount x Zeit

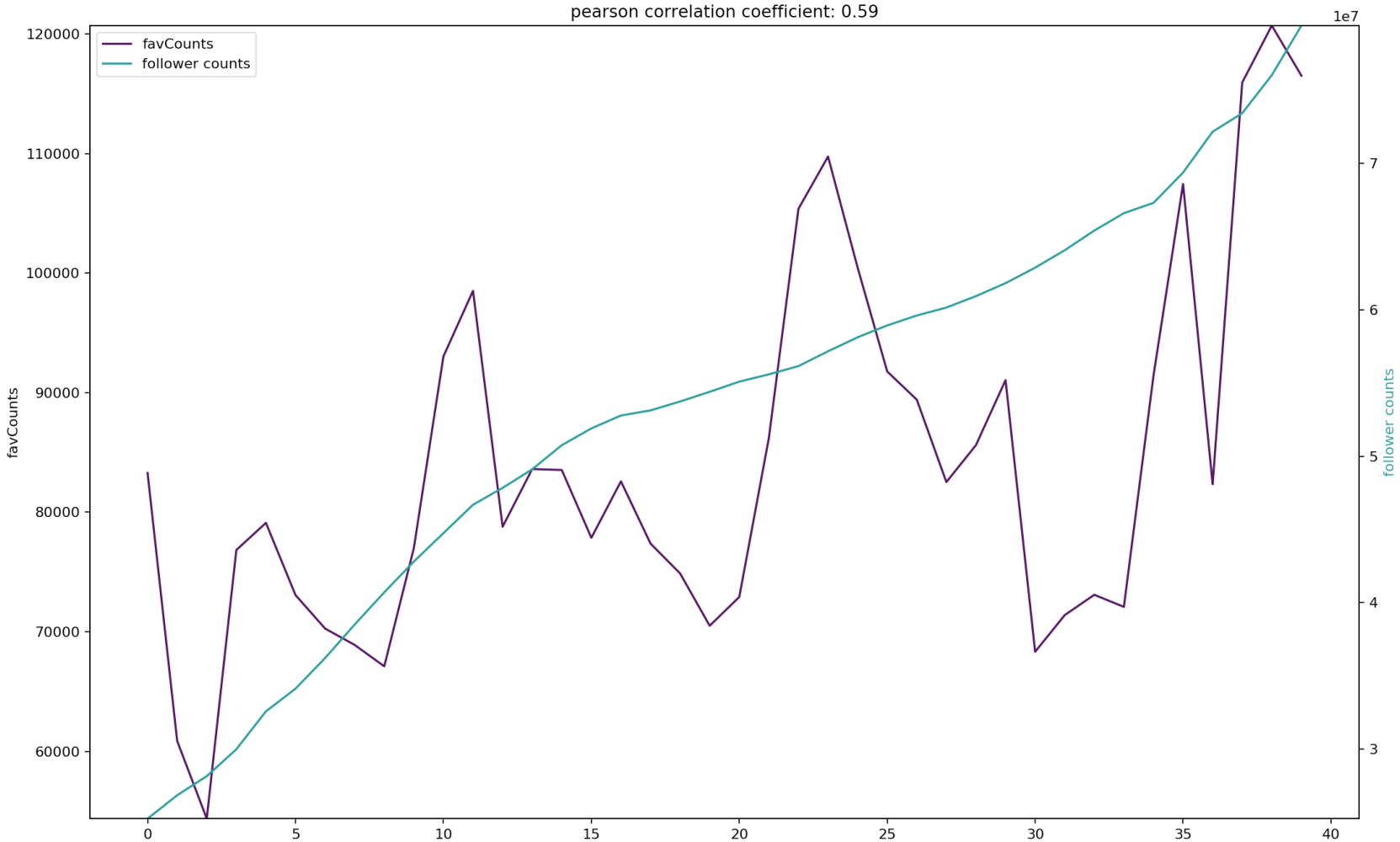
Die Anzahl an Tweets von Trump steigen mit der Zeit



=> warum?

FavCount x FollowerCount (bzw. Zeit)

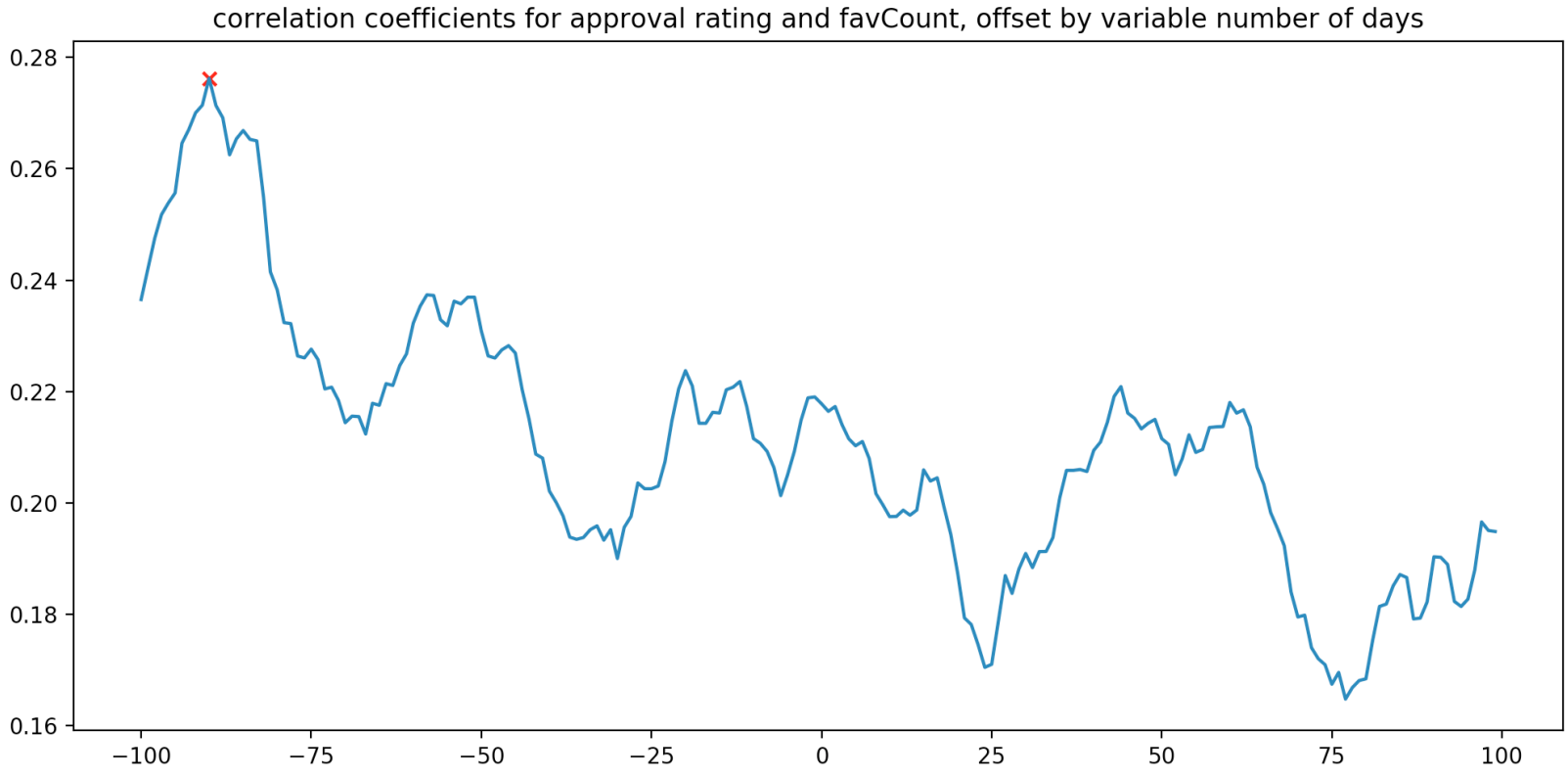
Die Anzahl seiner Follower steigt stetig mit der Zeit



=> Verknüpfung FollowerCount/FavCount? Unklar...

Approval Rating x Zeit

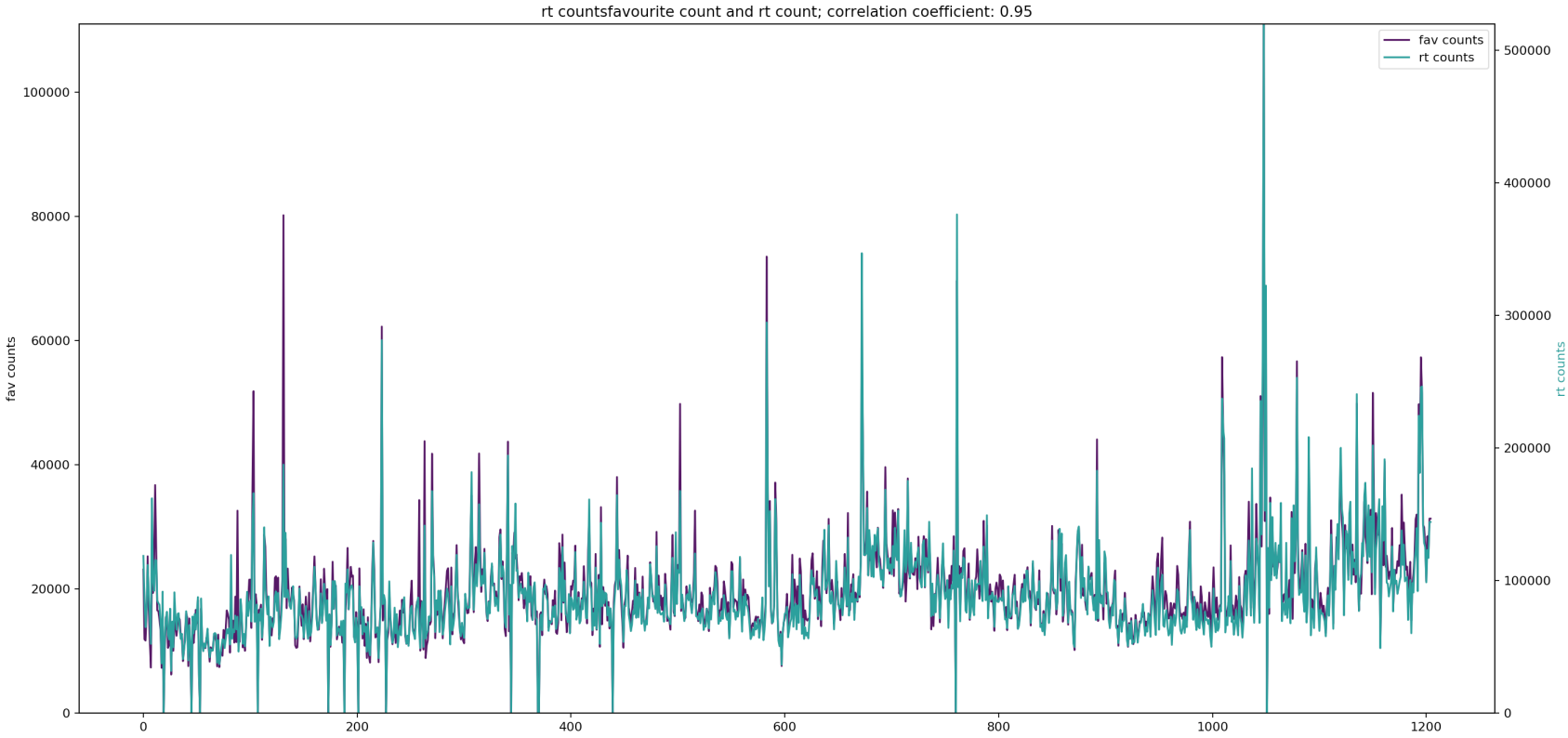
Die Approval Rating von Trump steigt schwach mit seiner FavCount



=> Wahrscheinlich Zufall

FavCount x RTCount

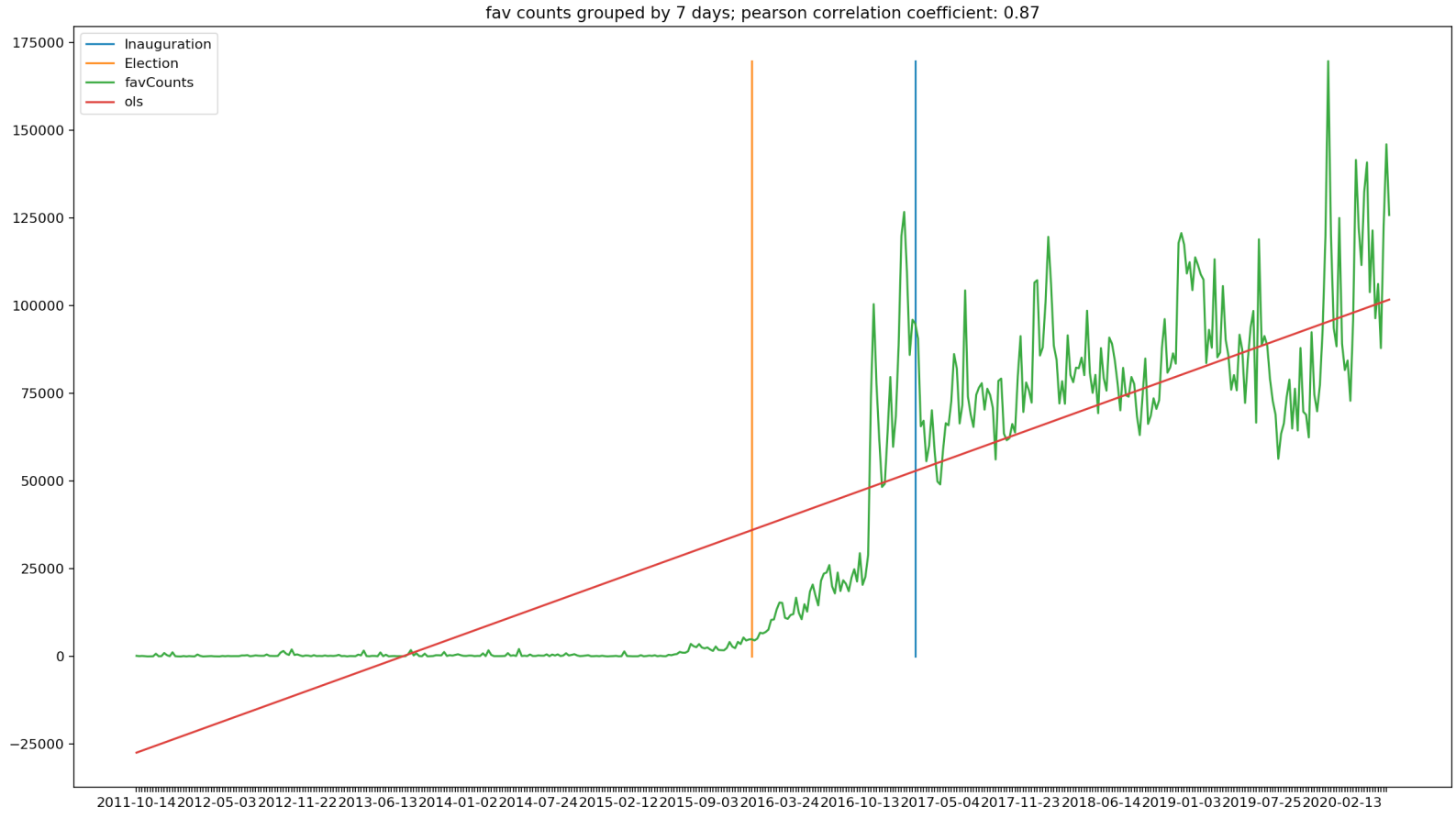
Fav Count $\overset{\propto}{\sim}$ RT Count



=> FavCount und RT Count sind quasi ein Feature

FavCount x Zeit (Präsident)

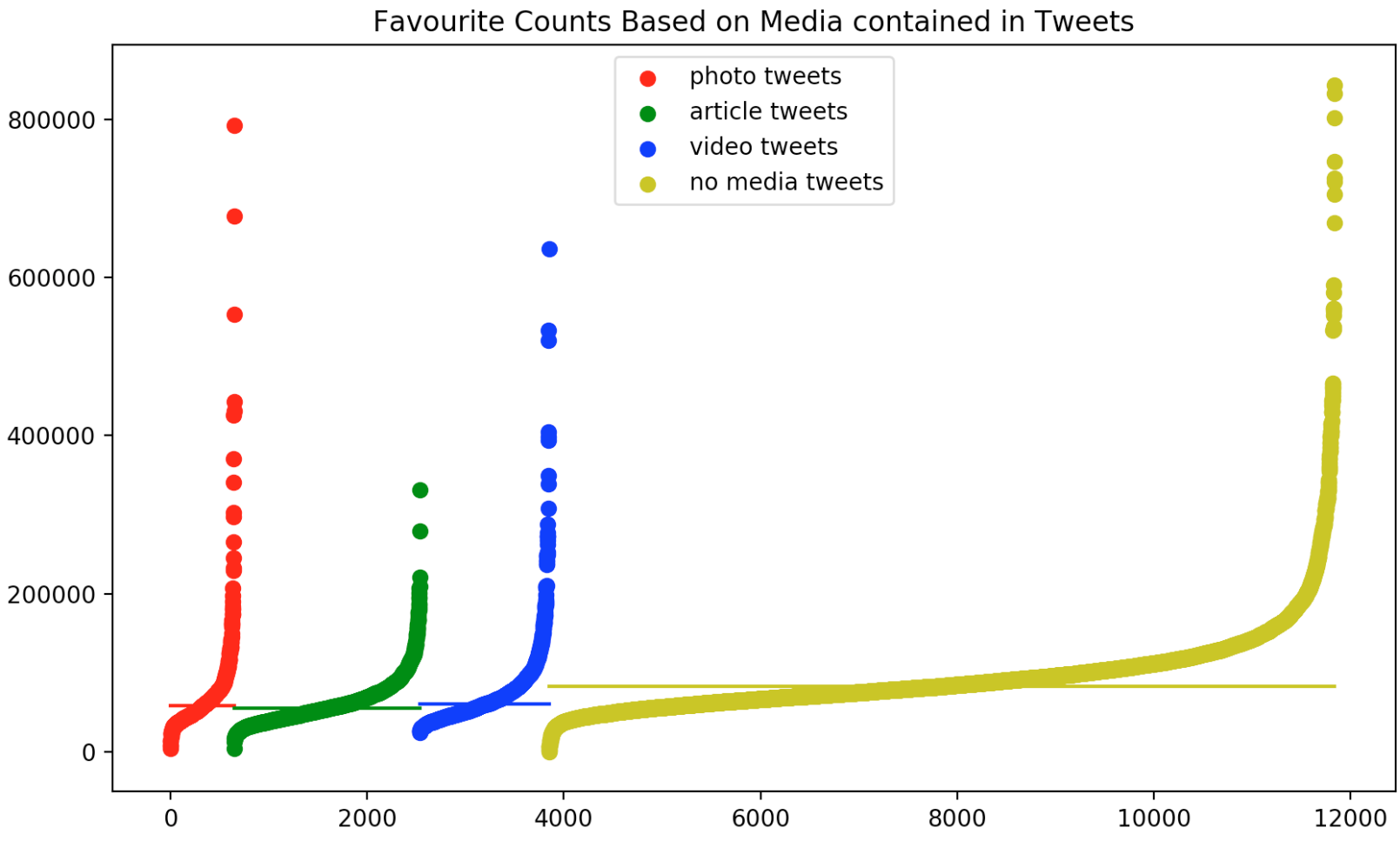
Die Reaktion auf seine Tweets hat sich seit seinem Amtsantritt wesentlich verändert



=> ich betrachten nur die Tweets seit Amtsantritt

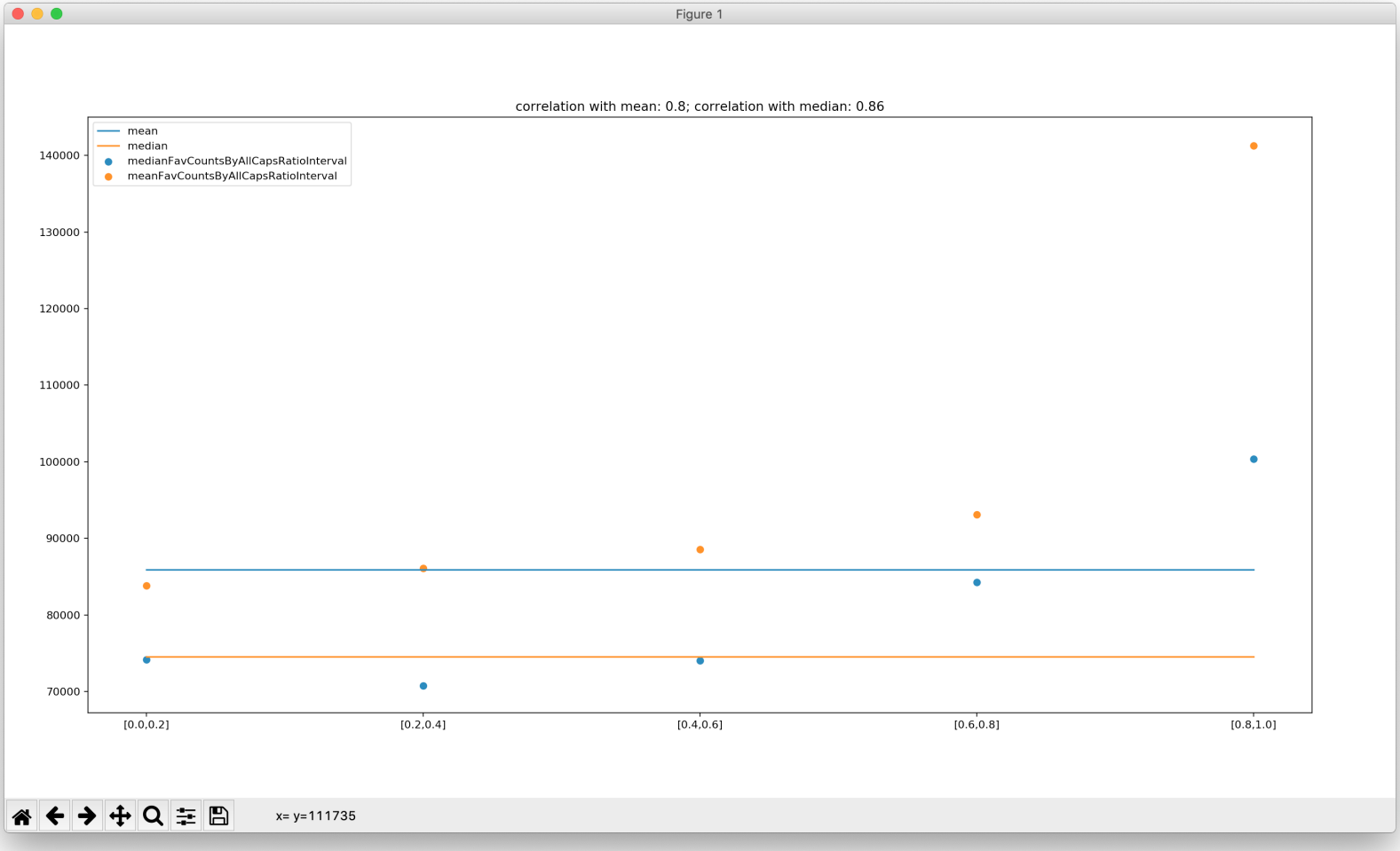
FavCount x MediaType

Die Art der beigefügten Medien variiert leicht mit der FavCount



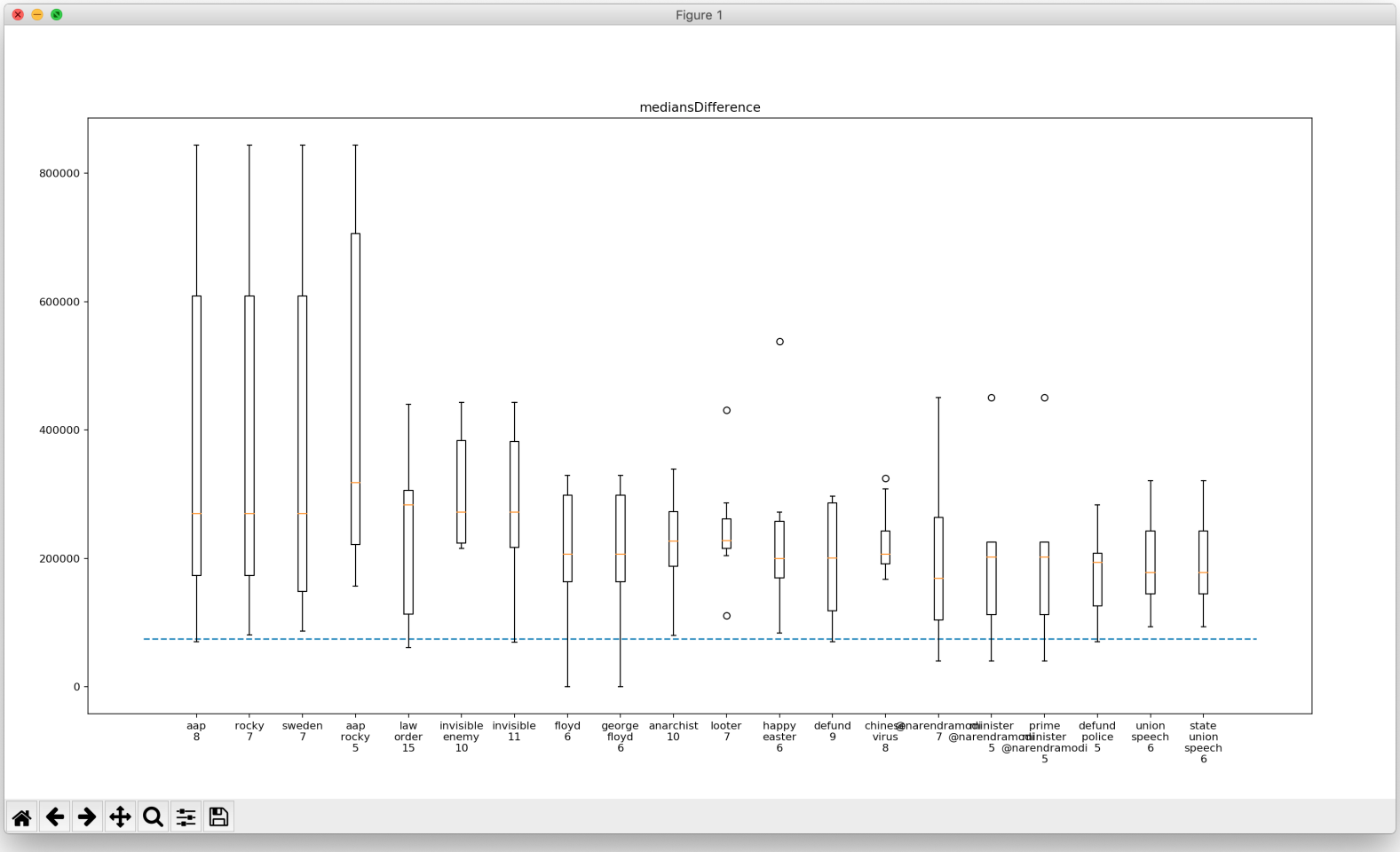
FavCount x allCaps Percentage

Die Anzahl an groß geschriebene Wörter hat einen starken Einfluß auf die Popularität



FavCount x nGrams Used

Die Wortwahl der Tweets hat ein Effekt auf deren FavCount



usw.

und so weiter und so weiter...[2]

[2] LADY HARRIET DURHAM

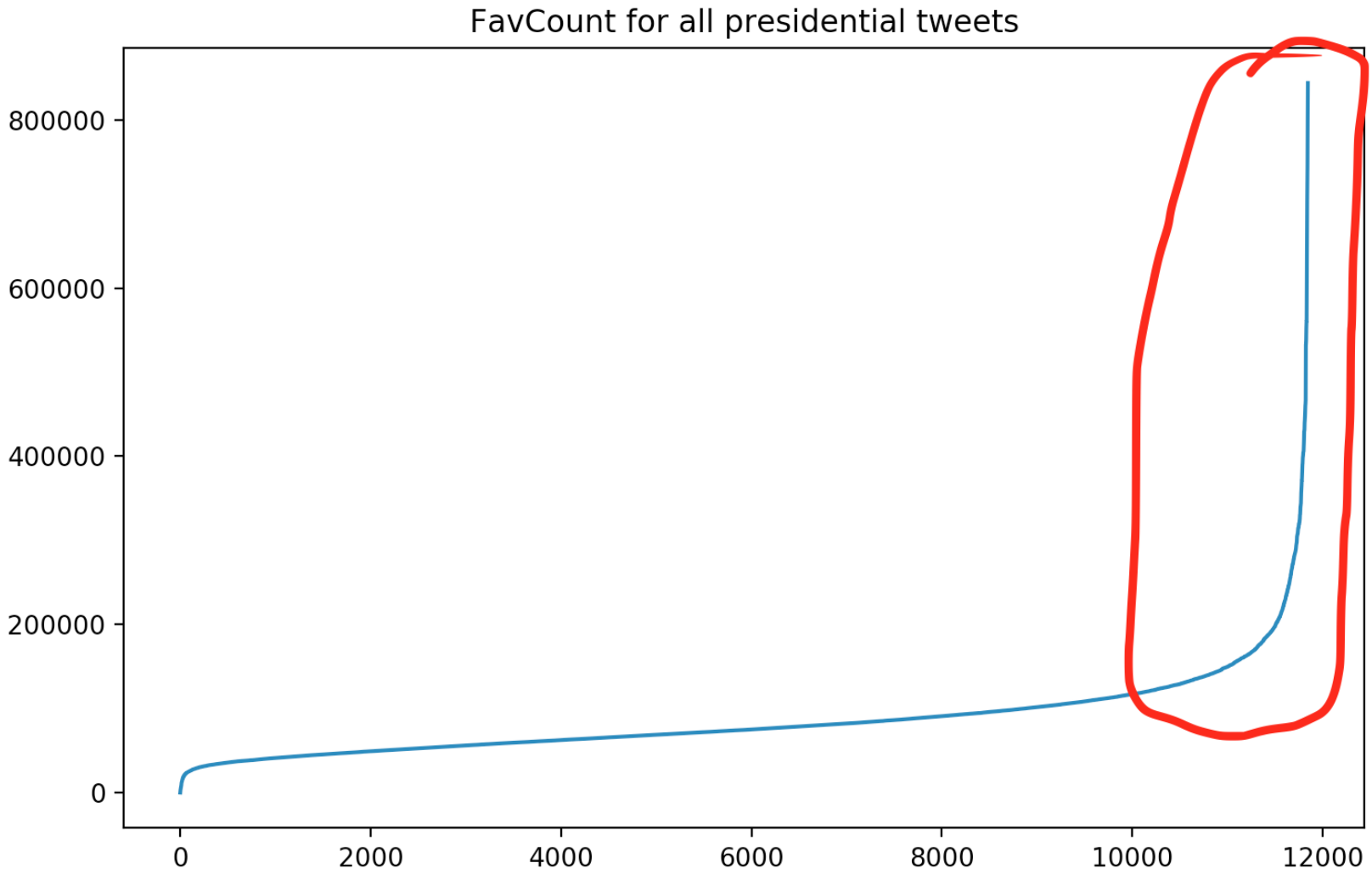
Table of Contents

Warum Classification Eigentlich?

- Eigentlich wäre Regression richtig
- Zu wenige Daten => Vorhersagen wären unpräzis
- Unpräzise Regression $\approx$ Klassifikation

Viral Classification Motivation

- Weitere Trennungen kommen später
- Z.B. Viral/Dud
- Hier: Flache Mitte, Viral Tail



N-Gram Classifier

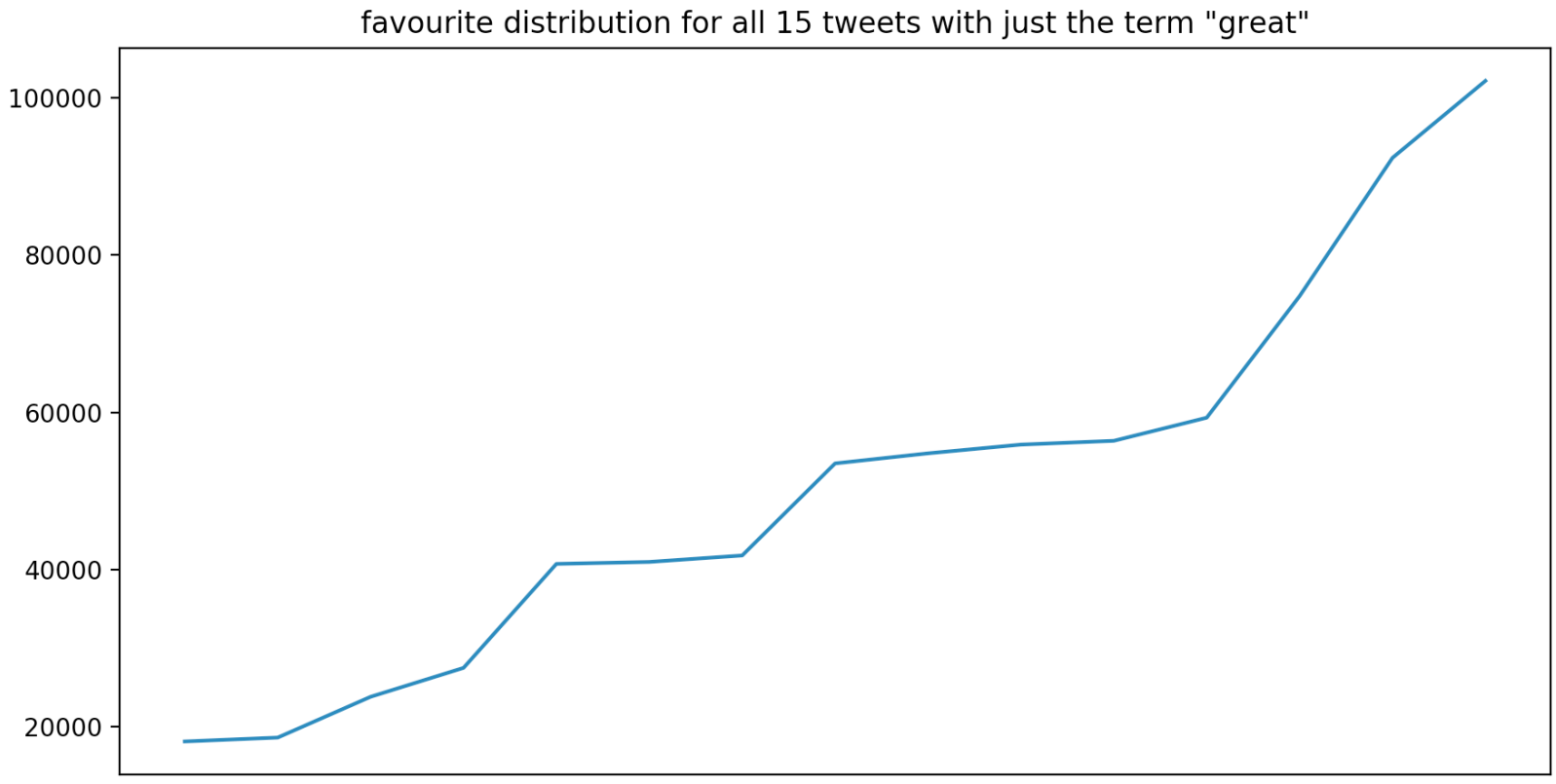
Erstmal aber Popular/Unpopular

- Idee: für jedes (relevante) N-Gram in einem Tweet, bestimme erwartete Skew
- Summiere (gewichtet) über alle Skews)

$$prediction(t) = c \sum_{nGram \in t} skew_{nGram} \cdot \sigma(favs_{nGram})$$

N-Gram Classifier Nachteile

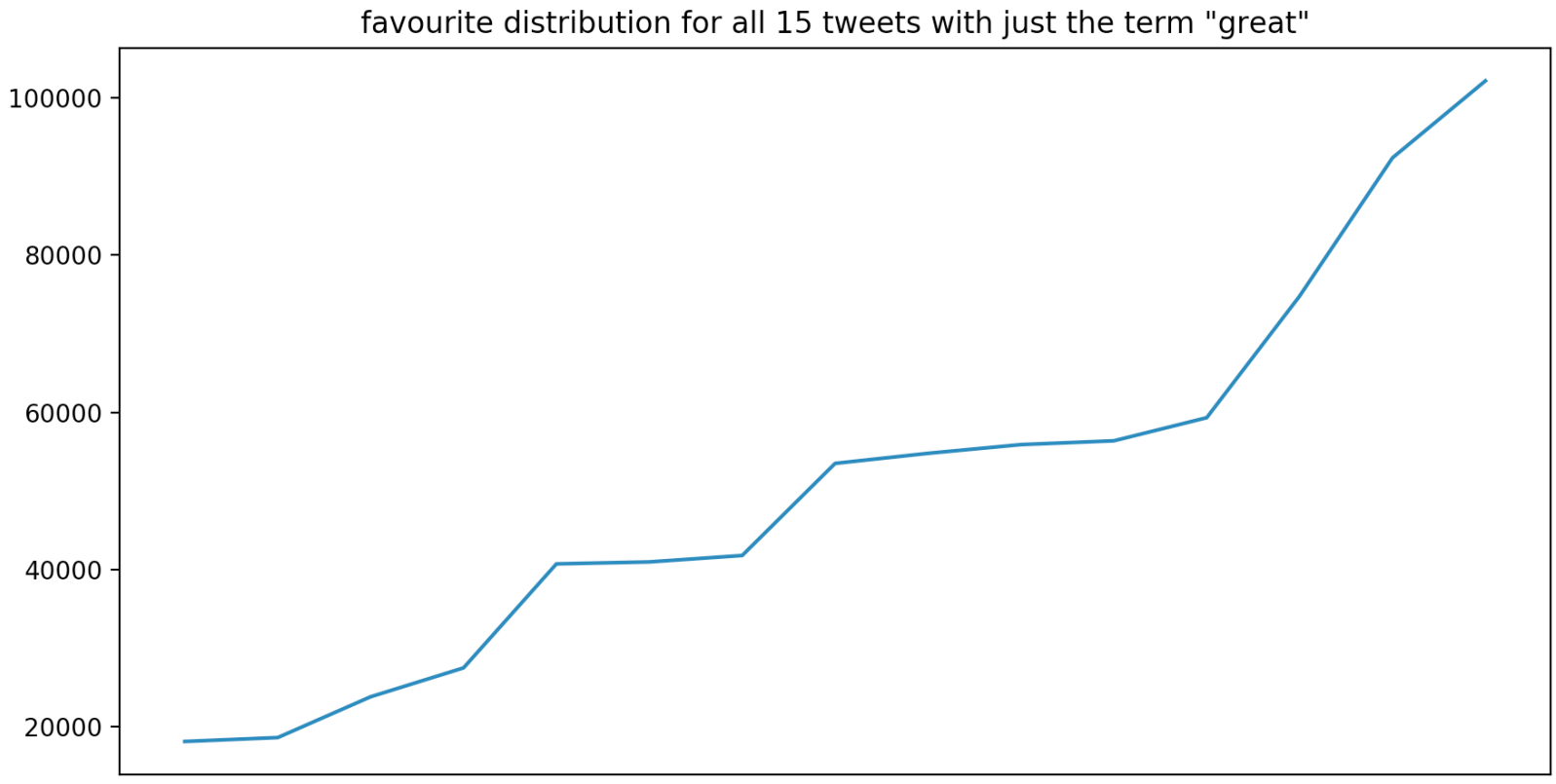
Viele Tweets haben sehr wenige Wörter



Wie unterscheidet man zwischen denen?

N-Gram Classifier Nachteile

Viele Tweets haben sehr wenige Wörter



Wie unterscheidet man zwischen denen?

Hyper- bzw. weitere parameter, z.B.

Weniger Interessante Versuche

Naive Bayes Classifier:

- Gleiche Daten, andere Perspektive als bei Boosting
- 65 % Genauigkeit, viel Overfitting

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Weniger Interessante Versuche

Naive Bayes Classifier:

- Gleiche Daten, andere Perspektive als bei Boosting
- 65 % Genauigkeit, viel Overfitting

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

OLS:

- Jeder Tweet bekommt sparsen Eingabevektor, -1 oder 1 als Target
- 68 % Genauigkeit, viel Overfitting, Hyperparameter müssen noch untersucht werden

Immerhin Fehler auf nicht den selben Tweets

Boosting/Classifier Committee

- Es gibt viele Merkmale, nicht nur auf n-Gram-skew basiert, die Tweet Populärität vorhersagen
- Allerdings nicht so stark
- Sie machen Fehler bei unterschiedlichen Tweets

[3] https://web.stanford.edu/ hastie/Papers/buehlmann.pdf

Boosting/Classifier Committee

- Es gibt viele Merkmale, nicht nur auf n-Gram-skew basiert, die Tweet Populärität vorhersagen
- Allerdings nicht so stark
- Sie machen Fehler bei unterschiedlichen Tweets

=> Classifier, der Ergebnisse aus vielen schwachen Classifiern zusammenträgt[3]

[3] https://web.stanford.edu/ hastie/Papers/buehlmann.pdf

Boosting/Classifier Committee p. 2

Z.B.:

- n-Gram median discrepancy classifier sagt -1 mit geringer Sicherheit
- all-Caps Classifier sagt 1 mit hoher Sicherheit
- ...
- Prediction: 1

Ergebnisse aus diesem Ansatz

- 72% test Genauigkeit
- Noch nicht viele Classifier involviert
- Verbesserungsbedürftigkeiten bei umfangreicherem Zusammentun
- Gewichtungen lernen!

Ergebnisse aus diesem Ansatz

- 72% test Genauigkeit
- Noch nicht viele Classifier involviert
- Verbesserungsbedürftigkeiten bei umfangreicherem Zusammentun
- Gewichtungen lernen!

Wichtig: Abspaltung von Stärken/Schwächen

Table of Contents

Die Hürde

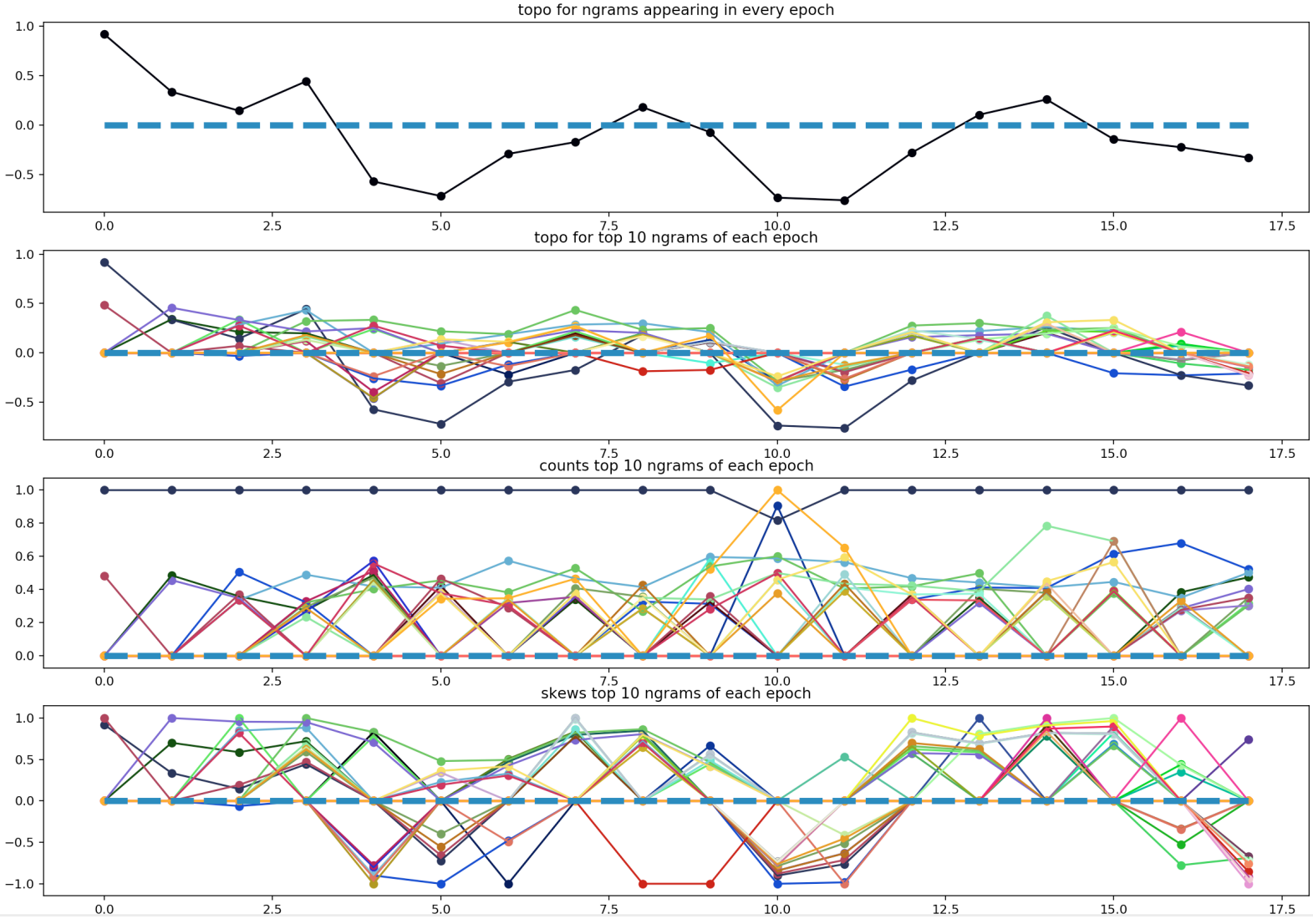
"It's no use going back to yesterday, because I was a different person then"[4]

- $reaktion(nGram_1, time_1) \neq reaktion(nGram_1, time_2)$
- Wörter bekommen veränderte Signifikanz/Geschmack
- Z.B.: Great beginnt als positives Wort, am Ende ist es negativ
- Verbleibt häufig genutzt

[4] Lewis Carroll

Veranschaulichung des Fluches

Schützen Sie Sich die Augen!



Das eventuelle Elixier

- Zeit-lokalisiert trainineren
- Wird er fähiger?
- Bzw. veränderte Wortwahl $\implies$ favCount $\uparrow$?

Das eventuelle Elixier

- Zeit-lokalisiert trainineren
- Wird er fähiger?
- Bzw. veränderte Wortwahl $\implies$ favCount $\uparrow$?

=> Das führt zu...

Table of Contents

Die Hürde

- Er hat seit seinem Amtsantritt nur etwa 12.000 mal getweetet
  - NLP Anwendungen aus der Forschung haben viel größere Corpora
- Tweets sind kurz
  - “bis auf die Knochen abgemagert” Dokumente
  - Viele sind Unschlüssig
  - Je raffinierter bei der Dokumentenwahl, desto wenigere Daten!

Das eventuelle Elixier

Daten aus den Replies mitnutzen.

- Seeehr viele Leute antworten auf seine Tweets
- Die Inhalte dieser Antworten könnten viel über den Tweet enthüllen

Table of Contents

Analyse der Antworten

- Großer Datensatz
- Algorithmen anwendern, die zuvor ausgeschlossen waren
- Z.B. Word-embeddings weitertrainienen
- Z.B. Topic Modelling
- => Aktuelle Forschung brauchbarer

Topic/Sentiment Analysis

- Tweets nach Sentiment/Topic deren Antworten filtern
- Daraus FavCount vorhersagen?
- Hänchen und Ei...

das wars

Vielen Dank, dass Sie zugehört haben

Und Danke an Stephi für die bisherige Betreuung!

das wars

Vielen Dank, dass Sie zugehört haben

Und Danke an Stephi für die bisherige Betreuung!

Ich freue mich über jeden Vorschlag :)