Goethe Universität
Institut für Informatik
Texttechnologie

SoSe 2023

NLP-gestützte Data Science

Übung 3

Manuel Stoeckel Prof. Dr. Alexander Mehler

Frist: 13. Juli 2023

Formalia

Die Übungen der Veranstaltung *NLP-gestützte Data Science* dienen der Vertiefung der in der Vorlesung behandelten Themen und einem praktischen Einblick in die diskutierten Probleme und Ihrer Lösungen. Es wird insgesamt **drei** Übungsblätter geben, mit denen Sie bis zu **150** Punkte erreichen können, welche Ihnen bei der Modulabschlussprüfung zu **einem Zehntel** als **Bonuspunkte** angerechnet werden. Ziehen Sie hierzu die entsprechenden Abschnitte Ihrer Prüfungsordnung zu rate.[1]

Allgemein gilt für Abgaben:

Die Abgabe erfolgt im **PDF Format** via OLAT. Bitte stellen Sie sicher, dass **Ihr Name und Ihre Matrikelnummer** auf jeder Abgabe vermerkt sind. Wir empfehlen Ihnen die Verwendung von LaTeX zur Erstellung Ihrer Abgabedokumente. Dazu liegt ein Template in OLAT bereit. Es gelten die gängigen Regeln für Plagiate: stellen Sie sicher, dass Sie Ihre Abgaben selbstständig erstellt haben und etwaige Fremdinhalte entsprechend gekennzeichnet und korrekt zitiert haben.

Für Programmieraufgaben gilt gesondert:

Neben der Abgabe Ihrer Ergebnisse im geforderten Format, ist zusätzlich der gesamte Quellcode ZIP-komprimiert einzureichen. Bitte achten Sie darauf, dass Sie **keine Binaries oder Bibliotheken** mit abgeben (z.B. `virtualenv`, Python Byte-Code, `git`-Repositories, etc.). Der Quellcode ist zu kommentieren. Fremdcode ist entsprechend zu kennzeichnen, auch hier gelten die gängigen Regeln für Plagiate. Wir empfehlen die Verwendung von `conda` zur Umgebungsverwaltung, z.B. mit `miniconda`. Bitte verwenden Sie eine Version von Python ≥ 3.10.

Bei Gruppenabgaben gilt gesondert:

Der individuelle Beitrag jedes Gruppenmitglieds muss aus der Abgabe ersichtlich sein.

[1] §36 Abs. 6 Satz 2 der Ordnung für den Bachelorstudiengang Informatik bzw. §35 Abs. 5 Satz 2 der Ordnung für den Masterstudiengang Informatik, jeweils in der Fassung vom 17. Juni 2019.

Übung 3: Online Information Propagation & Biases in NLP **max. 50 P**

3.1 Online Information Propagation **30 P**

Im Kapitel 09 „Online Information Propagation“ haben wir die *Kaskaden-artige Manifestation* von Tweets kennengelernt. Dabei wurden **vier** Metriken für solche Kaskaden vorgestellt:

- Kaskadentiefe
- Kaskadengröße
- Kaskadenbreite
- Viralität

Diese Kaskaden wurden nur explizit für Retweets eingeführt; hier wollen wir Sie auf Konversationen erweitern. Das heißt: alle Arten von Referenzen (Retweets, Replies & Quotes).

In den Materialien für diese Übung finden Sie Tweets im `JSON`-Format, die über die offizielle API von Twitter[2] heruntergeladen wurden. Jede `JSON`-Datei beinhaltet Tweets zu einem bestimmten Hashtag.

- Finden Sie für jede gegebene Datei die **fünf größten** Konversationen.
- Finden Sie für jede gegebene Datei die **fünf Wurzel-Tweets** der Kaskaden mit der
  - größten Kaskadentiefe,
  - größten Kaskadengröße,
  - größten Kaskadenbreite und
  - höchsten Viralität.
- Sammeln Sie die Antworten auf die obigen Fragen als Tweet- bzw. Konversations-IDs und reichen Sie sie mit Ihrer Abgabe als separate `JSON`-Datei ein, die folgendes Format hat:

```
[
  {
    "hashtag": "some_hashtag",
    "biggest_conversations": [{ "conversation_id": "123456789", "size": 123 }],
    "biggest_cascade_depth": [{ "tweet_id": "123456789", "value": 123 }],
    "biggest_cascade_size": [{ "tweet_id": "123456789", "value": 123 }],
    "biggest_cascade_breadth": [{ "tweet_id": "123456789", "value": 123 }],
    "biggest_cascade_virality": [{ "tweet_id": "123456789", "value": 0.123 }]
  },
  ...
]
```

Hinweise

- Sie können Graph-Algorithmus-Bibliotheken verwenden, z.B. `networkx`.[3]
- Bitte achten Sie darauf, dass manche Konversationen nicht vollständig enthalten sind.
- Achten Sie auch darauf, dass Kaskaden stets mit einem Wurzel-Tweet beginnen sollten, der keine Referenz auf einen anderen Tweet ist.
- Da es sich hier um echte Tweets handelt, können wir keine Garantien für deren Inhalte geben. Es können durchaus anstößige oder unangebrachte Tweets enthalten sein. Die Themen der Tweets wurden über die ”trendenden” Hashtags zum Download-Zeitpunkt bestimmt und nicht weiter eingeschränkt.
- Sie können das Format Ihrer Abgabe mit dem der Aufgabe beiliegenden `JSON`-Schema `ex3-json.schema` validieren.

[2] https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet
[3] https://networkx.org/documentation/stable/reference/index.html

3.2 Measuring Bias in Contextualized Word Representations **20 P**

In ihrem Paper greifen Kurita et al. (2019) den Bias-Test aus Caliskan et al. (2017) auf und wenden in auf kontextualisierte Sprachmodelle an. Dazu berechnen sie für eine Reihe an Template-Sätzen, ob ein positives bzw. negatives Attribut eher männlichen oder weiblichen Bezeichnungen zugeordnet wird, in den Sie die Ersetzungswahrscheinlichkeiten der Attribute für ein `[MASK]` Token miteinander vergleichen.

- Lesen Sie die Paper von Caliskan et al. (2017) und Kurita et al. (2019).
- Erstellen Sie die Template-Sätze wie in Tabelle 2 auf Seite 168 von Kurita et al. beschrieben.
- Schreiben Sie Code, der den *log probability bias score* berechnet.
- Schreiben Sie Code, der den *statistischen Signifikanzwert* p und die *Effektgröße* d (analog zu *WEAT*) für die *log probability bias scores* berechnet, wie in den Abschnitten 2 und 3 sowie Tabelle 3 von Kurita et al. (2019) beschrieben.
- Dokumentieren und interpretieren Sie Ihre Ergebnisse.

Hinweise

- Die Liste der Templates können Sie Kurita et al. (2019, 168) entnehmen.
- Die Liste der Attribute können Sie Caliskan et al. (2017, 3ff.) oder dem Code von Kurita et al. (2019) entnehmen.
- Achten Sie darauf sicherzustellen, dass Sie alle Wörter, für die Sie den Test durchführen, auch im als einzelnes Token im Vokabular des gewählten Transformer-Modells vorkommen!
- Verwenden Sie das `bert-base-uncased` Modell, um Ihre Ergebnisse mit denen von Kurita et al. (2019) vergleichen zu können.
- Lesen Sie das **ganze** Paper (Kurita et al., 2019) gründlich (insb. zur Vermeidung von Unklarheiten bezüglich der *log probability bias score* Berechnung sowie der Berechnung von Effektgröße und statistischer Signifikanz).

Bonus: WEAT for BERT **5 P**

Führen Sie die *WEAT for BERT* Experimentvariante aus Kurita et al. (2019) durch.

- Berechnen Sie die kontextualisierten Embeddings mit BERT wie von Kurita et al. (2019, §3.1) beschrieben.
- Berechnen Sie den *statistischen Signifikanzwert* p und die *Effektgröße* d für *WEAT for BERT*.
- Dokumentieren und interpretieren Sie Ihre Ergebnisse.

Literatur

Caliskan, Aylin, Joanna J. Bryson und Arvind Narayanan (2017). „Semantics derived automatically from language corpora contain human-like biases“. In: *Science* 356.6334, S. 183–186. DOI: 10.1126/science.aal4230. eprint: https://www.cs.princeton.edu/~arvindn/publications/language-bias.pdf. URL: https://www.science.org/doi/abs/10.1126/science.aal4230.

Kurita, Keita et al. (Aug. 2019). „Measuring Bias in Contextualized Word Representations“. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, S. 166–172. DOI: 10.18653/v1/W19-3823. URL: https://aclanthology.org/W19-3823.