Abschlusspräsentation

Exploration of News

Supervised by: Dr. Julian Kunkel

Tatyana Galitskaya, Sara Yüksel, Alexander Spikofsky

19.03.2018

Inhalt

1. Ziel

- Ausarbeitung der Eigenschaften und Vergleich von US-amerikanischen, britischen und australischen Newsartikeln

  - Fokus auf internationale Rubriken in englischer Sprache
  - Eigenschaften:
    - Schreibstil
    - Metadatenanalyse (z.B. Textlänge, Worthäufigkeit etc.)
    - Ähnlichkeitsanalyse

2. Tools/Knowledge

- Python
- Jupyter Notebook
- Crawler von Max + Beautiful Soup für Textextraktion
- Numpy und Pandas für Dataframes
- NLTK für Natural Language Processing
- ScikitLearn für verschiedene Textanalysemethoden
- Matplotllib.pyplot und Seaborn für Visualisierungen
- IPython.core.debugger
- …

3. Methodik

- Vorbereitung:
  - Erstellung RSS-Feed Liste, Artikel crawlen, Extraktion des Reintextes aus HTML im CSV-Format mit BeautifulSoup
- Preprocessing der Texte:
  - Füllwörter und Satzzeichen entfernen, Lemmatizing
- Analyse der Metadaten:
  - z.B. Anzahl der Artikel pro Paper, Durchschnittslänge von Wörtern, Part of Speech, Unique Words…
- Untersuchung der Ähnlichkeit von Artikeln:
  - z.B. Cosinus-Distanzen, euklidische Distanzen, MDS…

Allgemeine Informationen

UK		AU		US	
Daily Express	6.167	ABC Australia	4.237	ABC News	13.806
Daily Mail	30.541	Canberra Times	64	BBC US	1.525
The Guardian	10.015	News Australia	2.535	CBS	7.062
The Independent	6.882	The Advertiser	4.393	CNN	1.420
		The Daily Telegraph AU	2.499	Fox News	18.496
		The Mercury	28.400	NPR	4.669
		The west australian	4.909	New York Times	4.698
				Reuters	9.067
				Washington Post	18.505
				(mehrere Rubriken)	
Total:	53.605 Artikel	Total:	47.037 Artikel	Total:	79.248 Artikel

Total: 179.890 Artikel

Artikelzeitraum: 10.11 – 09.01 und 17.01 - 13.03

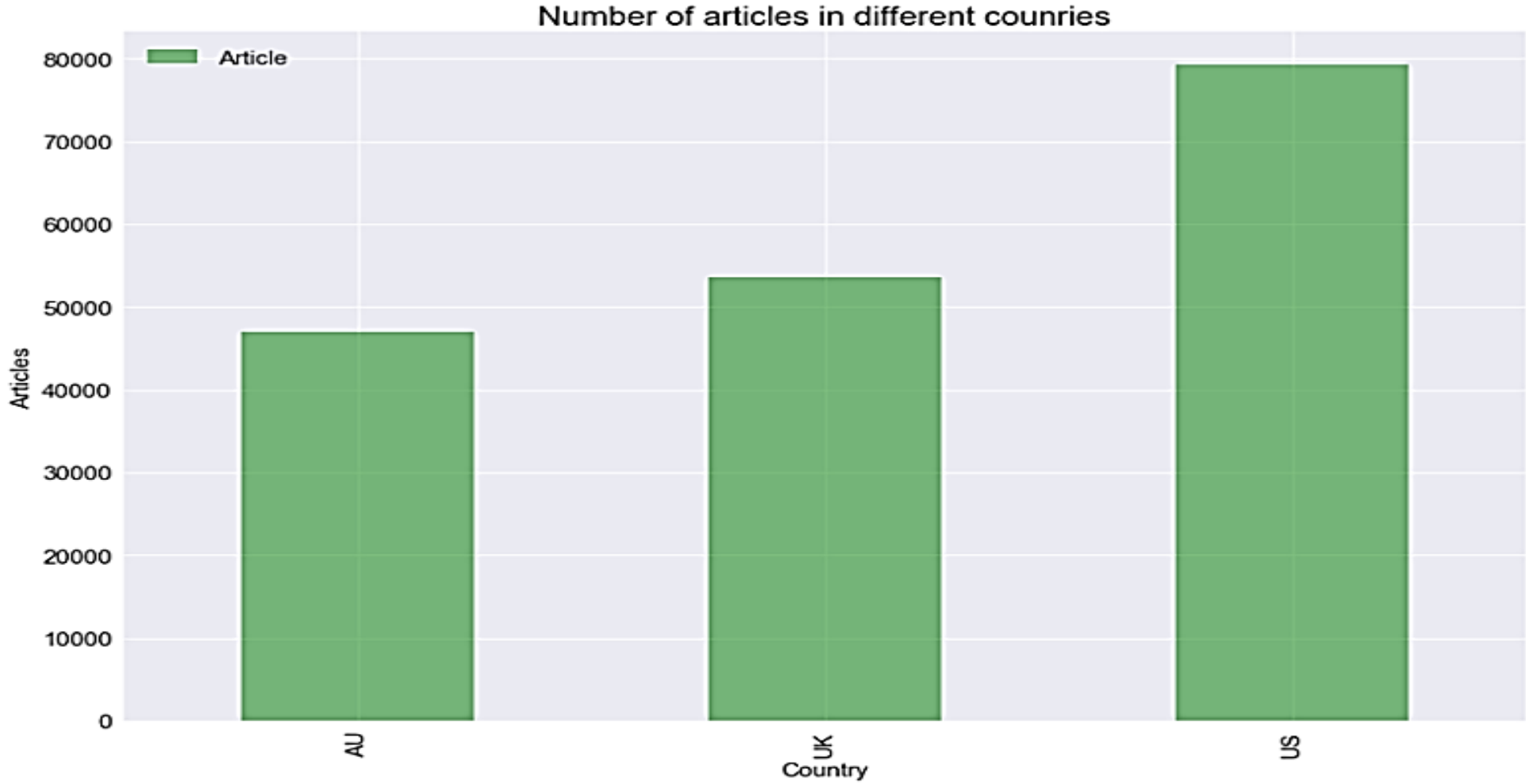
Beispiel bereinigter Artikel

NAIROBI, Kenya – The European Union is suspending funding for a water conservation project in Kenya after a tribesman was shot and killed by government forest service guards and another seriously wounded, the EU announced Wednesday. The shooting took place Tuesday in the Embobut Forest and came after the bloc of European countries warned it would suspend the $35 million project if force was used against locals, the European Union's ambassador to Kenya, Stefano Dejak, said. The conservation project was designed to protect land in the Mount Elgon and the Cherangani Hills areas of Kenya. Known as water towers, the terrain stores rainwater and enables regular river flows, among other benefits. "The EU insists on full respect for the rights of indigenous people, and the conservation work on the water towers was never expected to involve any evictions or use of violence," the bloc said in a statement. Staff members have been following up on reports that started coming in more than a year ago concerning abuses in the conservation areas and claims allegedly tying the evictions of Sengwer tribespeople to the EU's financial support, the EU said. The government-funded Kenya National Commission on Human Rights says forest guards have been carrying out forceful evictions of forest tribes since December. The commission, Amnesty International and two other human rights groups had called on the EU to suspend funding for the project. Earlier this month, three U.N. experts expressed concerns about the reports of evictions and urged the EU to stop funding the water project.

NAIROBI Kenya European Union suspend fund water conservation project Kenya tribesman shoot kill government forest service guard another seriously wound EU announce Wednesday shoot take place Tuesday Embobut Forest come bloc European country warn would suspend 35 million project force use local European Unions ambassador Kenya Stefano Dejak say conservation project design protect land Mount Elgon Cherangani Hills area Kenya Known water tower terrain store rainwater enable regular river flow among benefit EU insist full respect right indigenous people conservation work water tower never expect involve eviction use violence bloc say statement Staff member follow report start come year ago concern abuse conservation area claim allegedly tie eviction Sengwer tribespeople EUs financial support EU say governmentfunded Kenya National Commission Human Rights say forest guard carry forceful eviction forest tribe since December commission Amnesty International two human right group call EU suspend fund project Earlier month three UN expert express concern report eviction urge EU stop fund water project

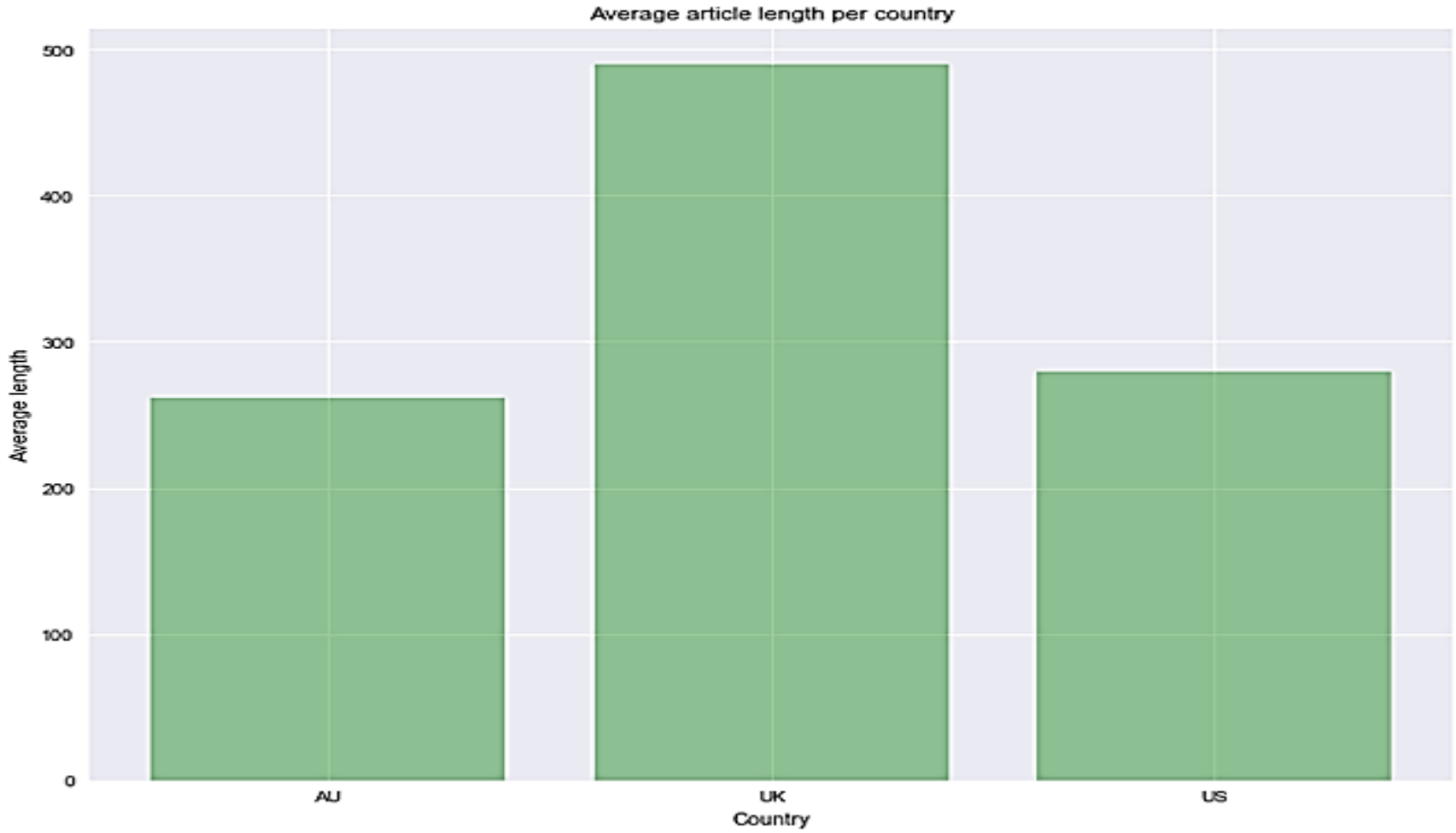
4.1 Analyse der Metadaten

- Anzahl der Artikel pro Land



4.1 Analyse der Metadaten

- Durchschnittliche Länge aller Artikel pro Land



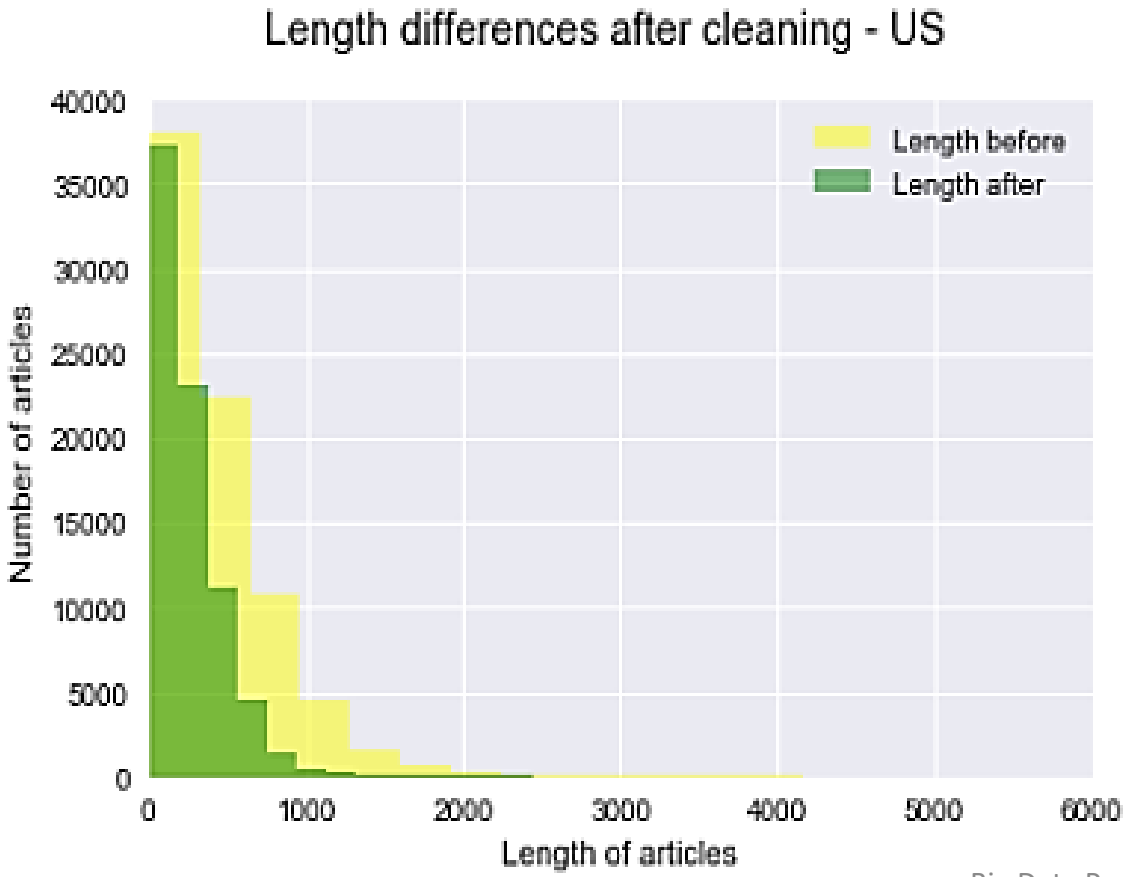
AU Artikel
Gesamtlänge:
12.327.069

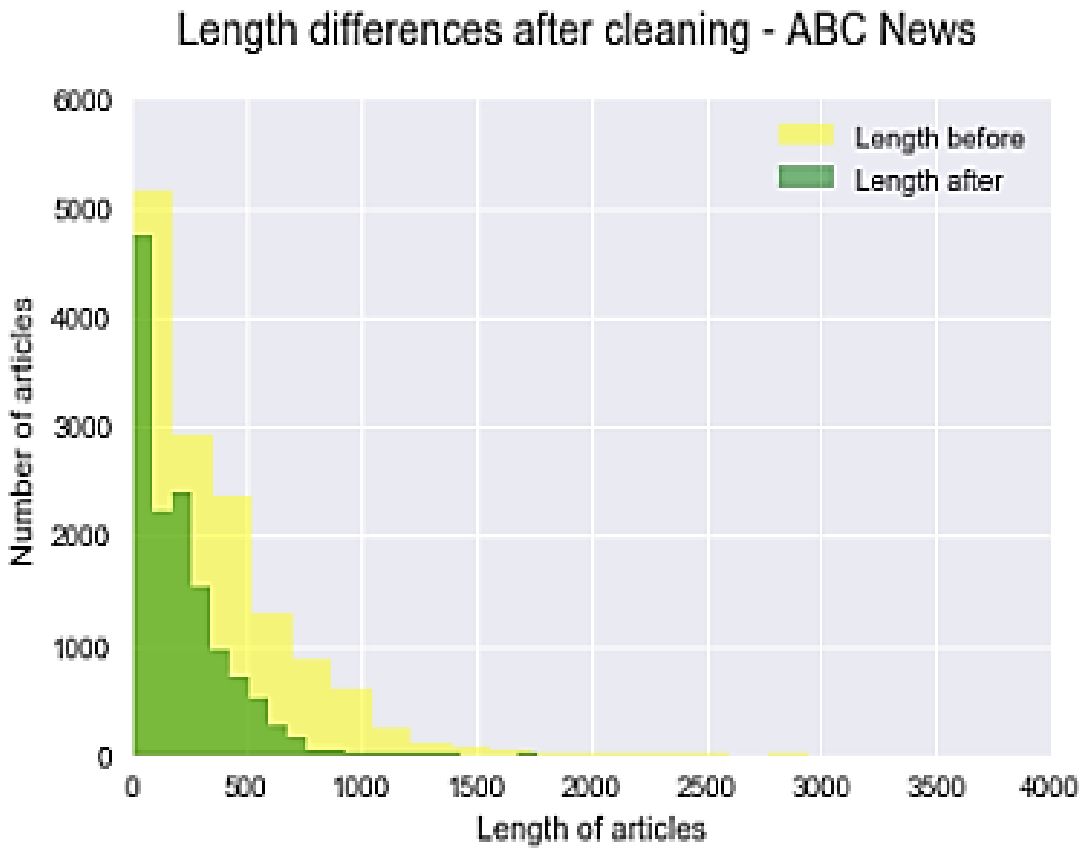
US Artikel
Gesamtlänge:
22.143.257

UK Artikel
Gesamtlänge:
26.297.698

4.1 Analyse der Metadaten

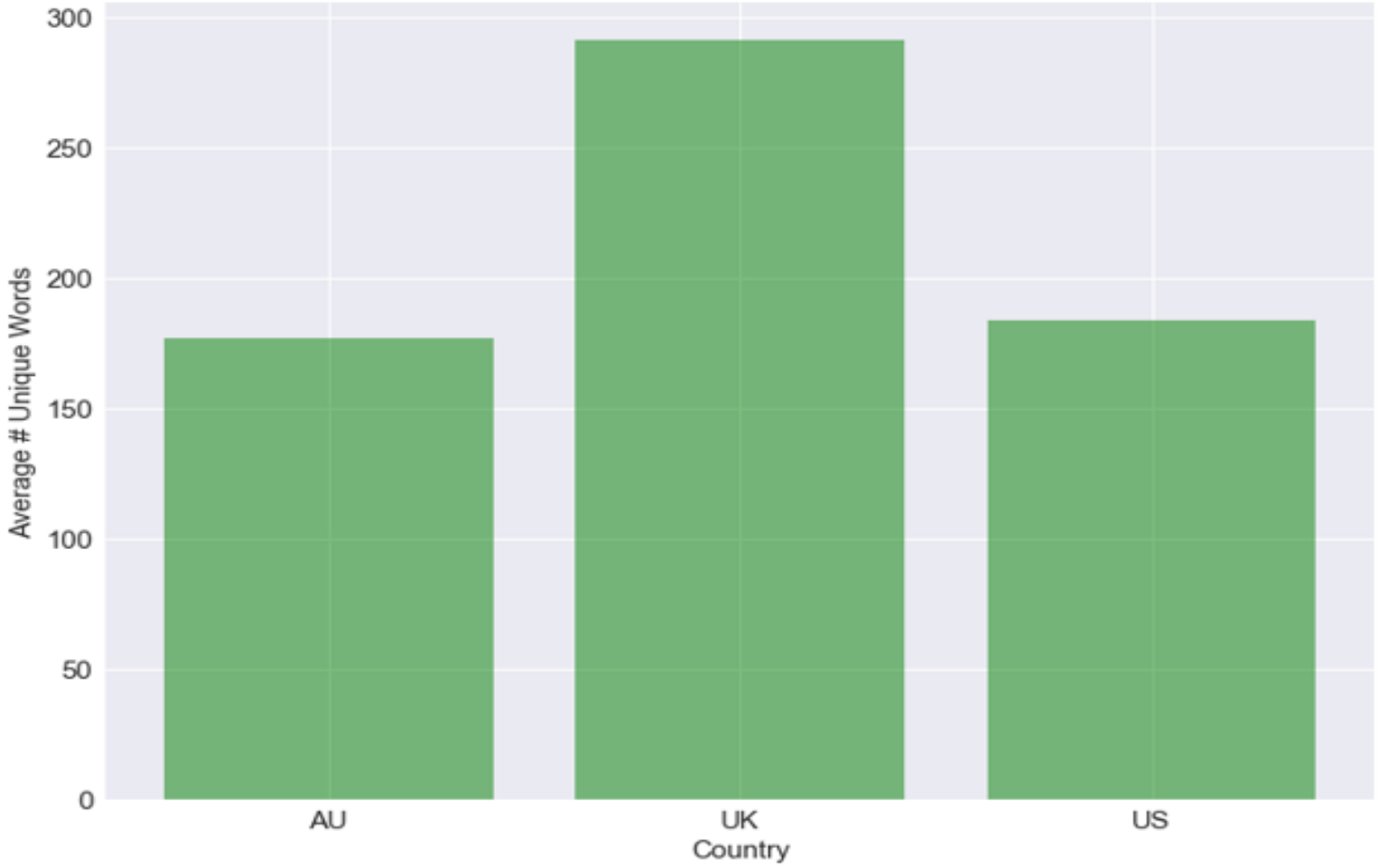
- Textlänge (pro Land und pro Paper) vor und nach dem Bereinigen





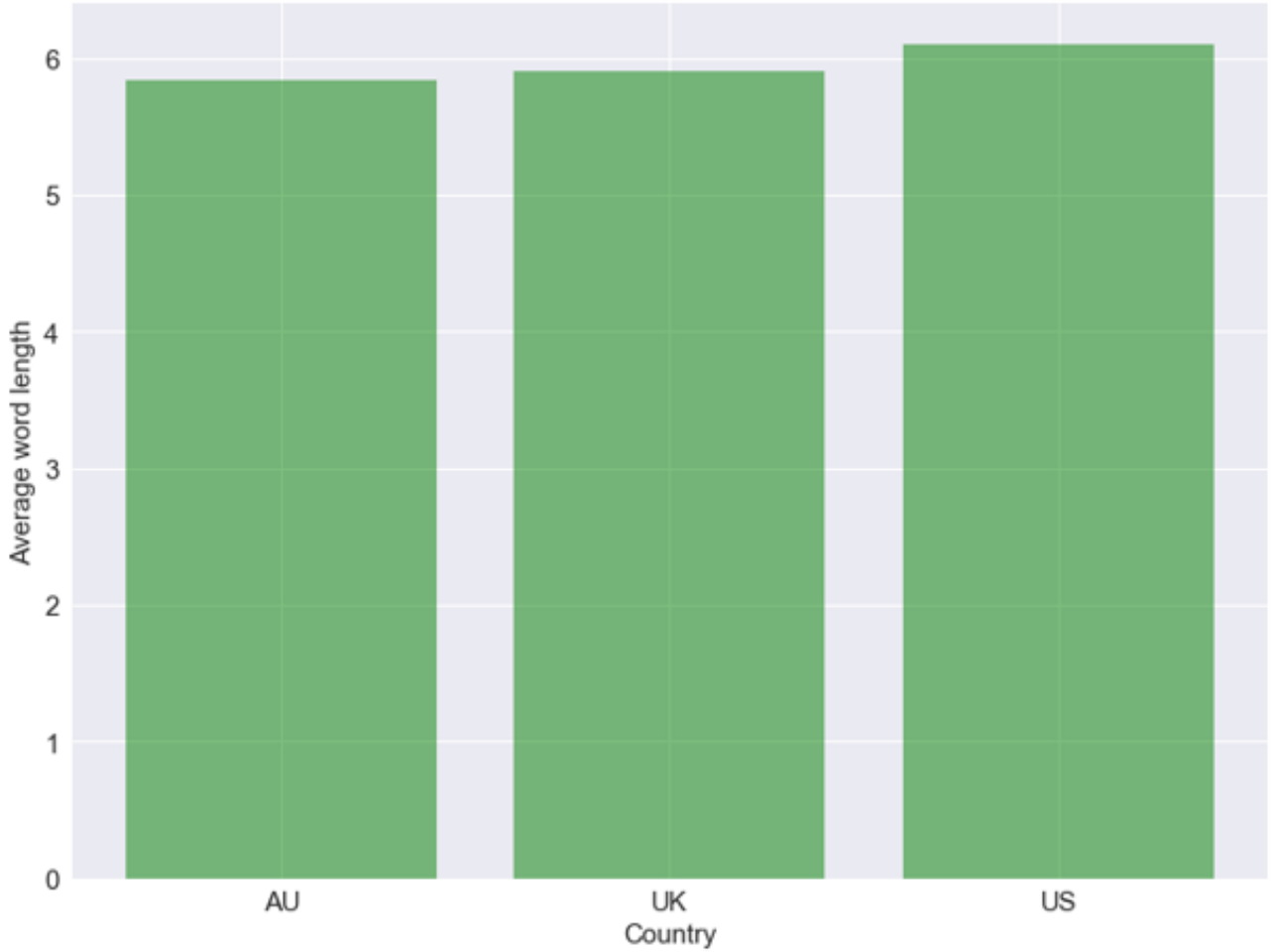
4.1 Analyse der Metadaten

- Durchschnittliche Anzahl an Unique Words pro Land



4.1 Analyse der Metadaten

- Wortlänge pro Land im Durchschnitt



4.2 Untersuchung der Ähnlichkeit

- Part of Speech – Gegenüberstellung der Länder:

Durchschnittliche Anzahl an Wortarten pro Artikel in jedem Land

AU

	NN	NNP	JJ	VBP	CD	RB	VB	IN	VBD	NNS
Number	94.01452	64.778132	34.849672	16.30223	13.476901	11.484406	7.865638	4.219572	3.529243	3.346599

UK

	NN	NNP	JJ	VBP	RB	CD	VB	IN	VBD	NNS
Number	188.423804	117.53676	62.125772	30.083462	21.155079	20.926294	15.883145	8.345154	5.684395	5.620987

US

	NN	NNP	JJ	VBP	RB	CD	VB	IN	NNS	VBD
Number	106.094905	66.671272	37.693216	18.21333	11.283452	9.990233	8.643385	4.363277	3.52289	3.451229

NN - Noun, singular or mass
NNP - Proper noun, singular
JJ - Adjective
VB - Verb, base form
VBP - Verb, non-3rd person singular present
RB - Adverb
CD - Cardinal number

4.2 Untersuchung der Ähnlichkeit

- Part of Speech – Gegenüberstellung der Länder:

Durchschnittliche Anzahl an Wortarten pro Artikel in jeder Zeitung aus AU

The Mercury

	NN	NNP	JJ	VBP	CD	RB	VB	IN	VBD	NNS
Number	87.815563	61.166127	32.81081	15.147958	12.205211	11.225704	7.568768	4.014824	3.422887	3.000599

The Daily Telegraph AU

	NN	NNP	JJ	VBP	RB	CD	VB	IN	VBD	NNS
Number	122.877551	72.431373	45.813125	21.237695	14.533413	11.228491	10.139656	5.256102	4.244898	4.185274

News Australia

	NN	NNP	JJ	VBP	RB	CD	VB	IN	VBD	NNS
Number	142.067061	107.470217	51.997239	24.189744	15.631164	13.800789	10.881657	6.47574	5.216963	4.712032

The Advertiser

	NN	NNP	JJ	CD	VBP	RB	VB	IN	VBD	NNS
Number	122.238334	98.668564	43.862053	29.437742	20.867744	14.732529	10.253358	5.159117	4.671523	4.404052

ABC Australia

	NN	NNP	JJ	VBP	CD	RB	VB	NNS	IN	VBD
Number	109.046023	65.182204	40.701204	20.008025	14.333963	12.14515	8.525608	5.029974	4.921643	3.662497

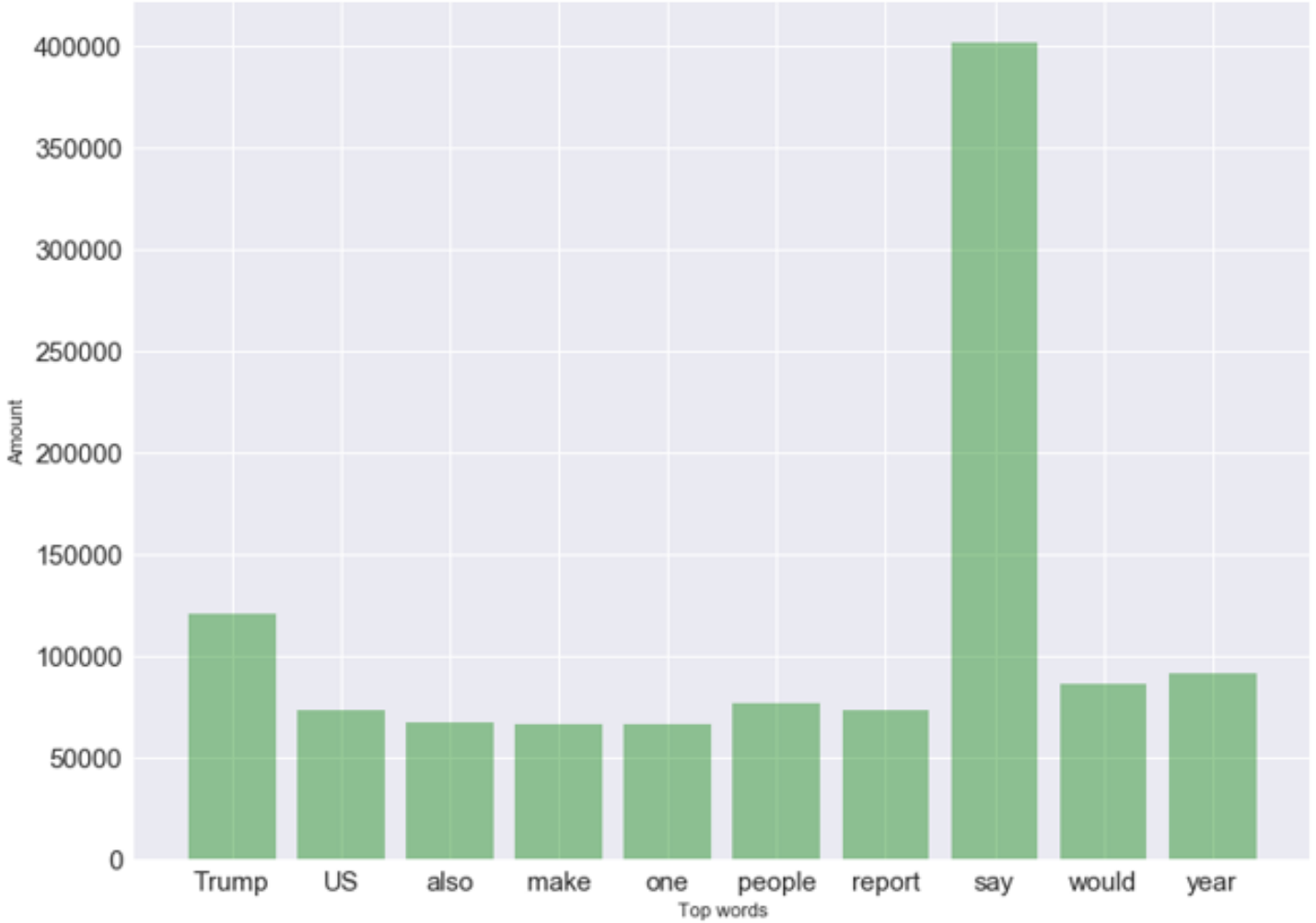
The west australian

	NN	NNP	JJ	VBP	CD	RB	VB	IN	NNS	VBD
Number	51.133021	28.443879	18.693217	8.932777	6.777348	5.729069	4.110409	2.23243	1.781422	1.74231

NN - Noun, singular or mass
NNP - Proper noun, singular
JJ - Adjective
VB - Verb, base form
VBP - Verb, non-3rd person singular present
RB - Adverb
CD - Cardinal number

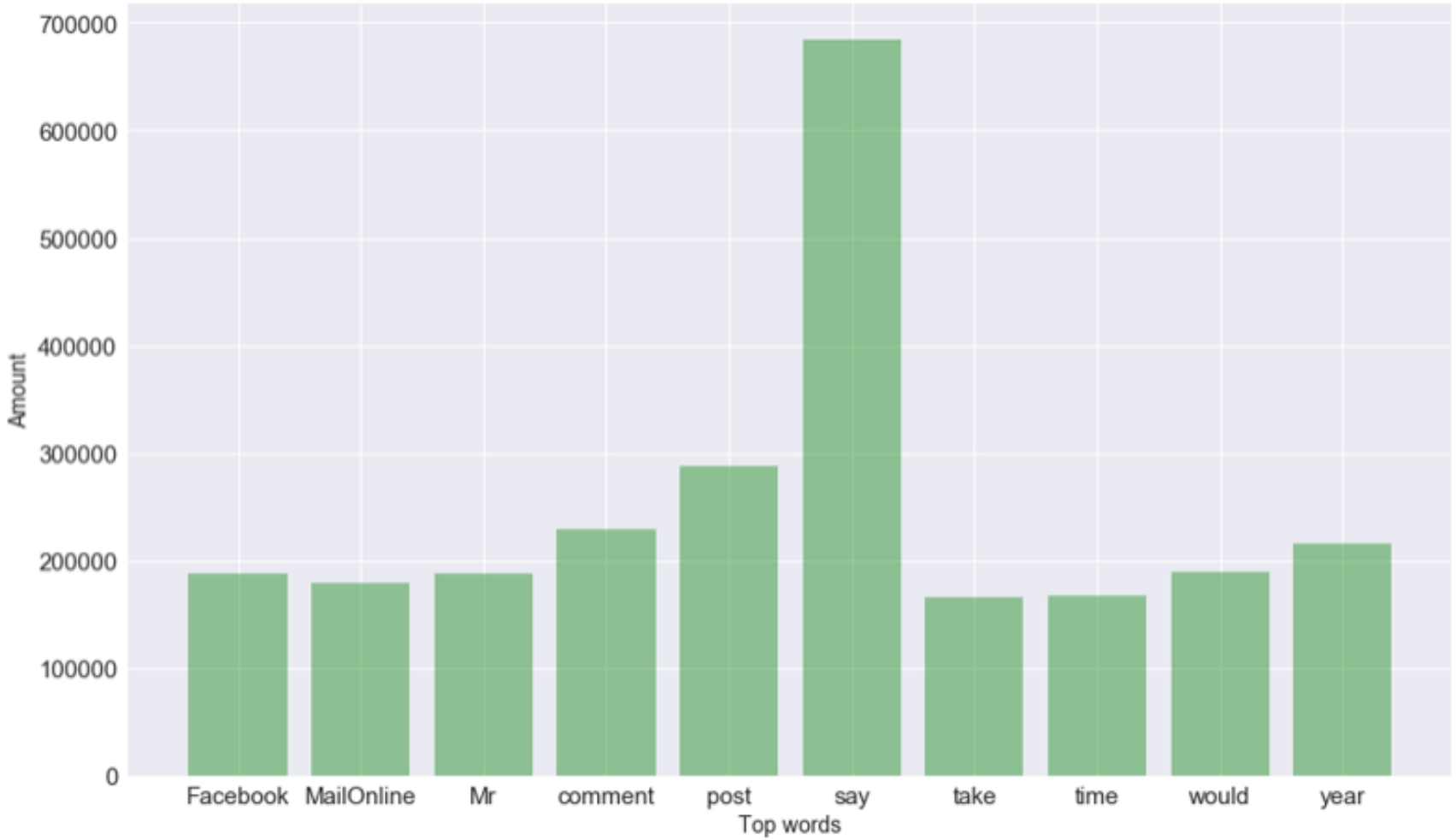
4.2 Untersuchung der Ähnlichkeit

- Bag of Words – Top 10 Wörter US



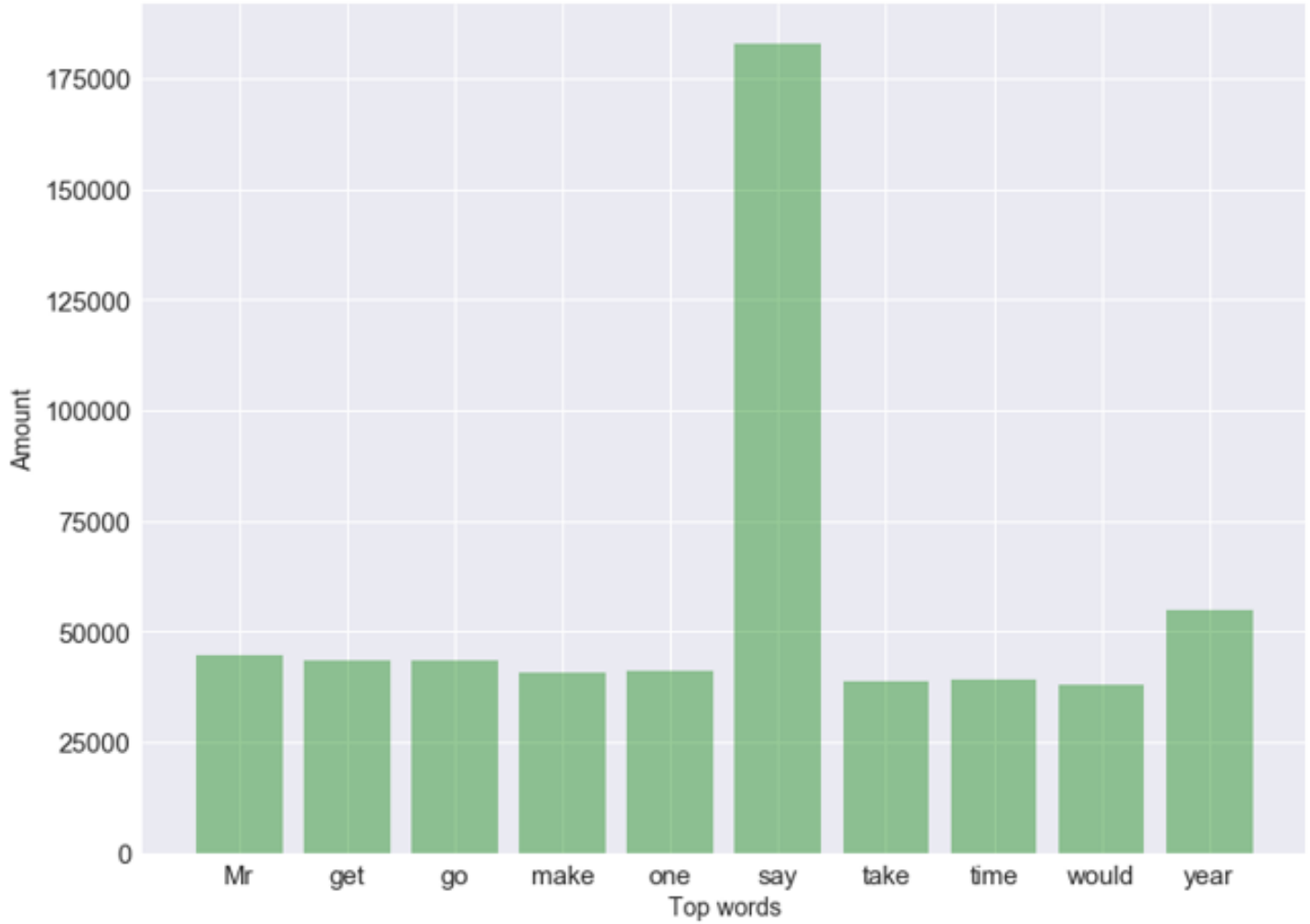
4.2 Untersuchung der Ähnlichkeit

- Bag of Words Top 10 Wörter UK



4.2 Untersuchung der Ähnlichkeit

- Bag of Words Top 10 Wörter AU



4.2 Untersuchung der Ähnlichkeit

- Vektorisieren der Artikel
- darauf basierend Messung der Distanzen
  (euklidische Distanz, Cosinus-Distanz)

→Memory Error (Chunking?)

Euklidische Distanzen

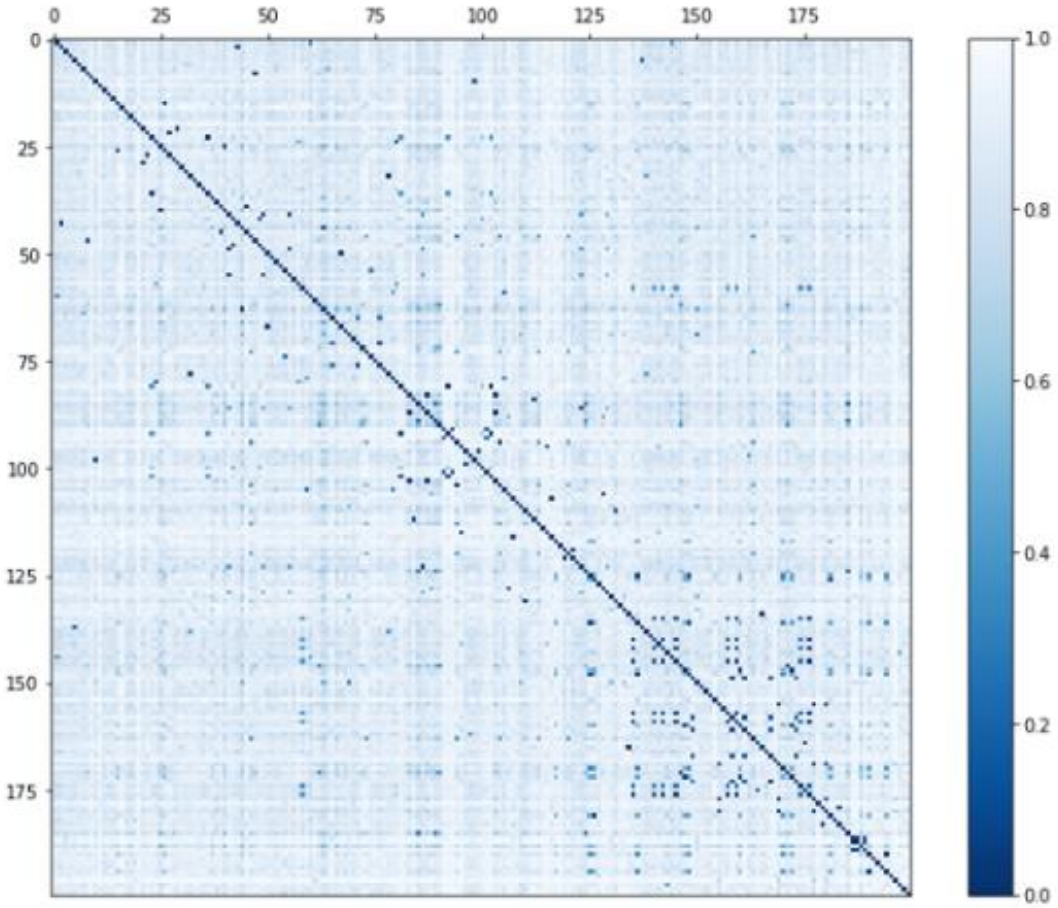
```
array([[  0.   ,  30.984,  34.337, ...,  18.52 ,  20.174,  20.224],
       [ 30.984,   0.   ,  35.93 , ...,  27.258,  27.946,  27.221],
       [ 34.337,  35.93 ,   0.   , ...,  31.016,  31.591,  31.591],
       ...,
       [ 18.52 ,  27.258,  31.016, ...,   0.   ,  10.77 ,  11.402],
       [ 20.174,  27.946,  31.591, ...,  10.77 ,   0.   ,  14.142],
       [ 20.224,  27.221,  31.591, ...,  11.402,  14.142,   0.   ]])
```

Cosinus - Distanzen

```
array([[ 0.   ,  0.873,  0.89 , ...,  0.925,  0.914,  0.877],
       [ 0.873,  0.   ,  0.748, ...,  0.883,  0.867,  0.78 ],
       [ 0.89 ,  0.748, -0.   , ...,  0.904,  0.883,  0.864],
       ...,
       [ 0.925,  0.883,  0.904, ...,  0.   ,  0.932,  0.876],
       [ 0.914,  0.867,  0.883, ...,  0.932,  0.   ,  0.885],
       [ 0.877,  0.78 ,  0.864, ...,  0.876,  0.885,  0.   ]])
```

4.2 Untersuchung der Ähnlichkeit

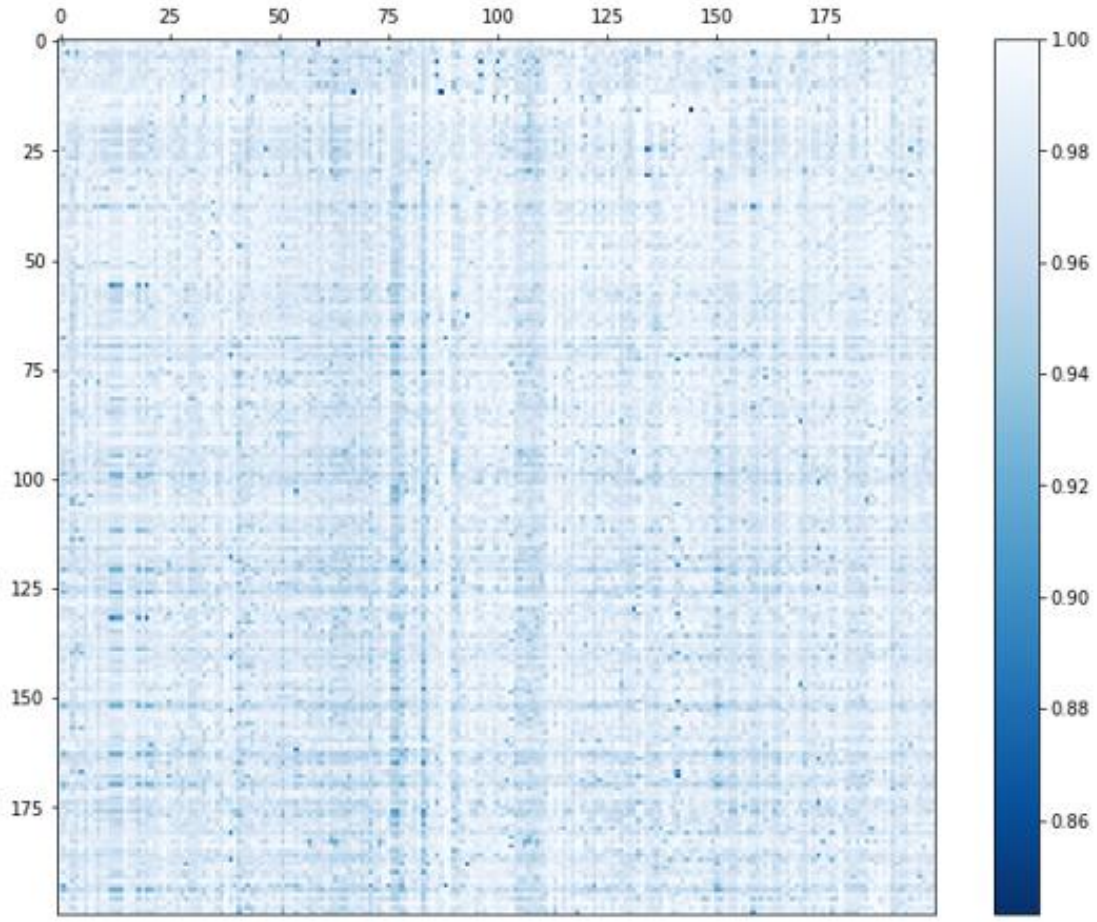
- Ähnlichkeit von Artikel abbilden, basierend auf Distanz
- Cosinus – Distanzen zwischen US Artikeln:



→ Memory Error (Chunking?)

4.2 Untersuchung der Ähnlichkeit

- Cosinus – Distanzen zwischen AU und UK Artikeln:

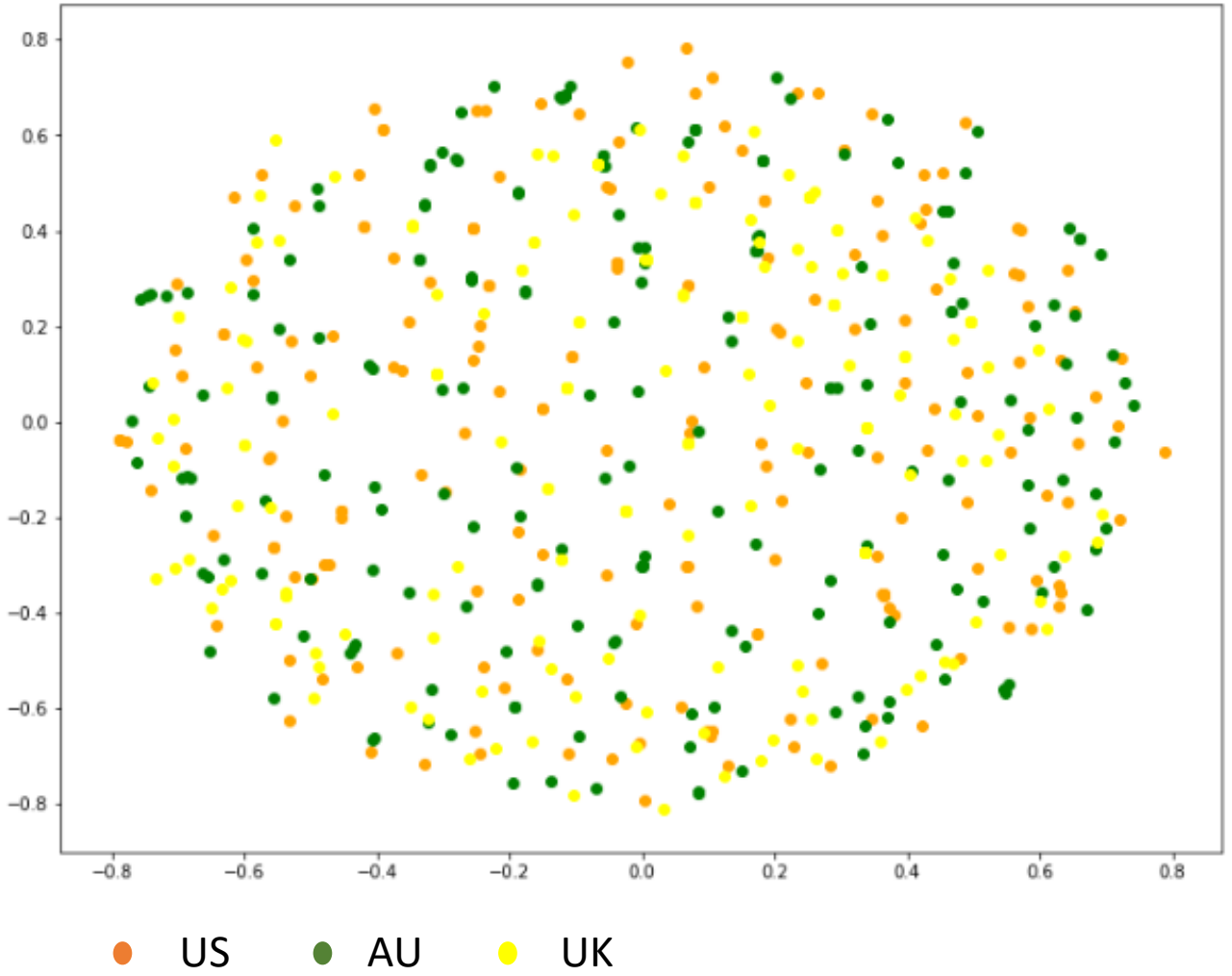


→ Memory Error (Chunking?)

4.2 Untersuchung der Ähnlichkeit

- Ähnlichkeit von Artikel abbilden, basierend auf Distanz

- Multidimensional Scaling (MDS): Entfernung zwischen Punkten ist proportional zur paarweisen Entfernungen

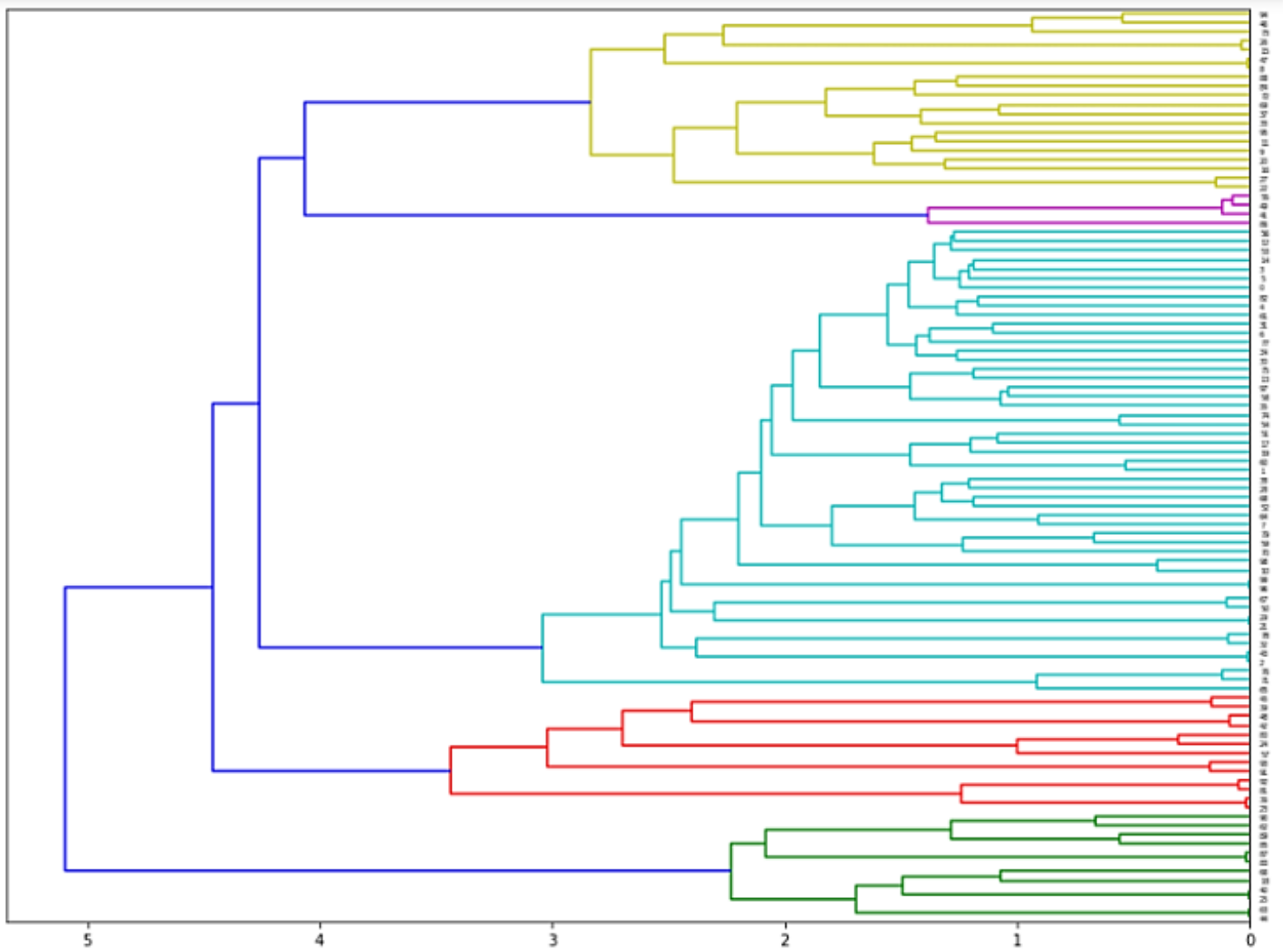


→ Memory Error (Chunking?)

→ Abhängigkeit zwischen Punkten aus verschiedenen Ländern

4.2 Untersuchung der Ähnlichkeit

- Ward – Methode: Hierarchisches Clustering der US-Artikel in Gruppen anhand von Cosinus-Distanzen



→ Memory Error (Chunking?)

5. Offene Punkte

- Problem: Memory Error (z.B. schon beim Vectorizing, können deshalb auch weiterführend nicht die Distanzen zwischen Ländern berechnen)
- TF-IDF implementieren (zusätzlich zu BOW)
- MDS – Abhängigkeit zwischen den Ländern hinzufügen
- Performance im Auge behalten

Vielen Dank für eure Aufmerksamkeit!

Anhang – Arbeitsaufteilung

Aufgrund dessen, dass wir unser Verständnis für die Inhalte, sowie das Programmieren des Codes nach und nach aufgebaut haben, lässt sich nicht feststellen, wer was gemacht hat, da der Code in Zusammenarbeit entstanden ist. So hat bspw. jeder an dem Crawler Code gearbeitet, bis dieser funktioniert hat. Dabei hat jeder etwas beigetragen.

Die Arbeit nun fair aufzuteilen scheint uns unmöglich. Wenn wir jemandem ein Kapitel zuordnen, das vermeintlich weniger Wert ist als ein anderes, ergibt sich unberechtigter Weise eine schlechtere individuelle Note, was unserer Meinung nach sehr unfair wäre, da jeder den gleichen Aufwand geleistet hat.

Sara: Interpolation – Matrix, Ward – Methode, Cosinus – Distanzen, POS, TF-IDF, Metadatenanalyse, Visualisierung

Alex: MDS, euklidische Distanzen, Lemmatizing, BOW, Chunking, Metadatenanalyse, Visualisierung

Tatyana: BeautifulSoup – Teil, Cluster , Vectorizing, Visualisierung, Perfomance – Probleme, Metadatenanalyse, Dataframe – Verwaltung (Pickle und co.)

Alle: Einstieg in Python