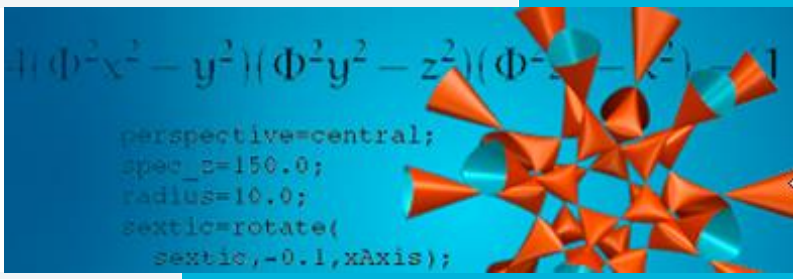
Philipps Universität Marburg

Simple language and NLP

A concept worth considering



Question:
Is simple language relevant when using NLP procedures?

- Can we optimize our features for simple language?
- Do our known features perform as well on simple language as on normal language?
- Should we train with simple language when we intend to classify simple language?

Data

- Spiegel.de
- Nachrichtenleicht.de
- Prelabeled

	simple language				normal language			
	culture	sport	politics	**total**	culture	sport	politics	**total**
# of articles	1304	1230	2020	**4554**	2357	2659	2233	**7249**
avg length of article (words)	151	137	174	**157**	93	93	89.8	**91.7**
avg reading ease	64	63	61.5	**62.77**	42	46.3	33	**40.6**
# of tokens	716642				714773			
# of unique words	47602				76126			

Table 1: Overview over the datasets

Approach and Methods

- Feature
  - TF-IDF as control
  - Modified TF-IDF as trial
  - Train-test split of 70-30
  - Classification using simple or normal data both for training and classification
- Classification
  - Different configurations between countvectors, TF-IDF and three classifiers(NB, MLP, DT)
  - Different setups which data is used for train and for test

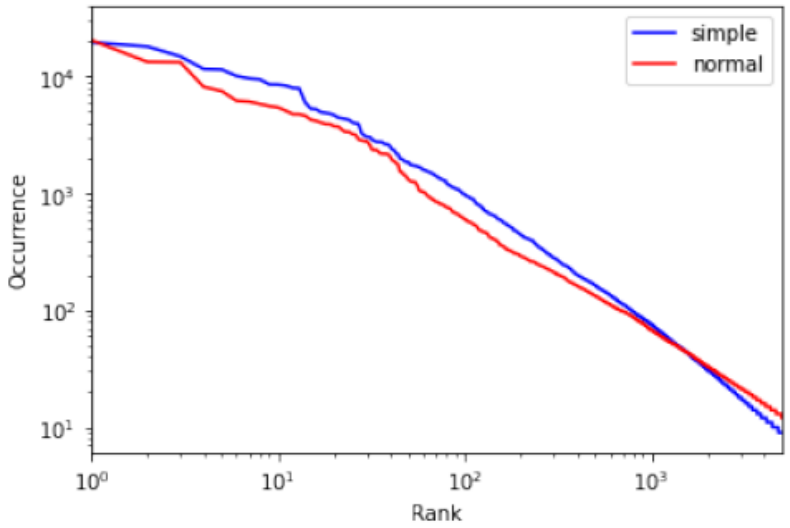


Figure 1: Rank Frequency curve.

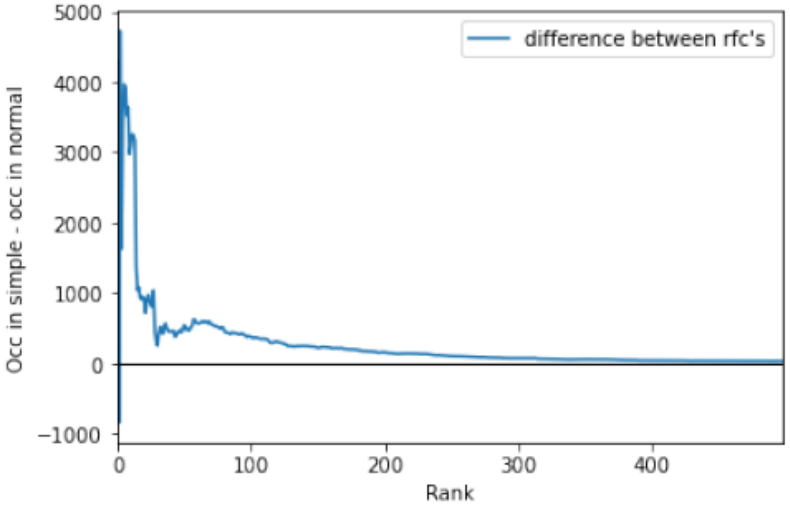


Figure 2: Difference between the two lines

- „[…]. Gar kein Plastik zu benutzen ist sehr schwierig. Darum sagt die EU-Kommission: Wenn eine Firma aus altem Plastik neue Dinge macht, geben wir der Firma Geld dafür. […]“

- „[…]. Die Kanzlerin zeigt damit Größe – offenbart aber auch ihr eigenes Scheitern. Sie erinnern sich, Jens Spahn, vor gut einem Jahr? »Wir werden in ein paar Monaten einander wahrscheinlich viel verzeihen müssen.«[…]“

Results

- Minor improvements for simple language when using the modified feature
- No visible change for normal language

- Classifier trained on normal language

-> Performance worsens when testing against simple language

Results - Classification

	precision	recall	f1 score	#articles
culture	0.90	0.93	0.91	708
politics	0.94	0.92	0.93	655
sport	0.98	0.96	0.97	812
accuracy		0.94		2175
macro avg	0.94	0.94	0.94	2175
weighted avg	0.94	0.94	0.94	2175

Table 2: MLP classifier, TD-IDF vectors, trained on conventional, tested on conventional

	precision	recall	f1 score	#articles
culture	0.74	0.86	0.80	1304
politics	0.89	0.83	0.86	2020
sport	0.98	0.93	0.96	1230
accuracy		0.87		4554
macro avg	0.87	0.87	0.87	4554
weighted avg	0.87	0.87	0.87	4554

Table 3: MLP classifier, TD-IDF vectors, trained on conventional, tested on simple lanuage

Discussion and outlook

- Simple language is different enough from normal language to have an impact
- Possibility to gear procedures towards it
- Necessary to consider when designing an application
- Future:
  - More complex tools
  - Machine translated simple language?

Conclusion

- Two guiding questions about the relation of simple language with NLP
- Experiments show value in considering simple language when applying NLP procedures
- Simple language is a different style of language entirely