1_Measures

Herleitung

Um die Güte des Clusterns zu bewerten benötigt man ein Measure. Inituitiv wird die Bewertung über eine 3D Visualisiserung vorgenommen, was jedoch sehr zeitaufwendig ist.

Recherche

Menge an möglichen Measures:

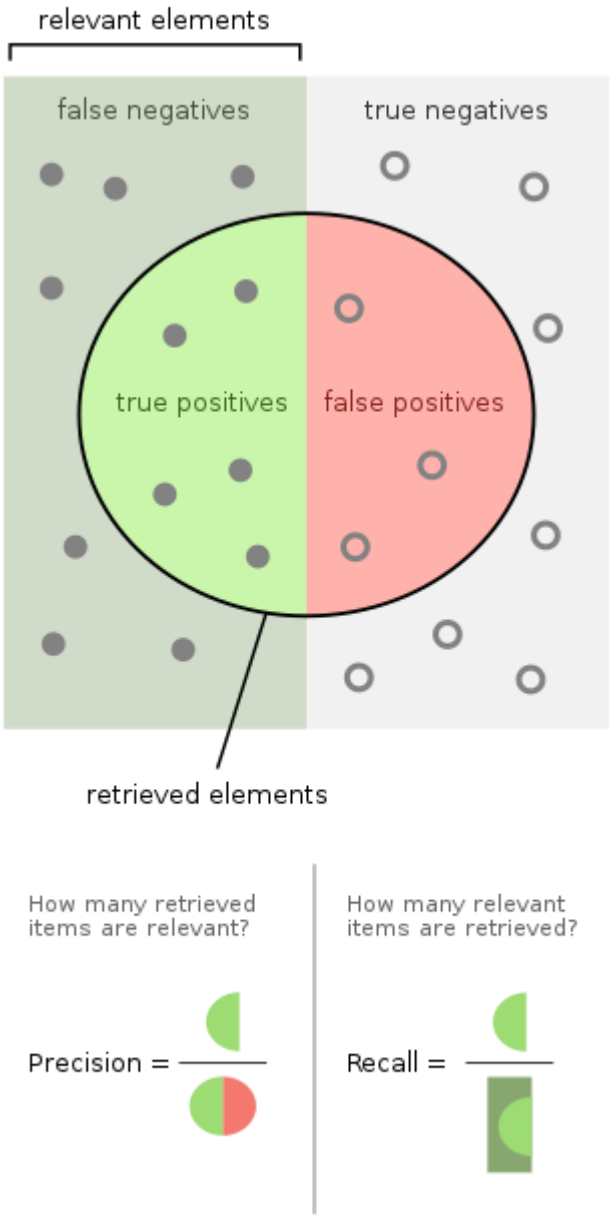
- True Positives
- True Negatives
- False Positives
- False Negatives
- Accuracy
- Precision
- Recall
- F1-Score
- Adjusted Rand Index (ARI)
- Adjusted Mutual Information (AMI)
- Fowlkes-Mallows Index (FMI)
- ggf. weitere (siehe Data Science Vault)

Hypothese

Es kann ein Measure identifiziert werden, was am besten geeignet ist für die Gütemessung

Ergebnisse

Folgende Grafik dient zur Argumentation



		Predicted condition		Sources: [4][5][6][7][8][9][10][11][12] view · talk · edit	
	Total population $= P + N$	**Predicted Positive (PP)**	**Predicted Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - TPR$
	Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

True Positives

- Führt dazu, dass eine Adjazenzmatrix gefüllt nur mit 1, die beste Alternative ist.
- Alle Wörter sind im selbem Cluster
- **ungeeignet**

True Negatives

- Führt dazu, dass eine Adjazenzmatrix gefüllt nur mit 0, die beste Alternative ist.
- Alle Wörter sind eigene Cluster
  ungeeignet

False Positives

- Führt dazu, dass eine Adjazenzmatrix gefüllt nur mit 0, die beste Alternative ist.
- Alle Wörter sind eigene Cluster
  ungeeignet

False Negatives

- Führt dazu, dass eine Adjazenzmatrix gefüllt nur mit 1, die beste Alternative ist.
- Alle Wörter sind im selbem Cluster
  ungeeignet

Accuracy

Da die Adjazenzmatrix eine sparse-Matrix ist, kann die Accuracy ebenfalls nicht genutzt werden. Die große Anzahl von True Negatives, die erreicht werden, wenn keine Verbindungen hergestellt werden, überwiegt bei weitem die wenigen Fälle, in denen True Positivs einen positiven Effekt hätten.
ungeeignet

Precision

Schon geeigneter, bezieht aber keine negativen Fälle mit ein. Kann aber genutzt werden.
ungeeignet

Hyperparameter Optimierung

Visuelle Beurteilung

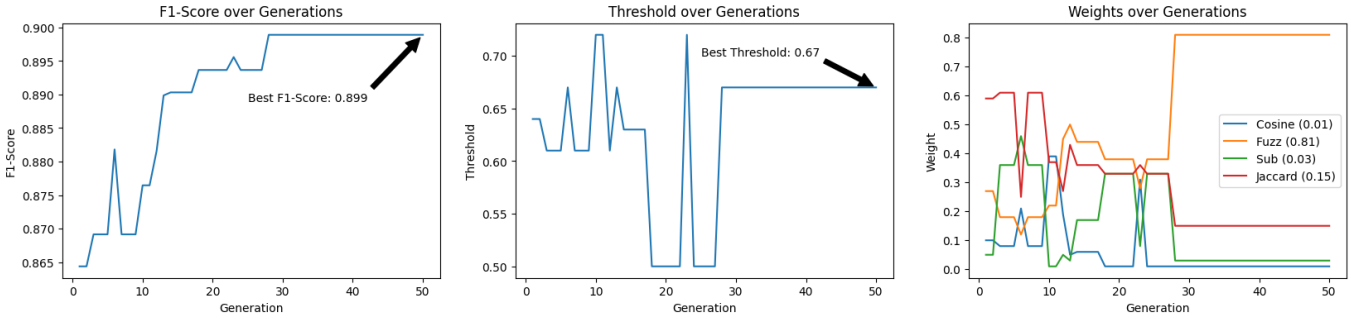
Recall

Nicht geeignet, da auch hier ein guter Recall erreicht wird, wenn ich alle Verbindungen zulasse.
ungeeignet

F1-Score

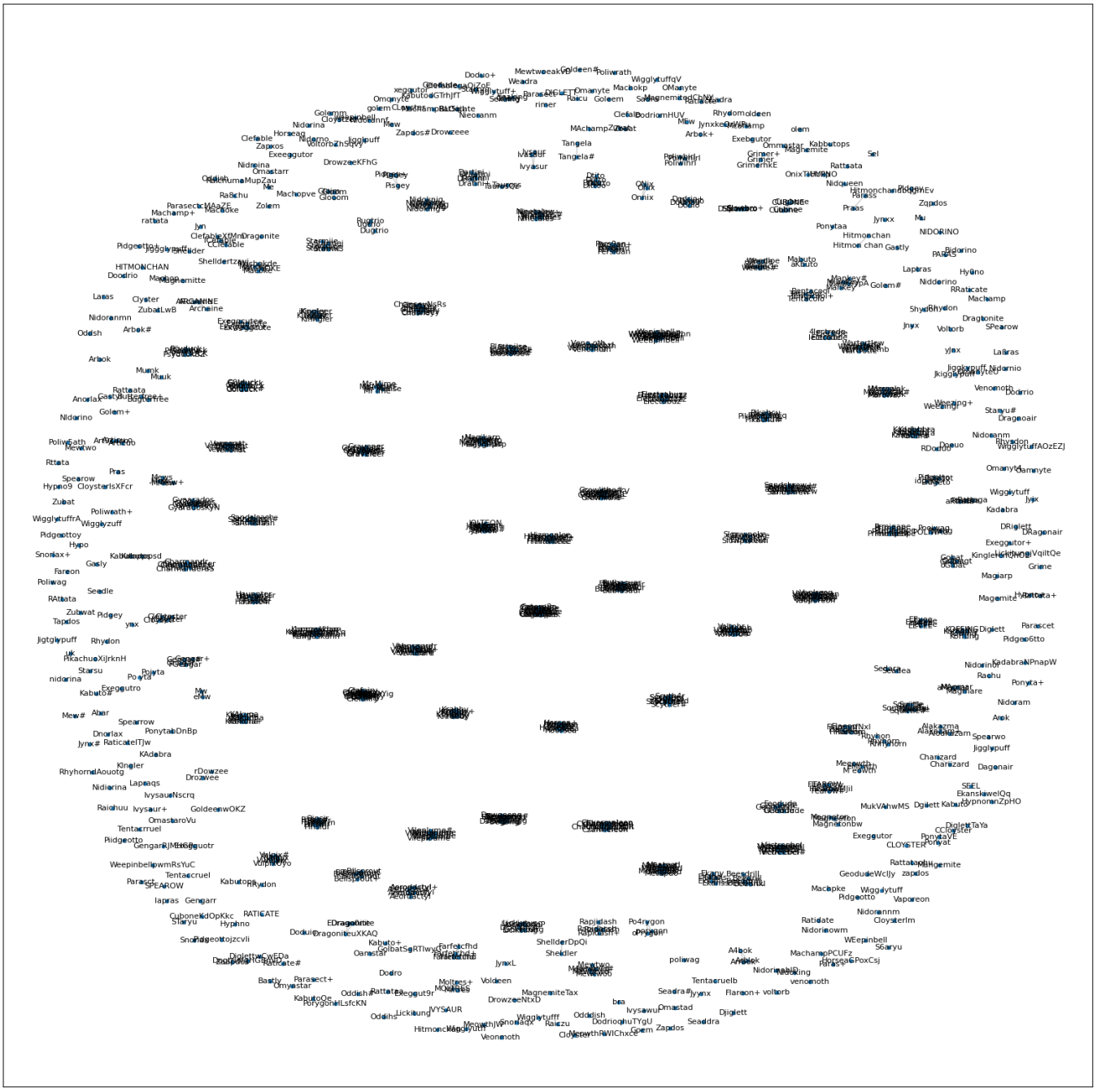
Schon geeigneter, bezieht aber keine negativen Fälle mit ein. Kann aber genutzt werden.
geeignet

Hyperparameter Optimierung



Merkwürdigerweise verläuft die Fitness nicht monoton steigend.

Visuelle Beurteilung



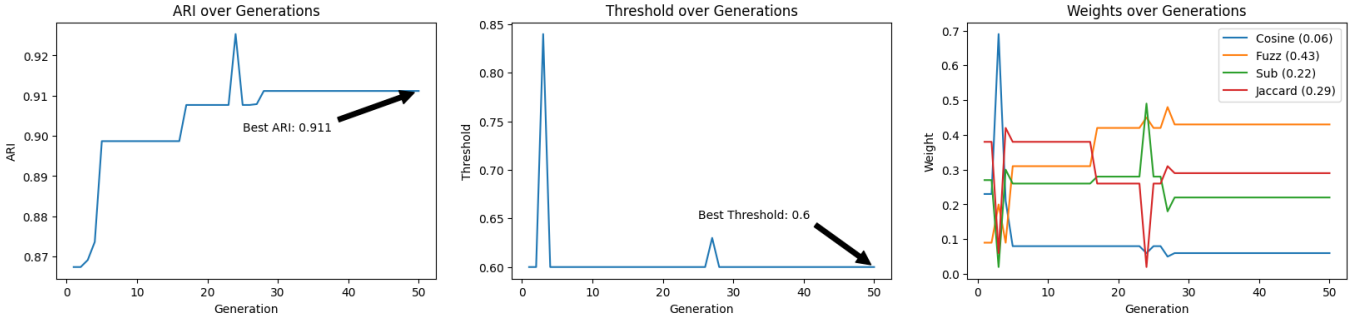
Größere Cluster und weniger einzelne Wörter

F1_Score.html

ARI

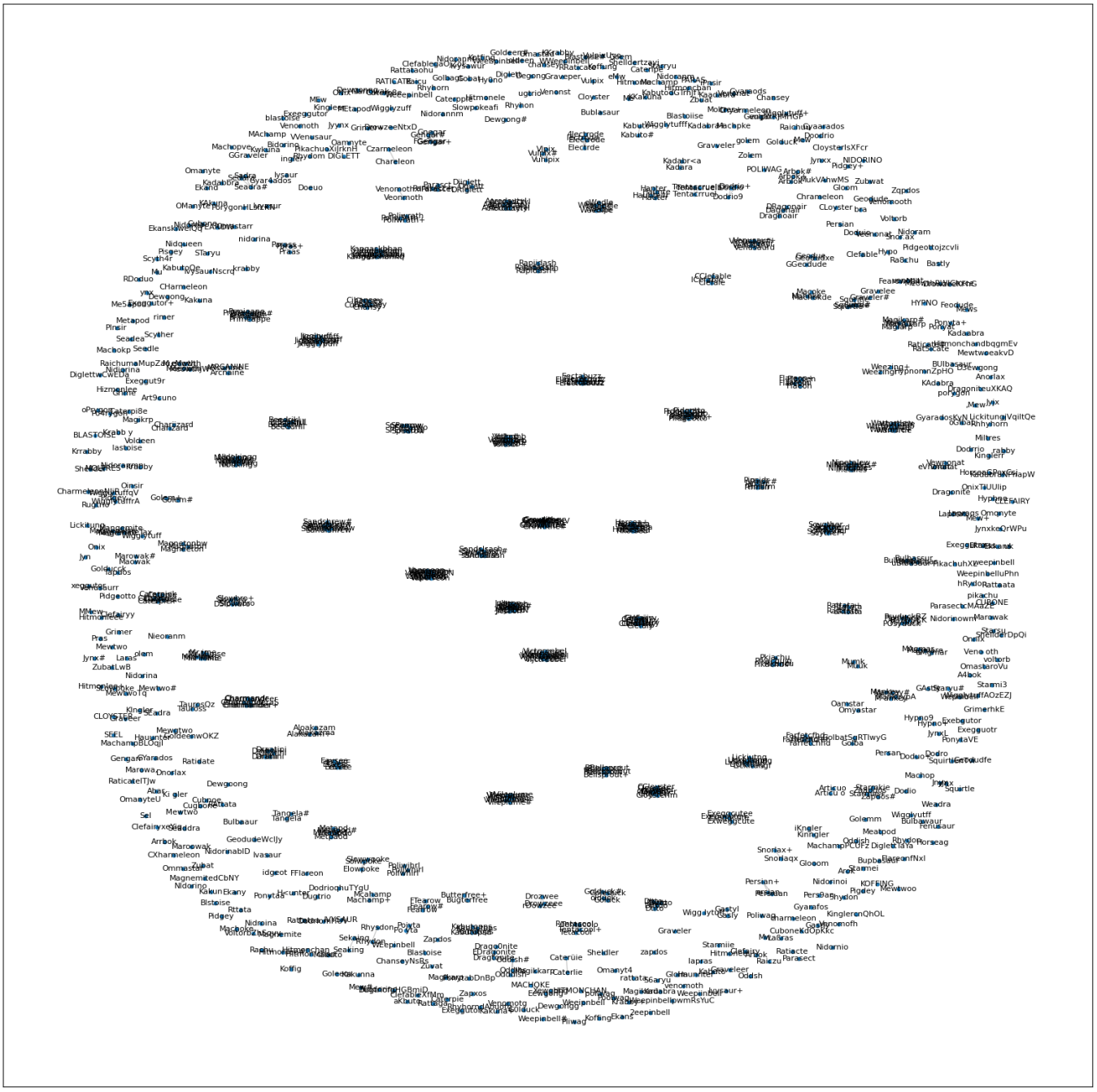
Geeignet für Clusteranalysen, bei denen die Form und Größe der Cluster variieren können und die Bewertung der Ähnlichkeit zwischen zwei Clusterzuordnungen erforderlich ist.
geeignet

Hyperparameter Optimierung



Merkwürdigerweise verläuft die Fitness nicht monoton steigend. Gewichte sind ähnlich zu AMI. Threshold ist wieder halbwegs gleich.

Visuelle Beurteilung



Wieder sehr viele einzelnen Wörter ohne Cluster.

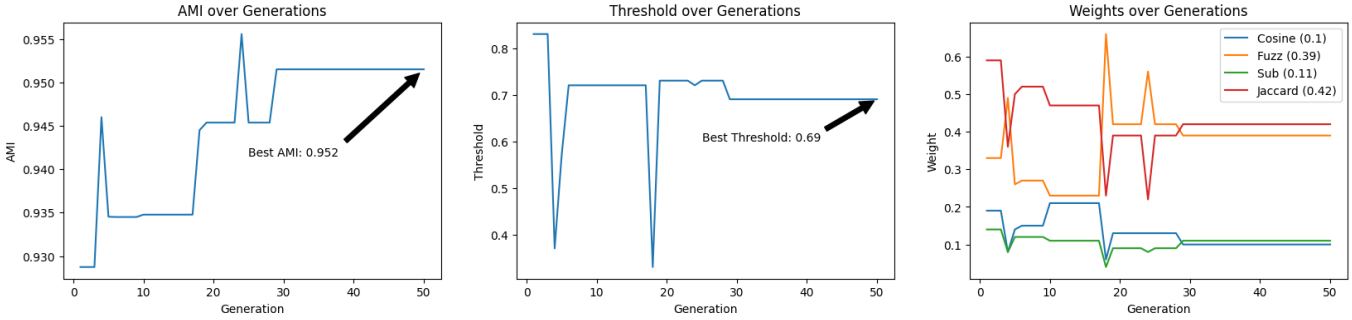
ARI.html

AMI

Geeignet für Clusteranalysen, bei denen die Form und Größe der Cluster variieren können und die Bewertung der Ähnlichkeit zwischen zwei Clusterzuordnungen erforderlich ist.
geeignet

Hyperparameter Optimierung

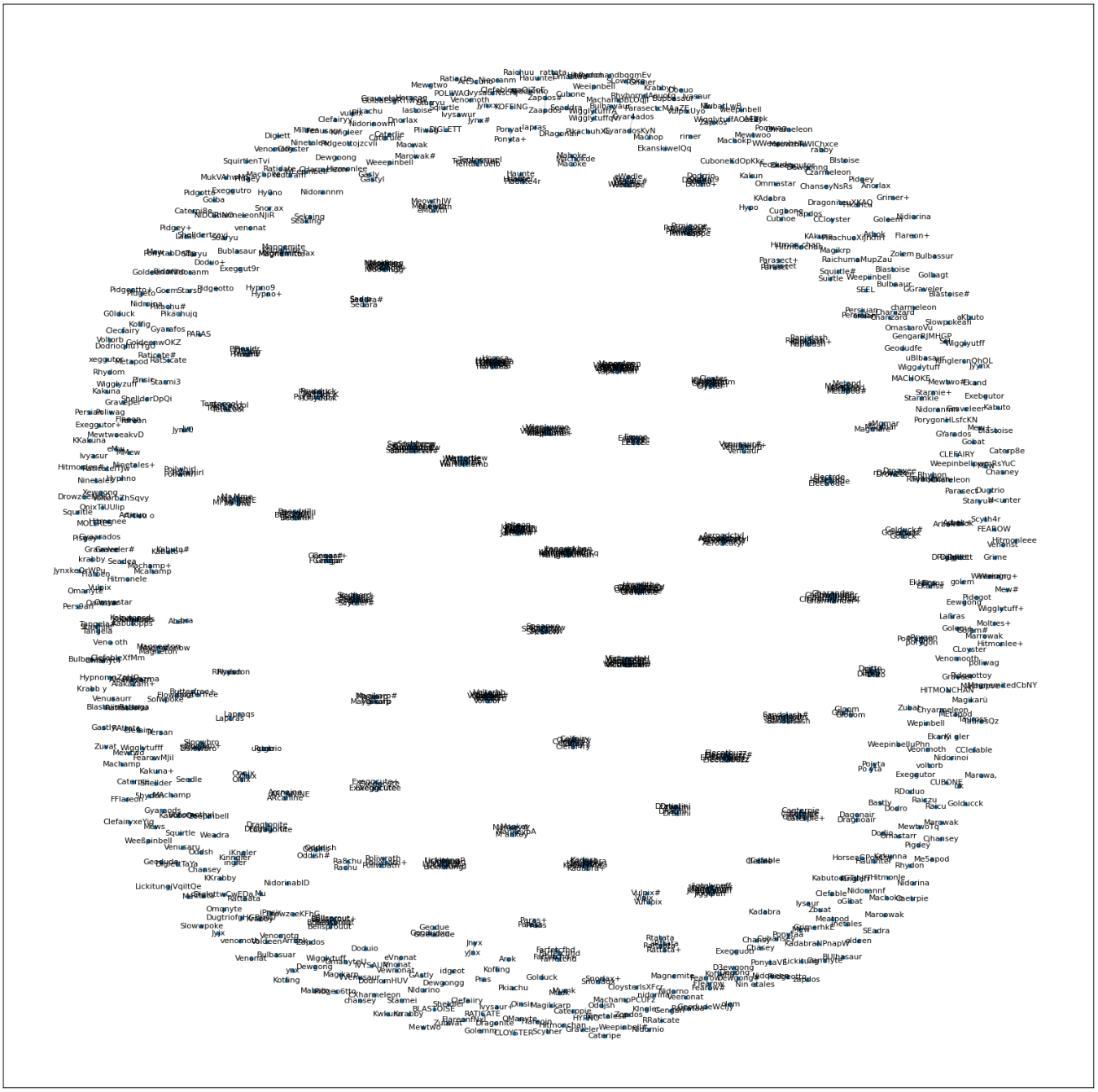
Recht langsame Performance



Merkwürdigerweise verläuft die Fitness nicht monoton steigend. Gewichte weichen zu F1-Score und FMI ab. Threshold ist ähnlich.

Visuelle Beurteilung

Viele einzelne Wörter ohne Cluster



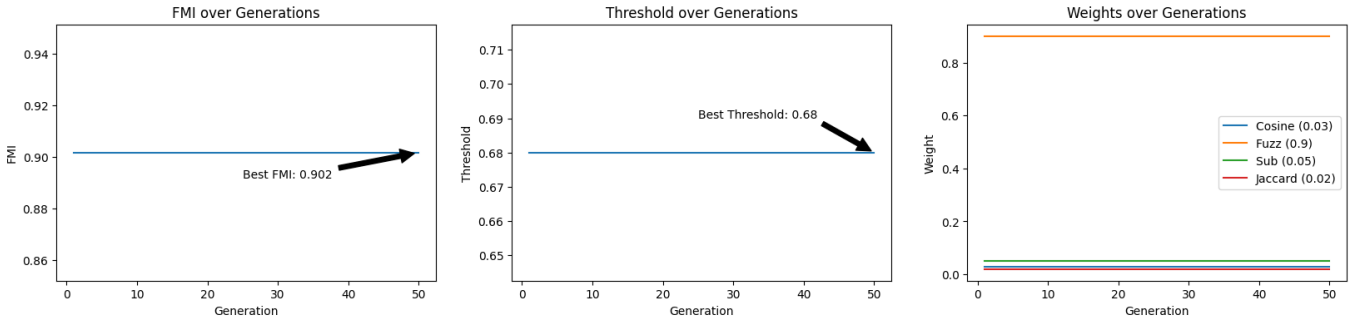
AMI.html

FMI

Gut geeignet, wenn die Größe der Cluster bekannt ist und die Bewertung der Ähnlichkeit zwischen den Clusterzuordnungen wichtig ist.

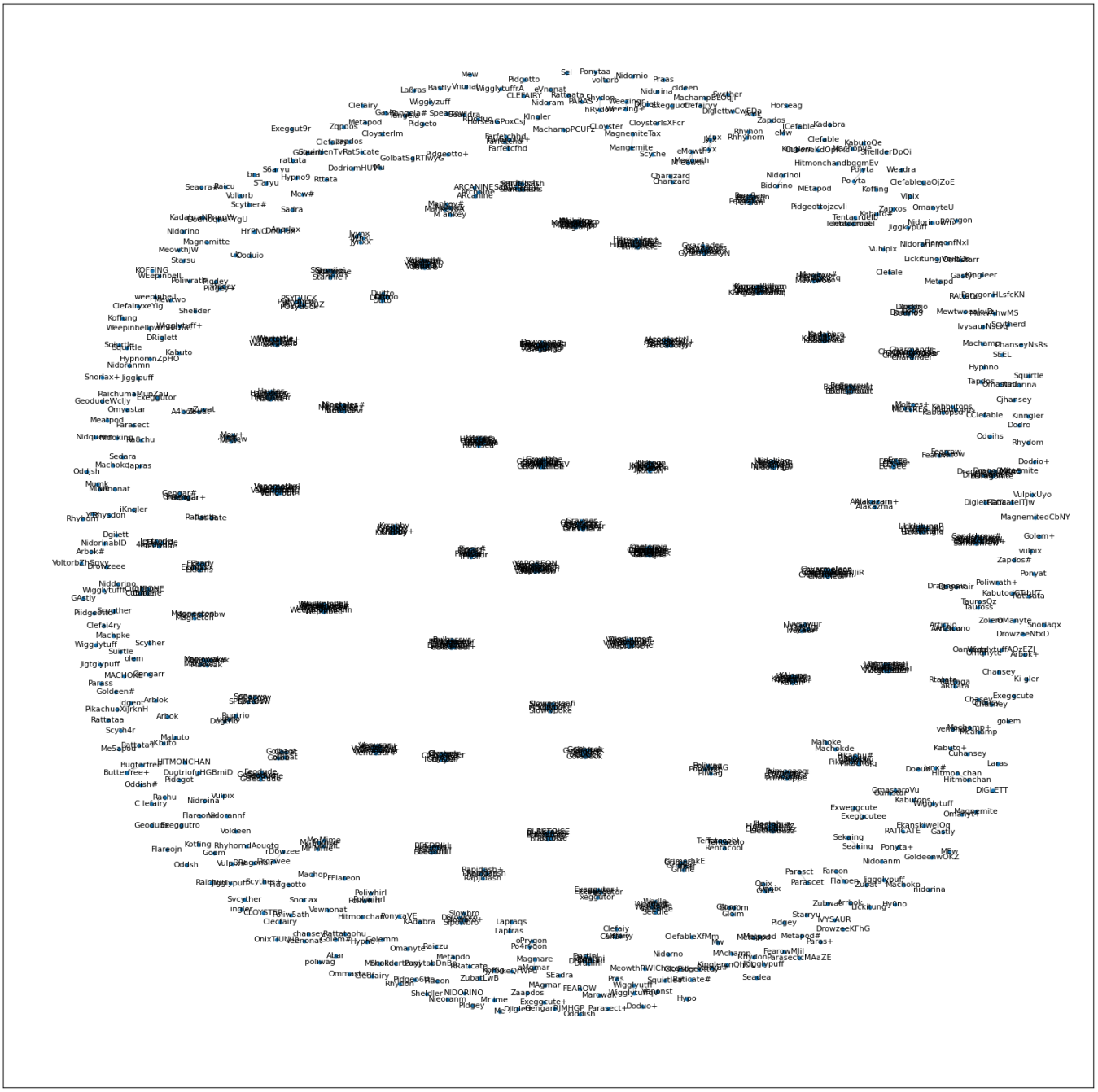
bedingt geeignet

Hyperparamter Optimierung



Funktioniert recht gut. Gewichte sind ähnlich zu F1-Score. Sehr schnelle Performance

Visuelle Beurteilung



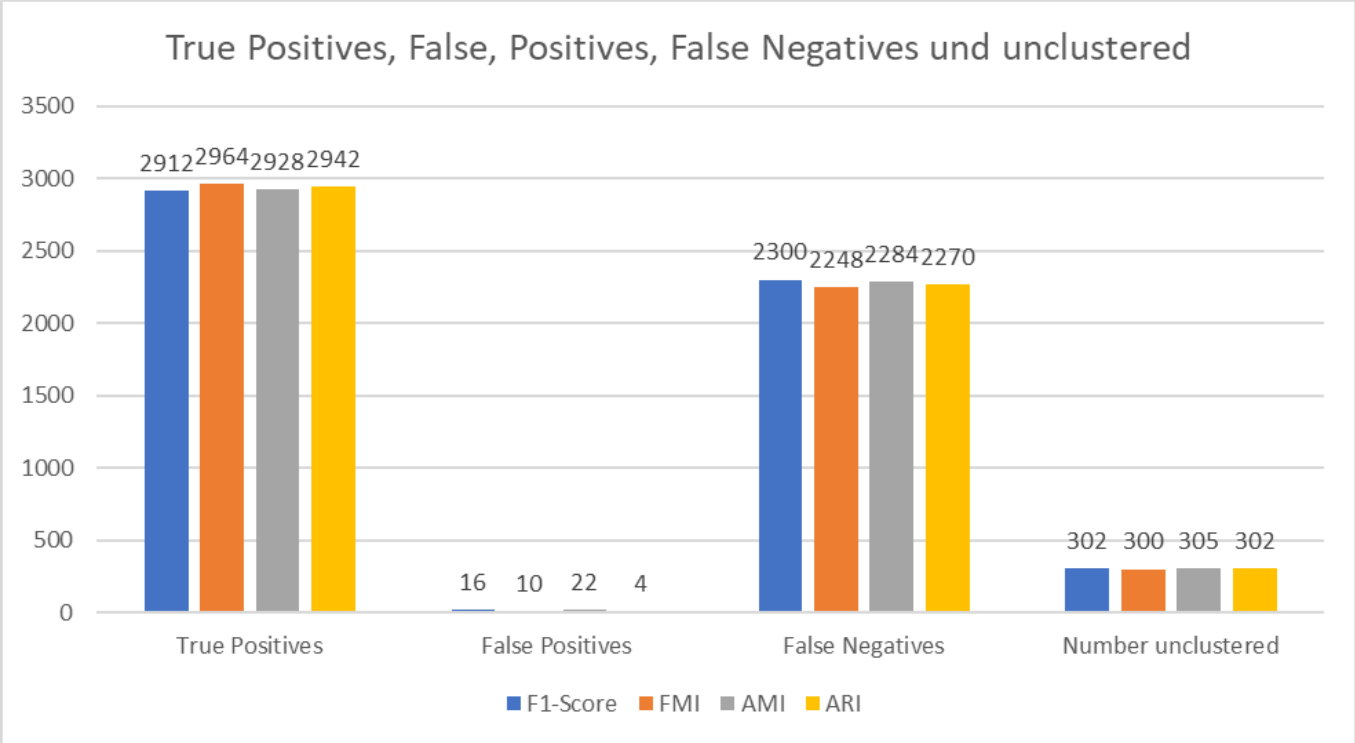
Bei optischer Bewertung, fallen einige nicht-geclusterte Wörter auf, die augenscheinlich beim

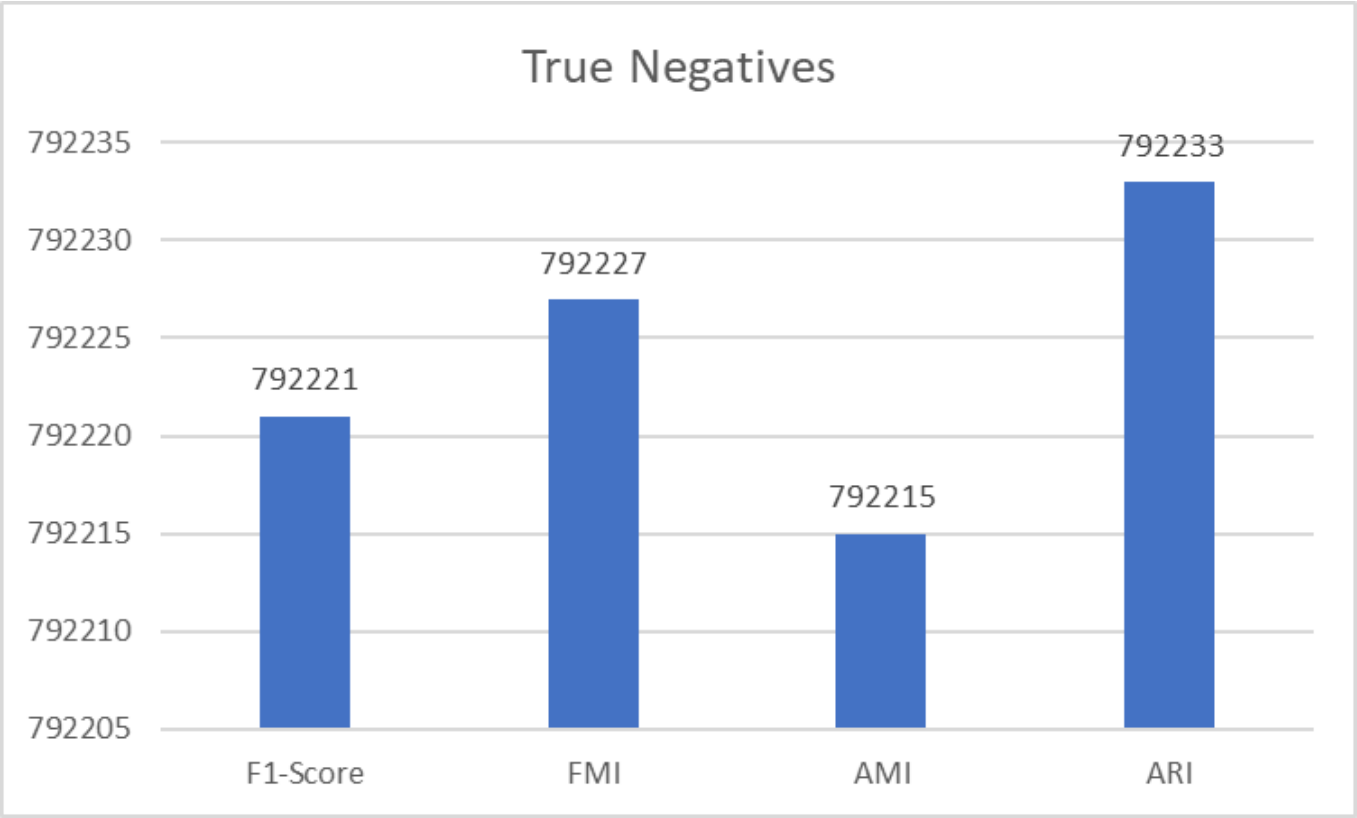
F1-Score geclustert waren (siehe Nidoran-Familie)

FMI.html

Vergleich

Betrachten wir die Konfusionsmatrix und die nicht geclusterten Wörter, haben wir eine Entscheidungsgrundlage zur Auswahl der besten Metrik. Die Ergebnisse basieren auf einer Hyperparameteroptimierung mittels Evolutionärem Algorithmus und sind deshalb zufallsbedingt. Neue Durchläufe können zu anderen Ergebnissen führen.





Ziel soll es sein alle True-Werte zu maximieren und alle False-Werte sowie die unclustered Menge zu minimieren. Dies erreicht am besten FMI oder ARI. In drei von fünf Fällen ist aber FMI besser (True Positives, False Negatives, Number Unclustered), weshalb sich für FMI entschieden wird.