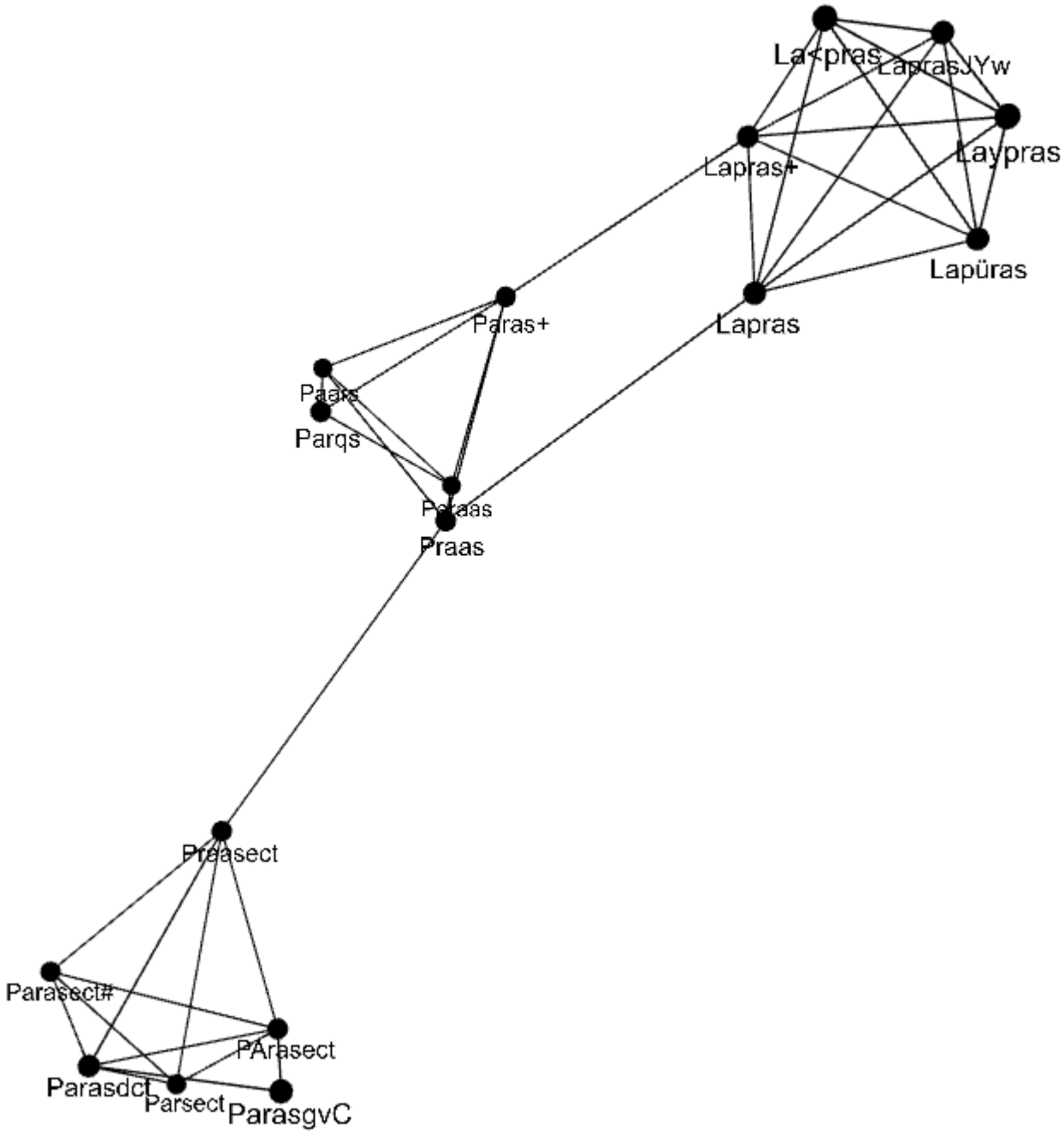
2_Cluster

Herleitung

Beim Clustern ist aufgefallen, dass zwar der Threshold eingehalten wird, aber immer nur paarweise und nicht für das gesamte Cluster. So kann es dazu kommen, dass die Clusterinhalte "wandern".



Dies führt dazu, dass die Cluster größer werden und zu viele Wörter beinhalten (viele false positives). Dies führt zu einer Verschlechterung des F1-Scores

Hypothese

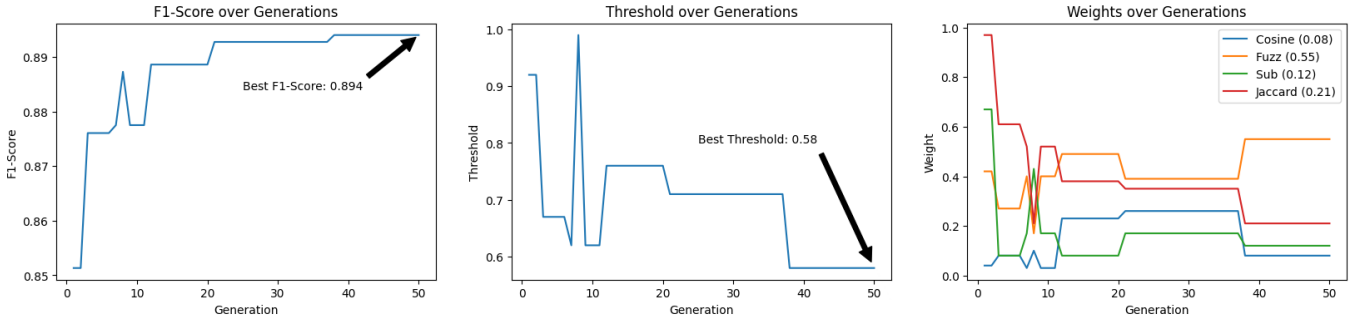
Wenn alle Elemente eines Clusters den Threshold einhalten müssen und nicht nur paarweise die Elemente, erhöht sich die Clusterqualität.

Messung

F1-Score

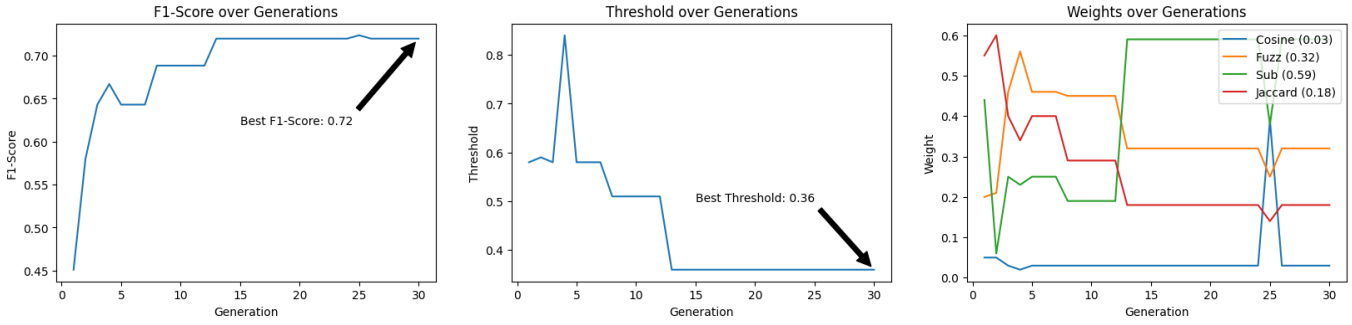
Ergebnisse

Vorherige Situation mit relaxierten Threshold



Best F1-Score = 0.894
Threshold = 0.58
Cos = 0.08
Fuzz = 0.55
Sub = 0.12
Jaccard = 0.21

Ergebnis mit strengen Threshold



Best F1-Score= 0.72
Threshold = 0.36
Cos = 0.03
Fuzz = 0.32
Sub = 0.59
Jaccard = 0.18

Da die Cluster in einer PowerApp validiert werden sollen, ist es einfacher zu große Cluster zu haben, als zu viele zu kleine Cluster. Es werden also lieber false positives, als false negatives akzeptiert