Deutsche Krebsgesellschaft
Classification Usefulness

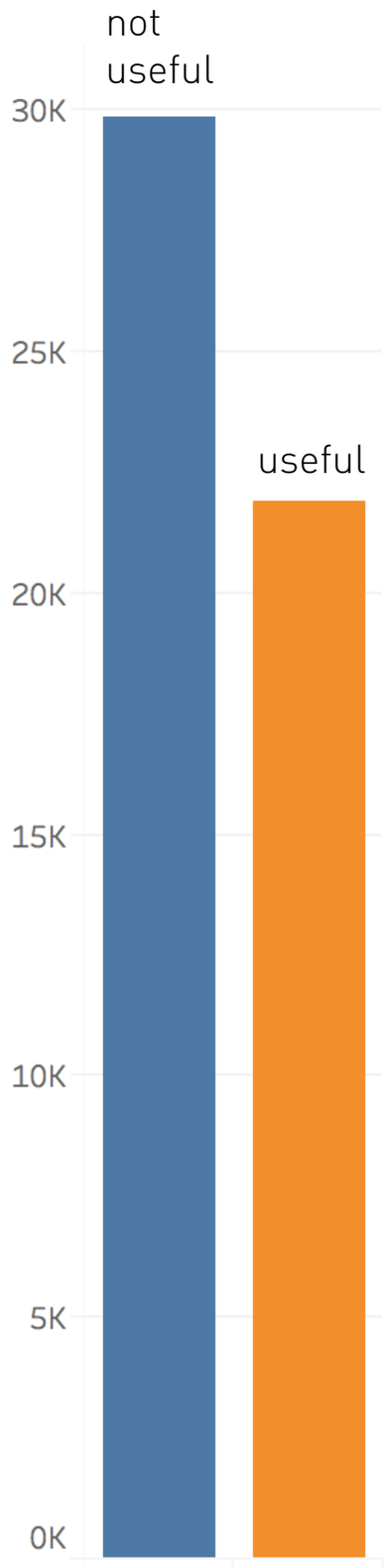
Use case

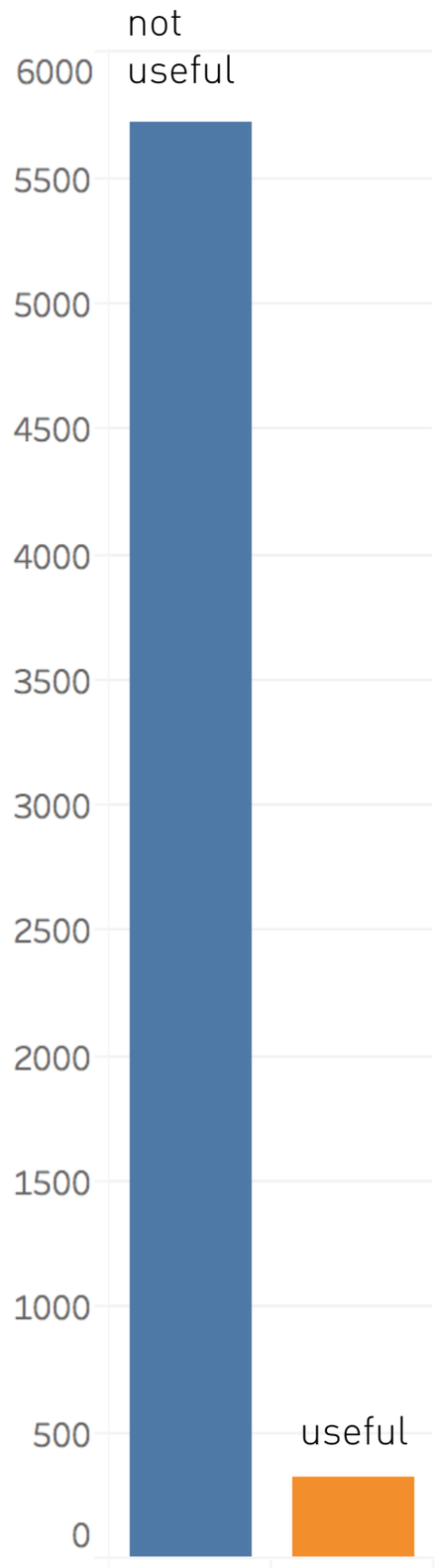
Which article / PDF is useful for the website?

GOAL:

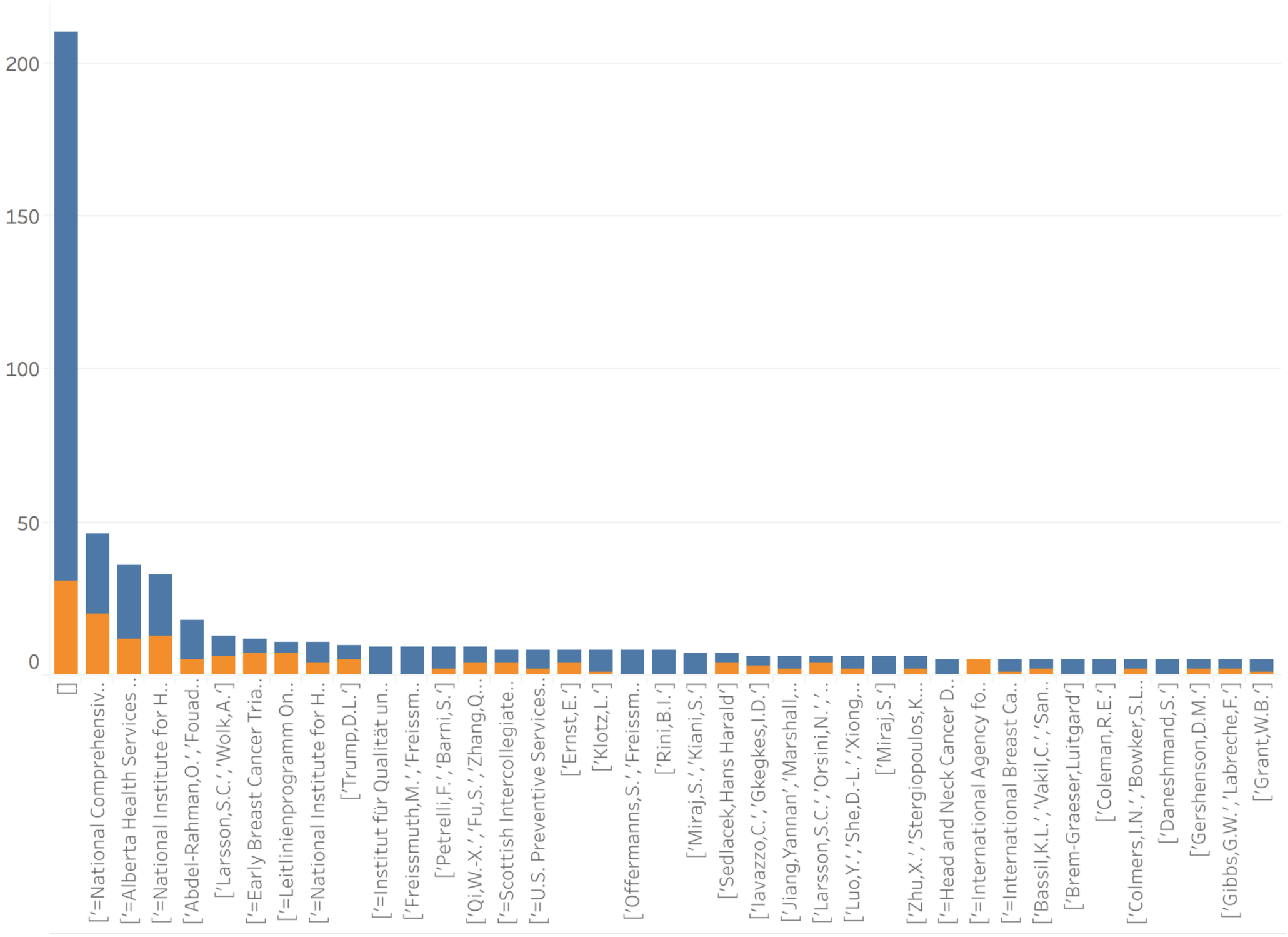
classification of articles with yes/no

CLASSES


not
useful
30K
25K
20K
15K
10K
5K
0K
useful


not
useful
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
useful

AUTHORS


200
150
100
50
0
[]
['=National Comprehensiv..
['=Alberta Health Services ..
['=National Institute for H..
['Abdel-Rahman,O.','Fouad..
['Larsson,S.C.','Wolk,A.']
['=Early Breast Cancer Tria..
['=Leitlinienprogramm On..
['=National Institute for H..
['Trump,D.L.']
['=Institut für Qualität un..
['Freissmuth,M.','Freissm..
['Petrelli,F.','Barni,S.']
['Qi,W.-X.','Fu,S.','Zhang,Q...
['=Scottish Intercollegiate..
['=U.S. Preventive Services..
['Ernst,E.']
['Klotz,L.']
['Offermanns,S.','Freissm..
['Rini,B.I.']
['Miraj,S.','Kiani,S.']
['Sedlacek,Hans Harald']
['Iavazzo,C.','Gkegkes,I.D.']
['Jiang,Yannan','Marshall,...
['Larsson,S.C.','Orsini,N.',...
['Luo,Y.','She,D.-L.','Xiong,...
['Miraj,S.']
['Zhu,X.','Stergiopoulos,K...
['=Head and Neck Cancer D..
['=International Agency fo..
['=International Breast Ca..
['Bassil,K.L.','Vakil,C.','San..
['Brem-Graeser,Luitgard']
['Coleman,R.E.']
['Colmers,I.N.','Bowker,S.L..
['Daneshmand,S.']
['Gershenson,D.M.']
['Gibbs,G.W.','Labreche,F.']
['Grant,W.B.']

First Step

We singled out and analysed variables which could possibly influence usefulness: Author, Abstract ...

BUT we shouldn‘t get the ideal score and instead build something which is scalable and easy to use for DKG in the future

Thus...

ElasticSearch

..we decided to classify on title, subtitle, author and article.

1. indexed the data in Elastic search continuously

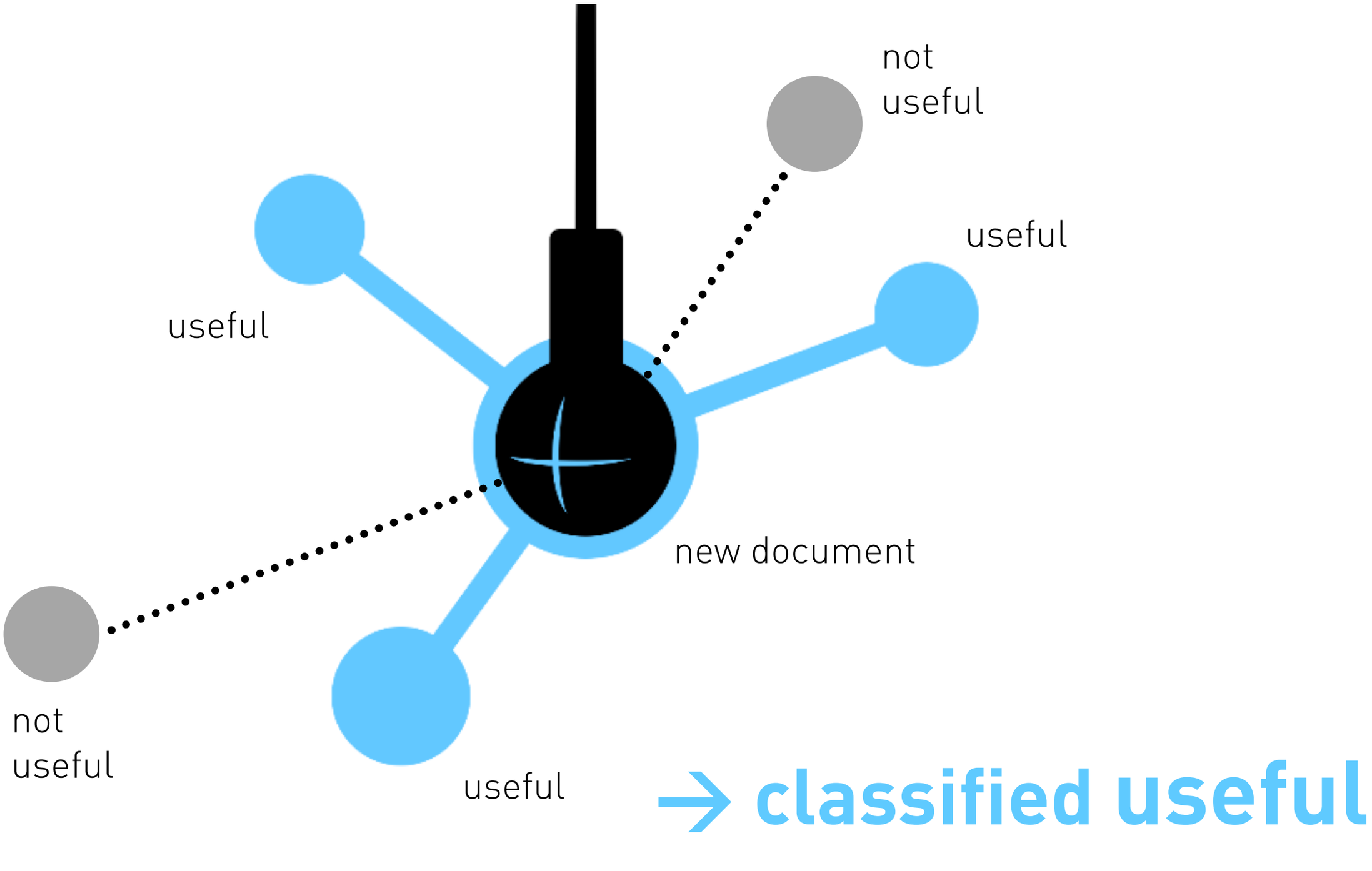
...

ElasticSearch

2. used **k-nearest Neighbor** based on »more like this« query

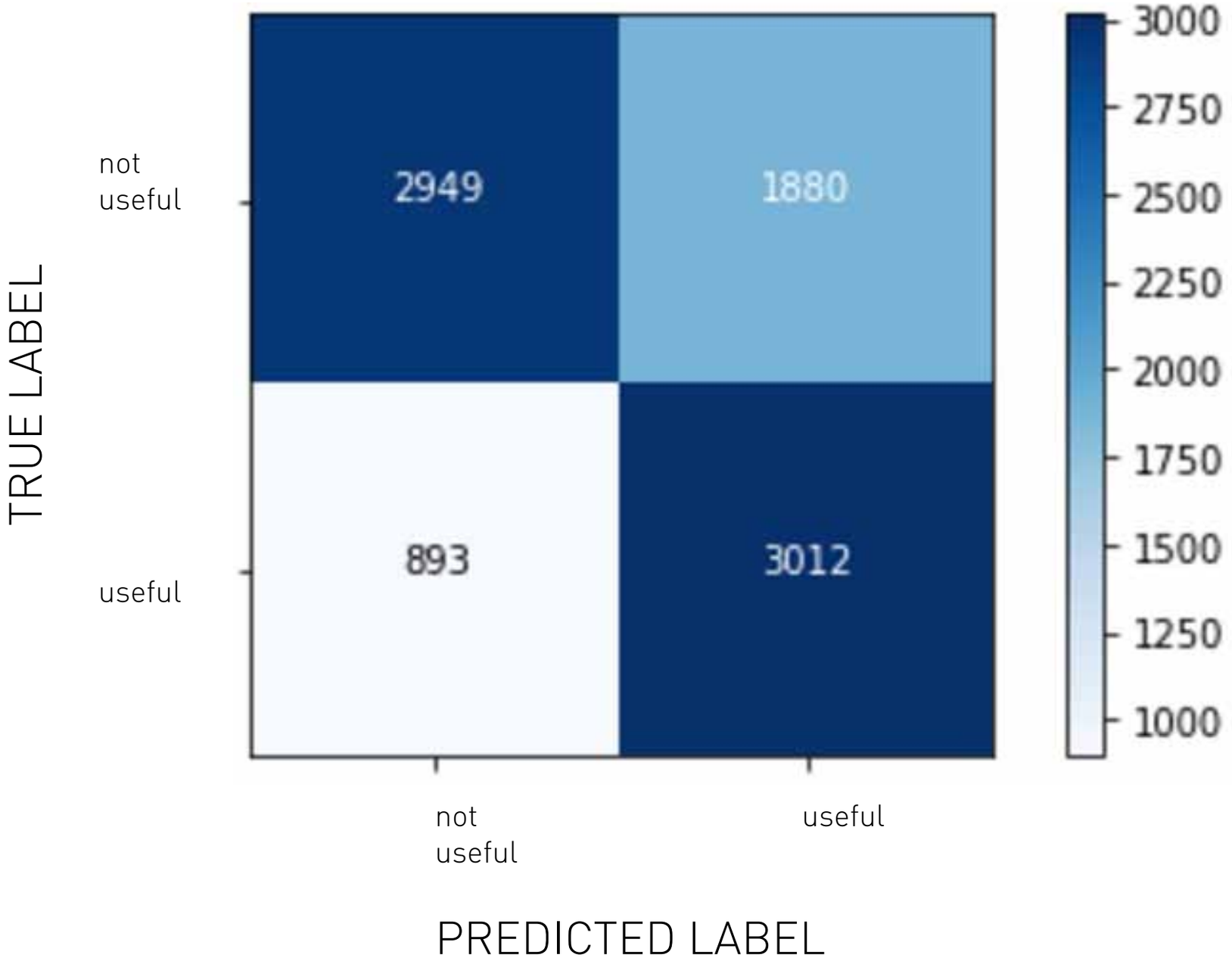
→ under the hood »more like this« applies tf–idf on the articles

k-nearest Neighbor



RESULTS

confusion matrix, without normalisation



Why ElasticSearch?

→ no extra training time needed

→ no additional infrastructure

→ adapts to new data in real time

→ updates runs continuously

Danke von unserem Team!

Till Breuer
breuertill@gmail.com

Anke Nowottne
anke.nowottne@gmail.com

Oliver Rieger
o.rieger@posteo.de

Franziska Wittleder
franziska.wittleder@gmail.com