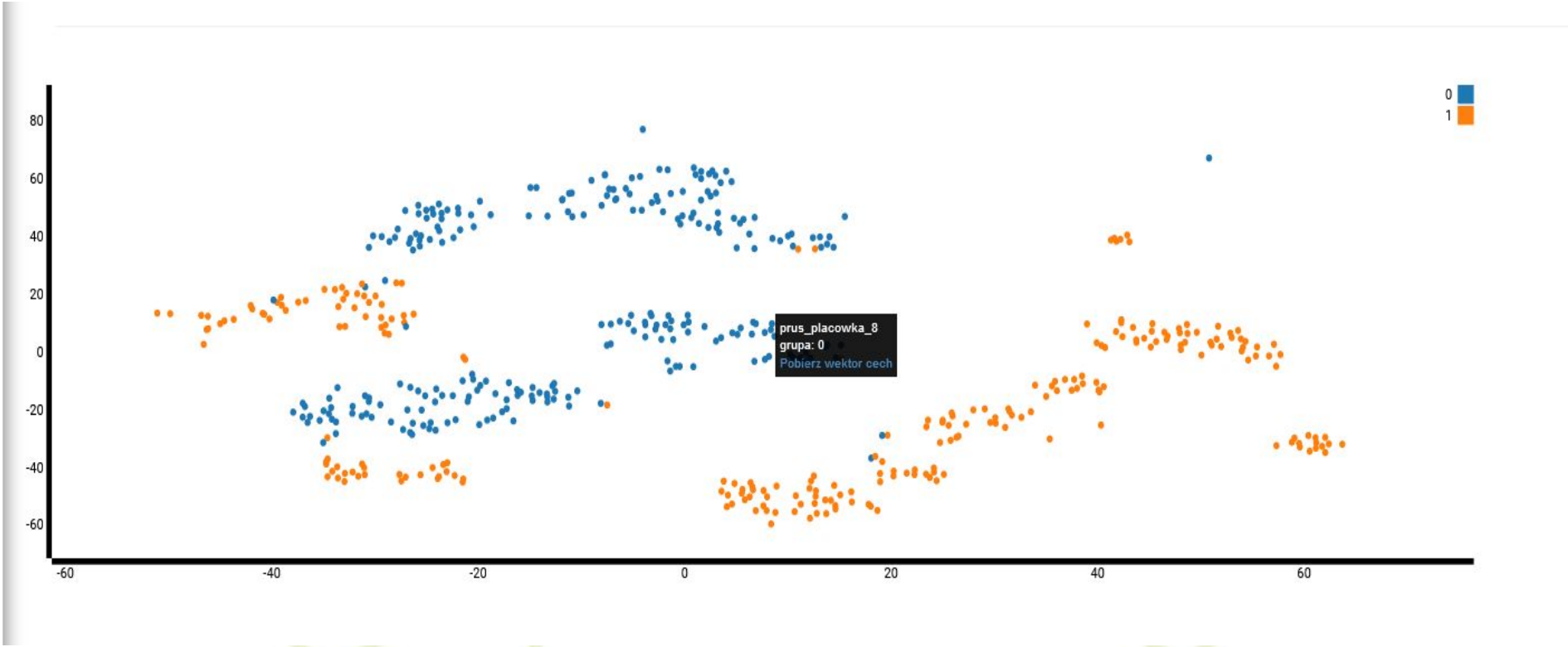
Evaluating the quality of dimensionality reduction algorithms

Author: Mateusz Gniewkowski

Wednesday 7 October, 12.45 - 13.45

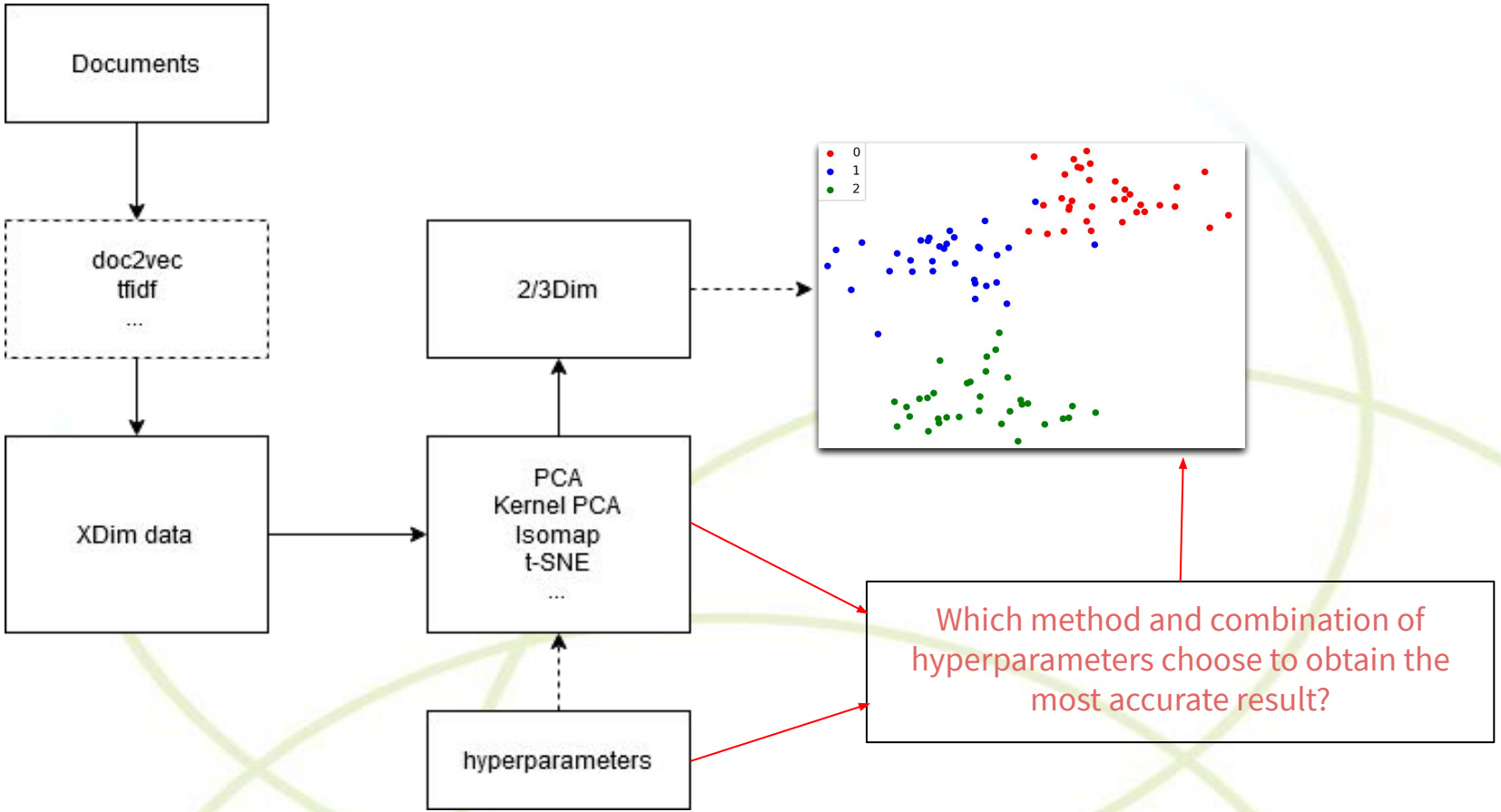
CLARIN

The problem



https://ws.clarin-pl.eu/websty.shtml?en

The problem


Documents
doc2vec
tfidf
...
XDim data
PCA
Kernel PCA
Isomap
t-SNE
...
hyperparameters
2/3Dim
0
1
2
Which method and combination of hyperparameters choose to obtain the most accurate result?

The solution

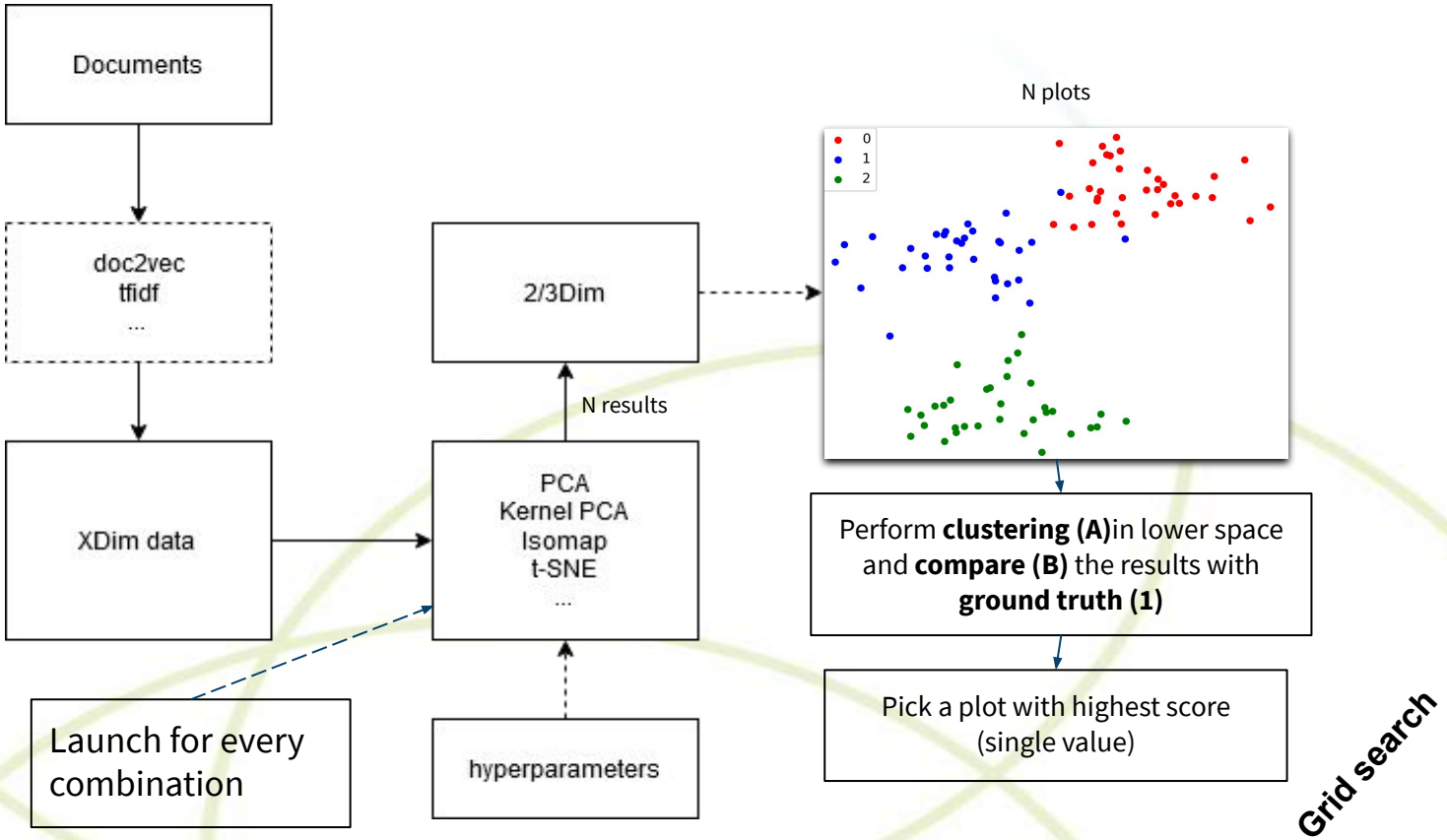
Prerequisites:

1. **Labeled** data

Notes:

A. Any method will do. We used **Agglomerative Clustering** with Euclidean distance (probably the distance measure) and average linkage criterion.

B. Any measure of the similarity between two data clustering. We used **ARI** (Adjusted Rand Index), and **AMI** (Adjusted Mutual Information).



The results (datasets)

Two collections of text documents

WIKI

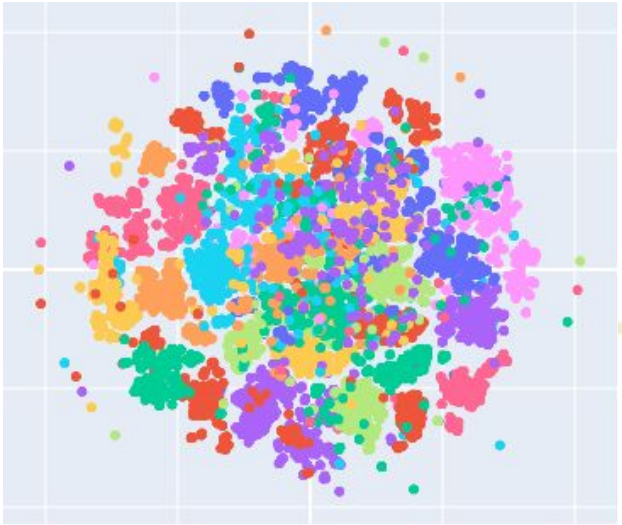
1. Language: Polish
2. 34 categories, doubtful assignments
3. 9,837 documents
4. some classes are underrepresented

PRESS

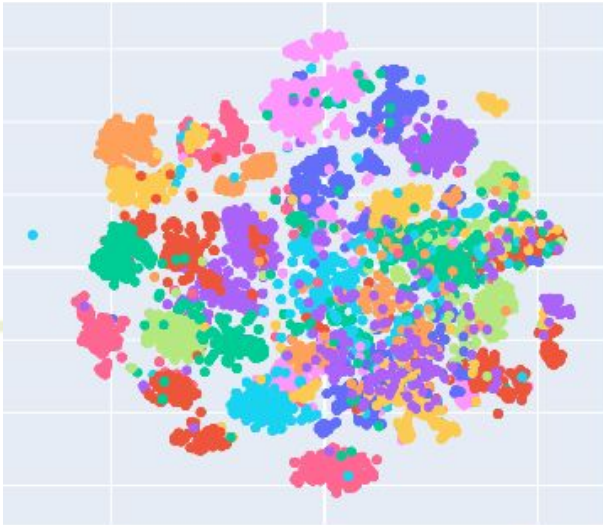
1. Language: Polish
2. 5 categories, well separated
3. 6,564 documents
4. well balanced

The results (WIKI)

score: 0.34

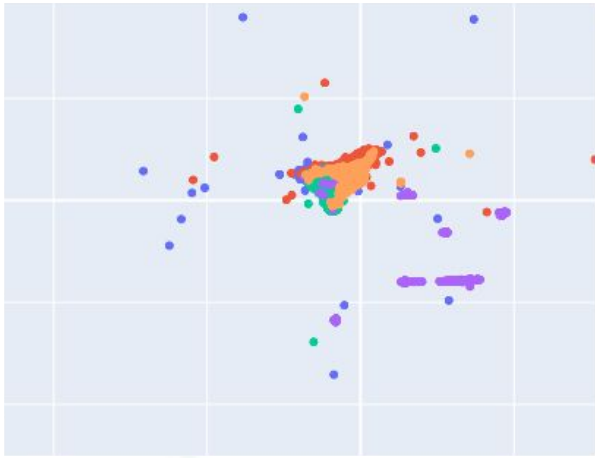


score: 0.46

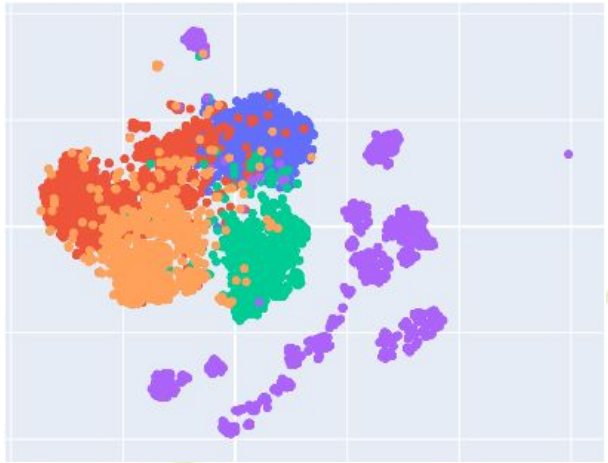


score: 0.56

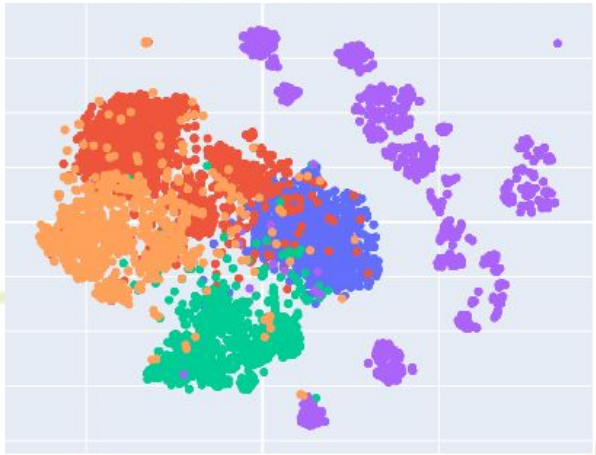
The results (PRESS)



score: 0.21



score: 0.39



score: 0.52

Todo list

1. Is the score fits people judgement?
    a. opinion poll
2. What with unlabeled data?
    a. You could use clustering algorithm in high dimensional space, but the obtained results might not fit the reality
3. How to adjust the clustering algorithm to emphasize the desired features
4. Fix the problem with multimodal groups

Ask me about those at poster-style discussion (14.50 - 15.50)

Evaluating the quality of dimensionality reduction algorithms

Thank you for your attention :)

Author: Mateusz Gniewkowski

Wednesday 7 October, 12.45 - 13.45

CLARIN