Input data
- subject
- body

→ ML classifier →

- Non-architectural
- Existence
- Technology
- Process
- Property

Input data
↓ raw data
Text preprocessing
- concatenate subject + body
- lowercase → NLTK, SKLEARN
- remove replies ← Handmatig
- remove HTML ?

↓ preprocesed data = clean string

Feature generation
- Bag of Words - CountVectorizer sklearn
- TFIDF - TfidfVectorizer sklearn

↓ vectors

ML classifier
- Decision tree classifier
- Random Forest classifier
- SVM - linear SVC
- Naive Bayes

Gridsearch CV sklearn
- Training ↔ k-fold cross validation
- Testing
- Optimizen ML classifier

↓

Resultaten

BOW

Sparse

Ik gooi een bal ] dataset → [1 1 1 1 0] → [0 1 2 3]
Ik gooi een steen ] → [1 1 1 0 1] → [0 1 2 4]

0 ik
1 gooi
2 een
3 bal
4 steen

Sparse vector

DF
1 Ik gooi een een bal 1
2 Ik gooi een steen 1/2 +

een: TF = 2
DF = 2

TF = 1
DF = 2