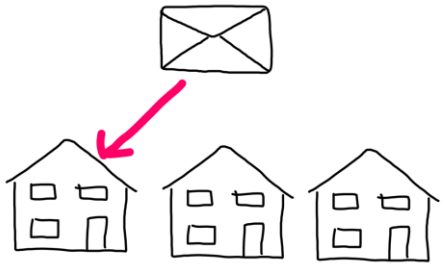
Determining department from email content using Fuzzy Logic

For our BSc project:
Fundamentals of Fuzzy Logic



Peter Heemskerk
Jim Kamans
Stefan Schenk

December 2017

Goal and relevance

For complex organisations it is difficult to answer email questions from customers / other parties in time.

Our goal:

use e-mail content to determine the correct department which should reply on an email.

Literature

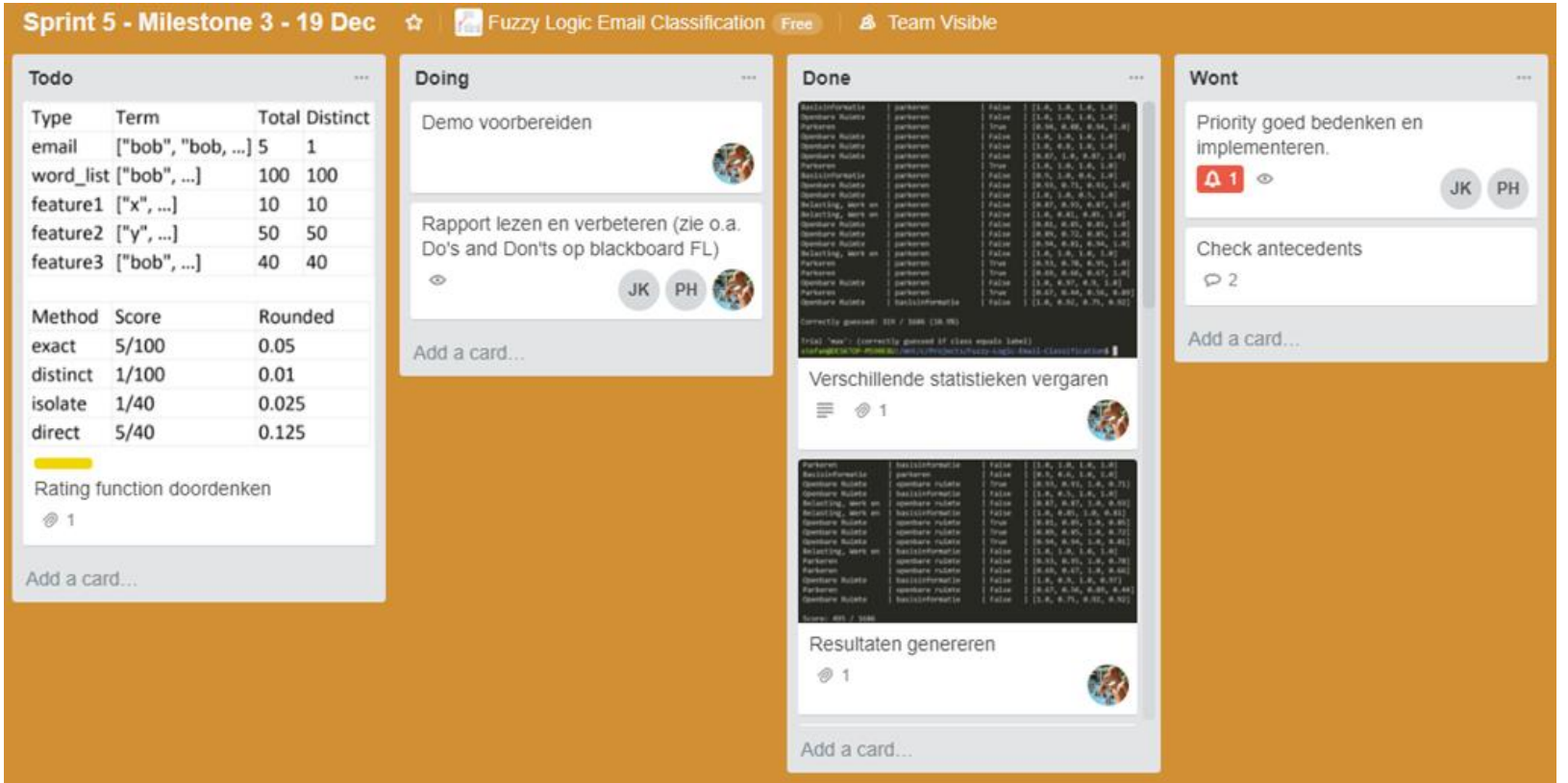
- Fuzzy Logic has been used for e-mail classification (spam, anti-phishing).

- Main idea from Ferolin:
  - extract from e-mail content linguistic generalised features (like “ technical “, “ financial “, “ emotional”).
  - use these linguistic features as input for a fuzzy logic system to determine the departements.

Collaboration

Stefan: code and experiments
Peter: fuzzy logic and reporting
Jim: “vliegende keep” (trouble shooting)

Collaboration with tools: Github, Trello, Google Drive



Approach

training corpus
of emails

term frequency /
inverse document
frequency

data preparation

feature
word- lists

email content
tokenized
and stemmed

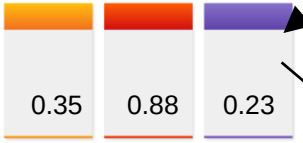
data preprocessing

cleaned
email words

overlap with
feature word list

ranking

feature
score



combination
of features
makes department
(fuzzy logic)

classification

department

Data after preprocessing and Ranking

{ ‘handhav’, ‘rijbewijs’, ‘kopie’, ‘daadwerk’, … }

⇒

{‘handhav’: 2, ‘rijbewijs’:1, ‘kopie’: 1, … }

⇒

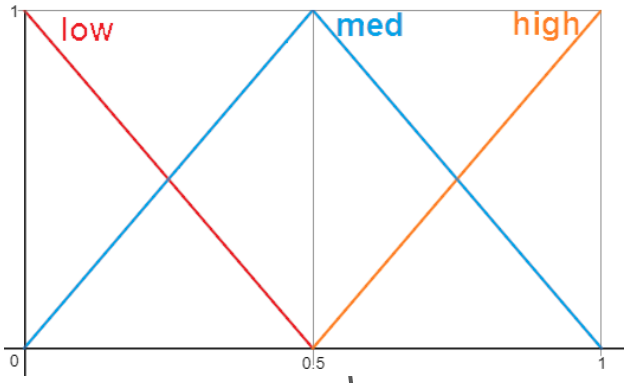
{‘handhav’: 0.0 0.0 0.2 0.0,
‘rijbewijs’: 0.0 0.0 0.1 0.0,
‘kopie’: 0.0 0.1 0.1 0.1, … }

⇒

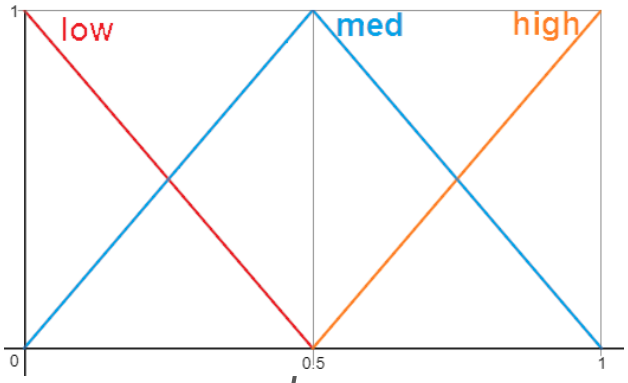
{'basisinformatie': 0.0, 'belasting, werk en inkomen': 0.5, 'openbare ruimte': 1.0, 'parkeren': 0.3}

Implementation

[0.0, 0.5, 1.0, 0.3]

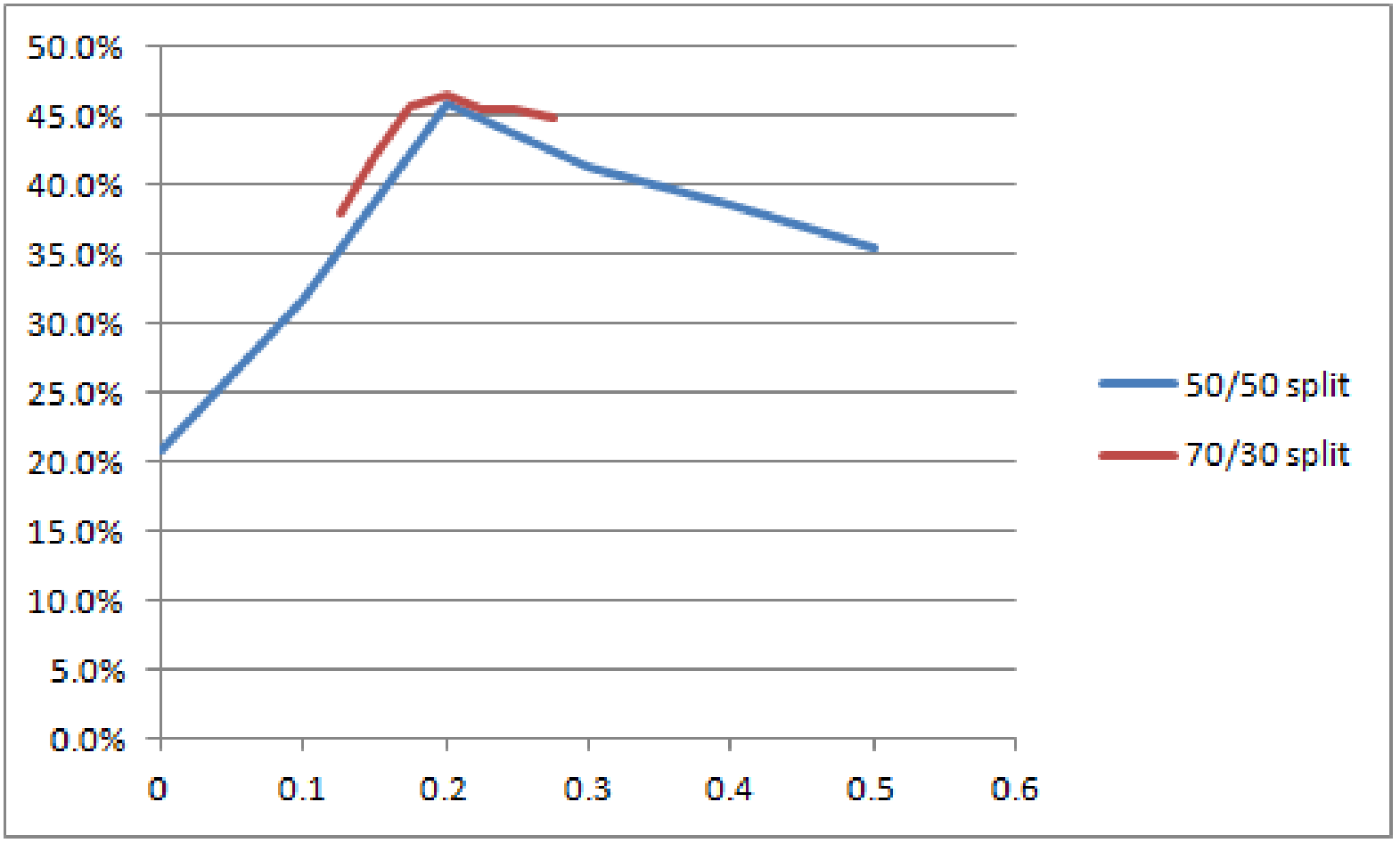


IF f1 = low THEN d1 = low
IF f1 = med THEN d1 = med
IF f1 = high THEN d1 = high
: : :
:
IF f1 = low & f2 = low & … THEN d1 = high



openbare ruimte

Experiments and Results



Results, Observations and Discussion

Accomplishments:

- baseline
- focus on feature word list generation
- basic fuzzy logic implementation
- for translating 4 features into 4 departments (feature “Parkeren” into department “Parkeren”)
- score of 46% correctly classified by 70/30 split

Improvements:

- extend feature baseline (based on word clustering of large email dataset)
- learn fuzzy logic rules from a large labeled training dataset.