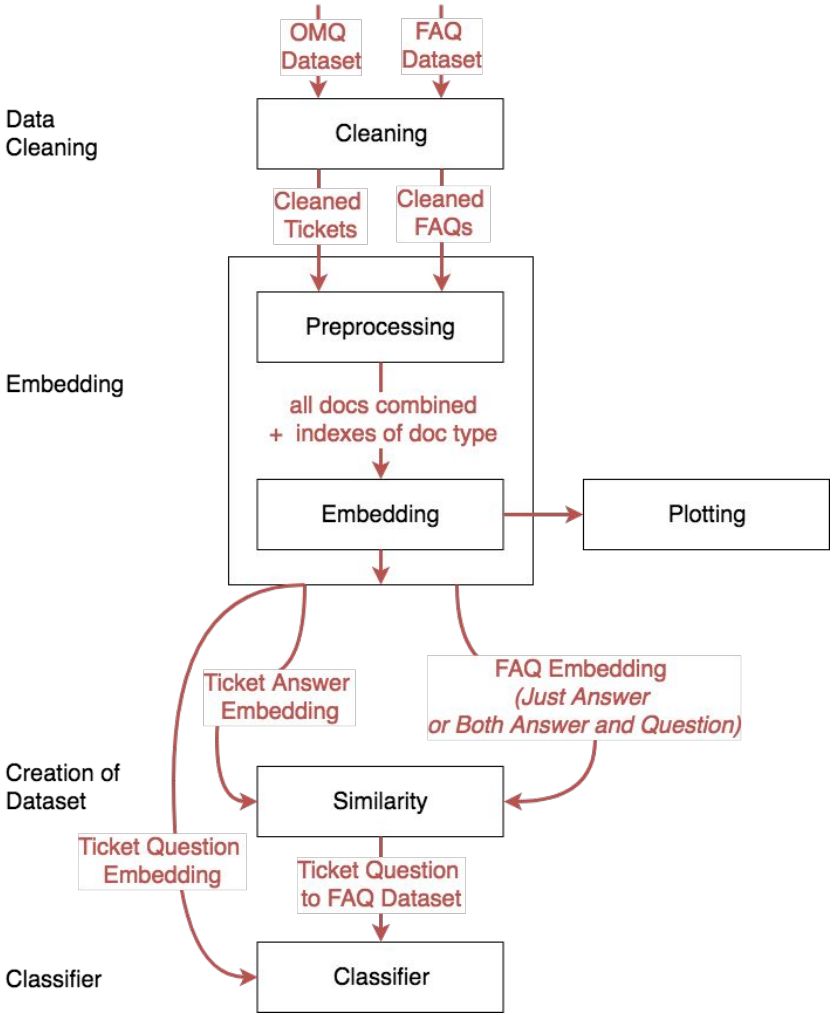
ETH IT Service and QA

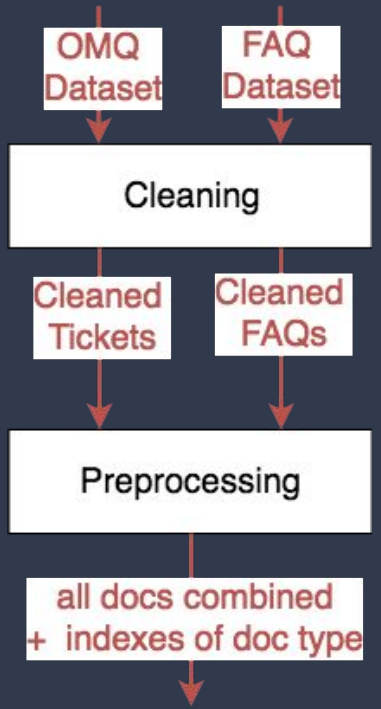
Costanza Calzolari
Margherita Rosnati
Brian Regan

11.04.2019

Overview *pipeline*



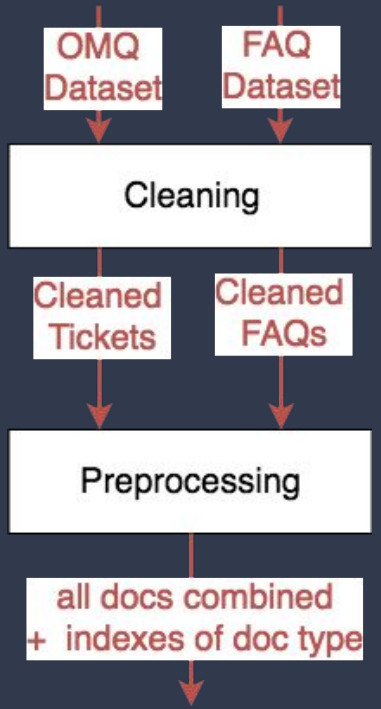
Cleaning



Heuristics - OMQ dataset:

- Drop system articles
  - System tagged
  - “Close!”, “Priority Update”, …
  - ***-> remove redundant articles***
- Drop appended previous messages
  - tokens : “--”, “_”, “From”, …
  - ***-> remove redundant article parts***
- Concatenate:
  - all client articles: questions
  - all agent replies: answers
  - ***-> create {question, answer} dataset***
- Drop empty questions / answers
- Translate (Google API)
  - ***-> consistency of dataset***

Cleaning

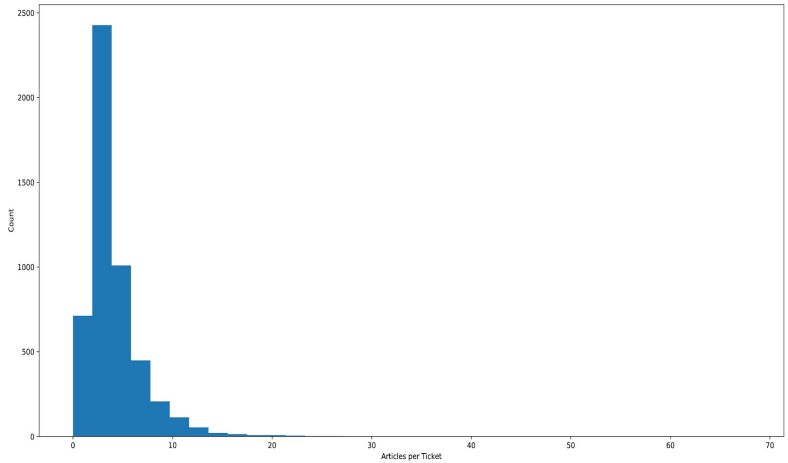


Heuristics - FAQ dataset:

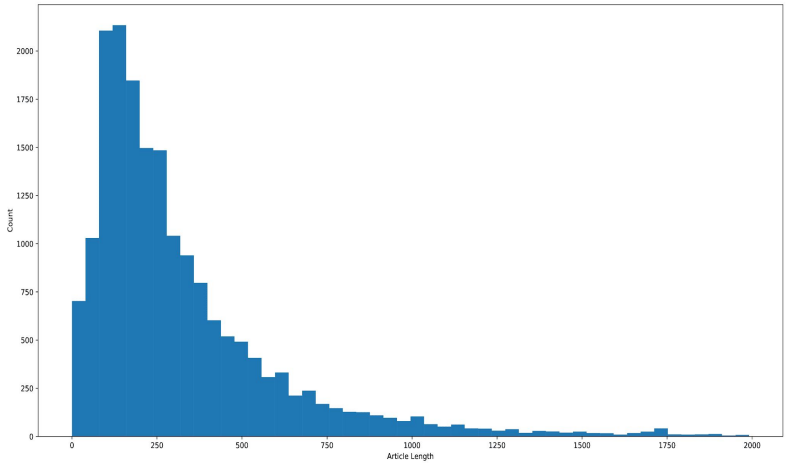
- Convert HTML into raw text
- Inner join of question and answer tables
  - ***-> create {question, answer} dataset***
- Translate (Google API)
- Drop duplicate questions / answers
- Drop empty answers
  - ***-> consistency of dataset***

Data Intuition

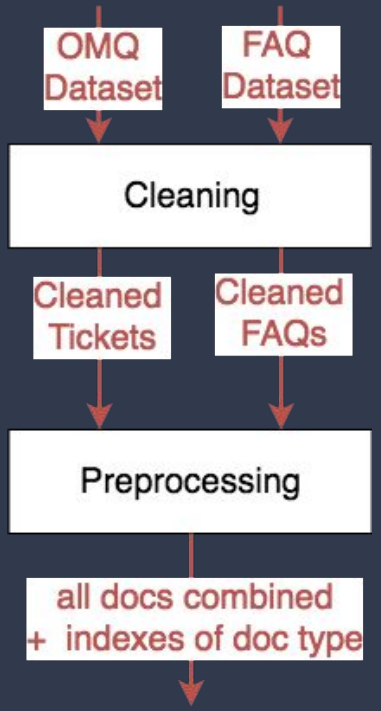
Articles per ticket - distribution



Article length - distribution



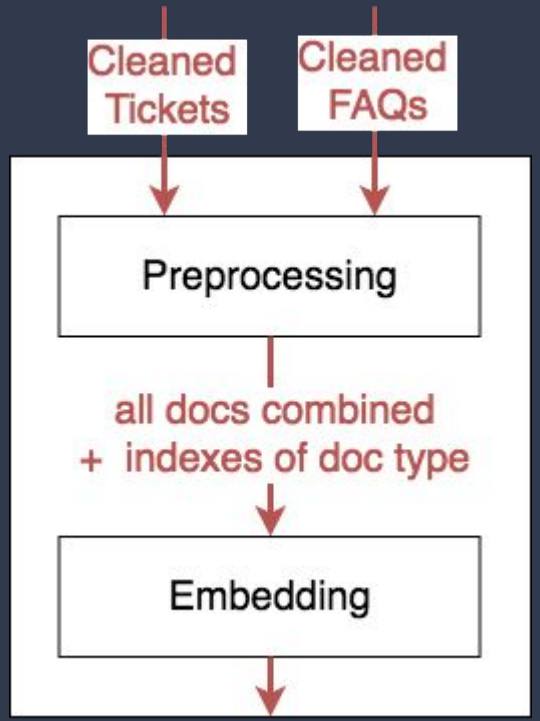
Preprocessing



- `Tf-idfVectorizer` (full details line 1310)
  - Remove accents
  - Set to lowercase
  - Remove spaces, punctuation
  - Tokenise
- Custom function (full details)
  - Same as `Tf-idfVectorizer` plus
  - Sub IP addresses with tag `<IP>`
  - Sub emails with tag `<email>`
  - Drop stopwords (“Dear”, “Sincerely”, ...)

-> limited word space

Embedding

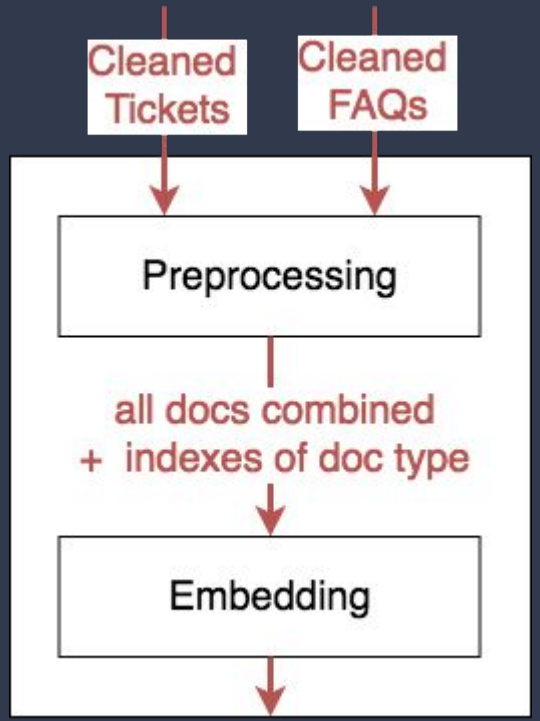


Tf-idf:

Frequency based approach

- Roughly: frequency of word in document over frequency of word over corpus
  - Eg: “VPN”: high score
  - Eg: “the”: low score
- Document = vector of frequency count for all words in dictionary

Embedding



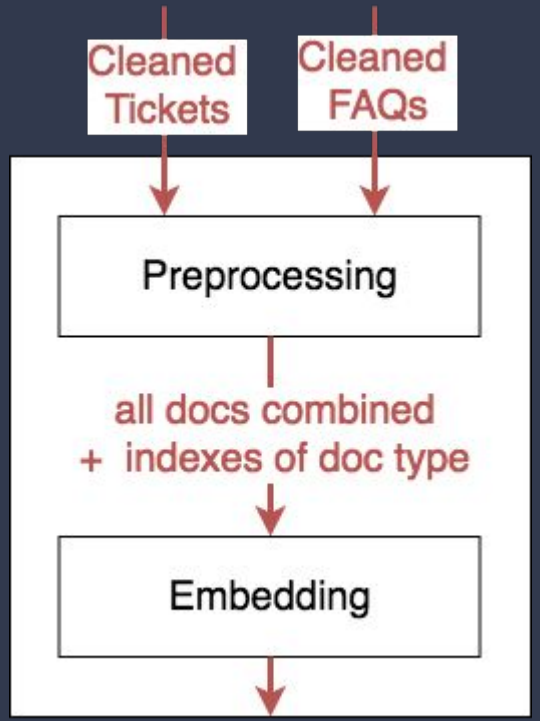
Word2Vec:

Frequency and semantic based approach

- Roughly: learn context of word in document corpus
  - Words with similar context have similar representations
  - Eg: “email” and “message”

- Word representation: needs further steps to create a document representation
- Learning method: more examples improve the optimisation

Embedding



Word2Vec - continuation:

Frequency and semantic based approach
From word rep to document rep

Document = comb. of word2vec vectors:

- Mean of word2vec vectors
- Tf-idf-weighted representations
  - Mean of 5 vectors representing highest scoring words
  - Mean of 5 (as above) with tf-idf weighting
  - Mean of all word2vec vectors with tf-idf weighting

Similarity tagging *heuristics*

Find the closest FAQ for each ticket

Validation data shows 75% tickets not answered by an FAQ

-> non-FAQ tickets

Similarity mapping to mirror validation dataset

For each ticket answer:

-> keep the FAQ answer with highest similarity

-> keep said FAQ similarity score

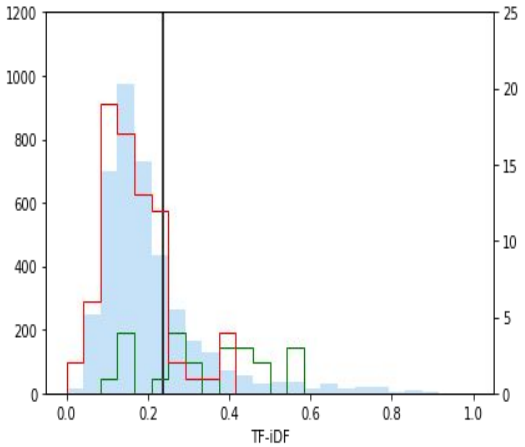
-> if similarity score belongs to 75th percentile

then ticket is tagged with non-FAQ

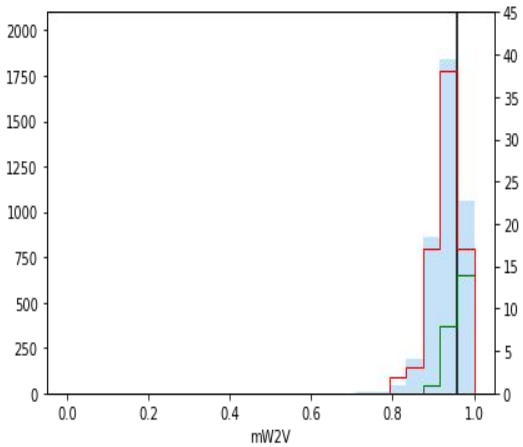
Document similarity (vector cosine)

Distribution of highest FAQ similarity score per ticket

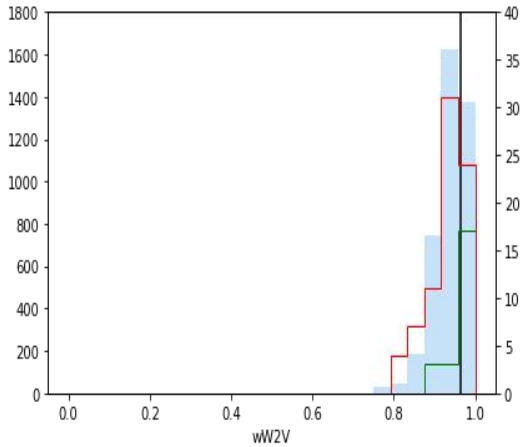
TFiDF



word2vec average



word2vec tf-idf weighted



Light blue bars	Untagged data
Red line	Validation data, non-FAQ class (right scale)
Green line	Validation data, FAQ class (right scale)
Black line	“non-FAQ” threshold

Examples of similar answers (Tf-idf)

Strong: (sim score 57%)

tick_ans: Dear xxx Could you tell us your number? It should be printed on a blue sign with the inscription “Betriebinformatik” or “ETH Zurich ID Services Delivery”. Your ID team Rudolf xxx. Dear xxx, A support ticket has been opened for your request. We will process your request as soon as possible. If you have any further questions, please answer directly to this e-mail.

FAQ_ans: ZO computer number Could you tell us your computer number? It should be printed on a blue sign with the inscription "Betriebinformatik" or "ETH Zurich ID Services Delivery".

Medium: (sim score 31%)

tick_ans: Dear xxx I see you have already been to ETH Zurich. Therefore, the old password will still be valid on [1] www.passwort.ethz.ch. If you do not know this, we would have to send you a new one. Your ID Team Joel Greuter Dear xxx Unfortunately, we can only regenerate the password. In this case we can either send it by mail or you can pick it up from us. Your ID Team Joel Greuter

FAQ_ans: Password forgotten Unfortunately we are not able to send your new password by email or by phone. You can pick up your new password at [...]

Weak: (sim score 13%)

tick_ans: Hoi Shiva I have ordered toner for p-hg-h-12 and p-hg-h-42. [...] Which printer do you mean with p-hg-j-33-1? [...] Did they get lost in the mail or did they end up at another printer? [...]

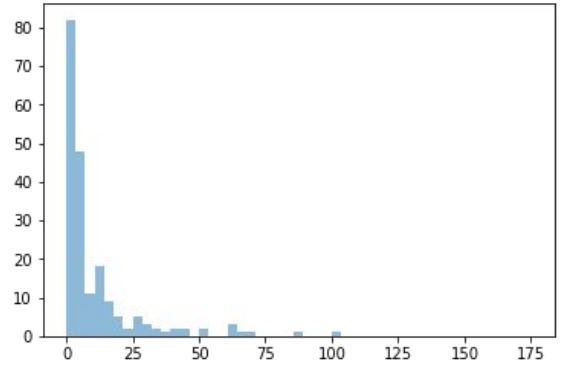
FAQ_ans: Unfortunately we are not allowed to give out the password by phone or e-mail. You can pick it up at our service desk (HG E 11) or we will send you the password home. In this case, please provide us with the following information: Date of birthAdress If you wish to pick up the password from us, we are open from 9.30 am to 11 am and from 1.30 pm to 4 pm. Take your Legi resp. Your employee badge with.

Similarity tagging *challenges*

- Too many classes with one element

  Classes distribution (excl. non-FAQ):

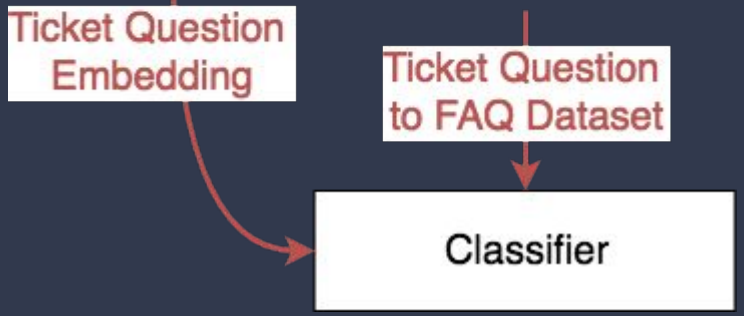
  Number of tickets associated to each class



- Large and diverse non-FAQ class

  75% of tickets, no commonalities

Classification *heuristics & challenges*



Heuristics:

Classify ticket questions using FAQ classes

Use a random forest classifier

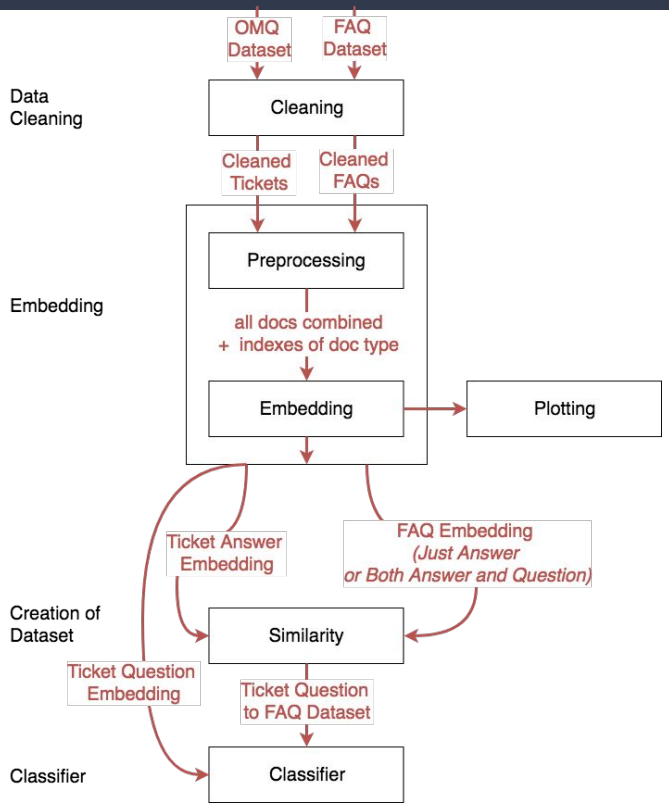
-> nonparametric

-> robust to overfitting

Challenge:

Max as good as the similarity tagging

Results



Model	Avg. Precision	Avg. Recall	F1-Score
Predict all as Non-FAQ	72.0%	36.0%	48.0%
Doc2Vec	71.1%	36.0%	47.8%
Mean Word2Vec	83.4%	36.4%	50.7%
TF-iDF	**89.4%**	**38.1%**	**53.4%**
TF-iDF weighted Word2Vec	84.0%	36.9%	51.3%

Tackling challenges *large non-FAQ class*

Idea: pre-classify non-FAQ tickets

Attempts:

- Random forest pre-classification
  Two classes: FAQ and non-FAQ
- Similarity pre-classification
  Max sim. between questions and FAQ thres

Tackling challenges *one-shot classes*

Idea: use distribution over class

Attempts:

- Instead of one FAQ per ticket answer, keep 5 highest scoring FAQ under similarity
- Classify using similarity weighted training set

Conclusions & further work

Conclusions:

Automate selection proposal of relevant FAQ given an incoming ticket

1. Unsupervised mapping from ticket answers to FAQ
2. Use mapping to train classifier on ticket questions

Further work → self updating FAQ list

- Deal with the “non-FAQ” class
  E.g. cluster and propose new FAQ automatically
- Remove FAQ no longer in use
- Detect tickets requiring action vs. informative tickets