Email classification for semi-automated reply generation

I256: Applied Natural Language Processing
Semester project

Hannes Hesse

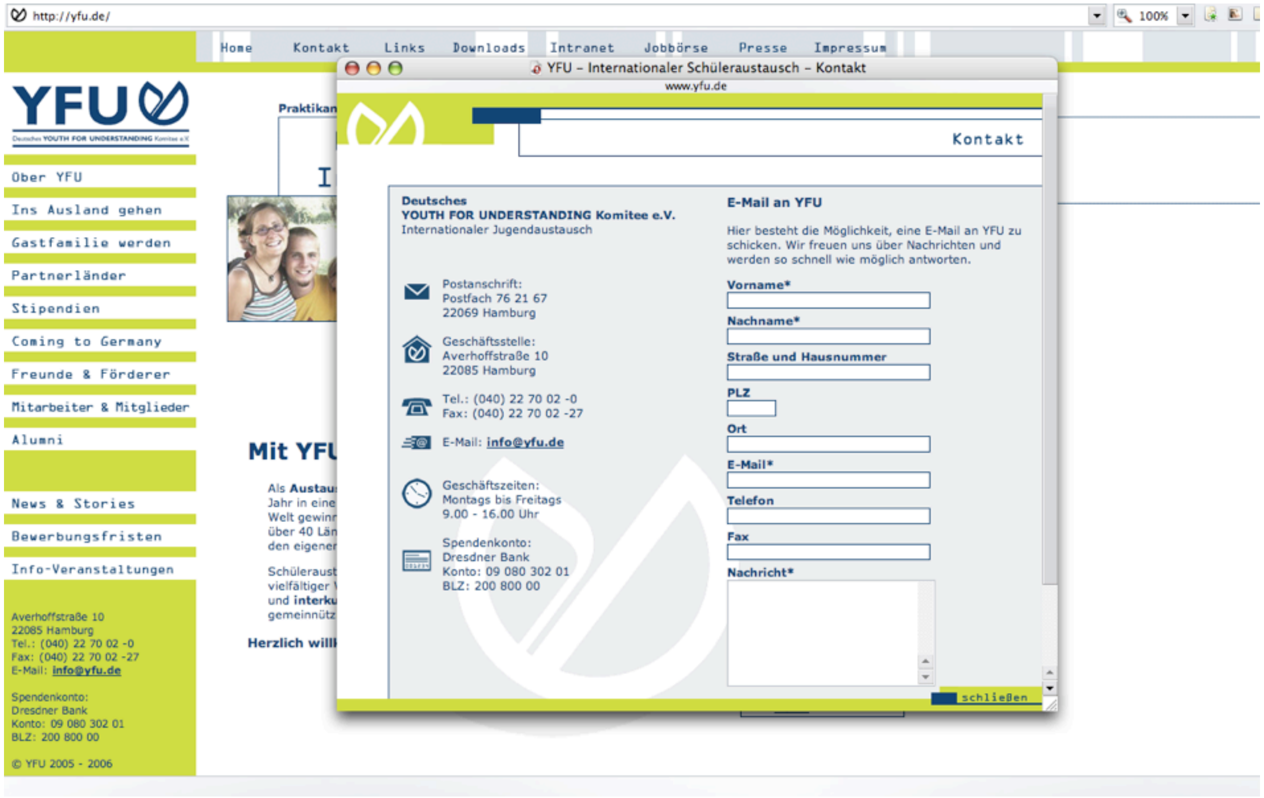
Background

Youth For Understanding,
organizer of student exchange

Sends ~1000 German students abroad each year

Largely run by volunteers, not heavily staffed

Background



Averhoffstraße 10
22085 Hamburg
Tel.: (040) 22 70 02 -0
Fax: (040) 22 70 02 -27
E-Mail: info@yfu.de

Spendenkonto:
Dresdner Bank
Konto: 09 080 302 01
BLZ: 200 800 00

© YFU 2005 - 2006

Deutsches
YOUTH FOR UNDERSTANDING Komitee e.V.
Internationaler Jugendaustausch

Postanschrift:
Postfach 76 21 67
22069 Hamburg

Geschäftsstelle:
Averhoffstraße 10
22085 Hamburg

Tel.: (040) 22 70 02 -0
Fax: (040) 22 70 02 -27

E-Mail: info@yfu.de

Geschäftszeiten:
Montags bis Freitags
9.00 - 16.00 Uhr

Spendenkonto:
Dresdner Bank
Konto: 09 080 302 01
BLZ: 200 800 00

E-Mail an YFU

Hier besteht die Möglichkeit, eine E-Mail an YFU zu schicken. Wir freuen uns über Nachrichten und werden so schnell wie möglich antworten.

Vorname*

Nachname*

Straße und Hausnummer

PLZ

Ort

E-Mail*

Telefon

Fax

Nachricht*

www.yfu.de: 10 - 15 inquiries per day

Typical questions

“Please send me application material or brochures”

“I am XY years old. Can I still apply?”

“Is it too late to apply now for year XY?”

“Can I go abroad for six months?”

“We would like to host an exchange student.”

“Where’s my stuff?”

Objective

Classify emails for

- Routing
- Auto-suggest feature in CRM system
- Analysis

Data Preprocessing

1. Convert HTML emails in .mbox format to text files

2. Extract author name and message body, discard headers and HTML

3. Convert special characters, remove capitalization

4. Tokenize words

5. Stemming, using German version of Porter algorithm

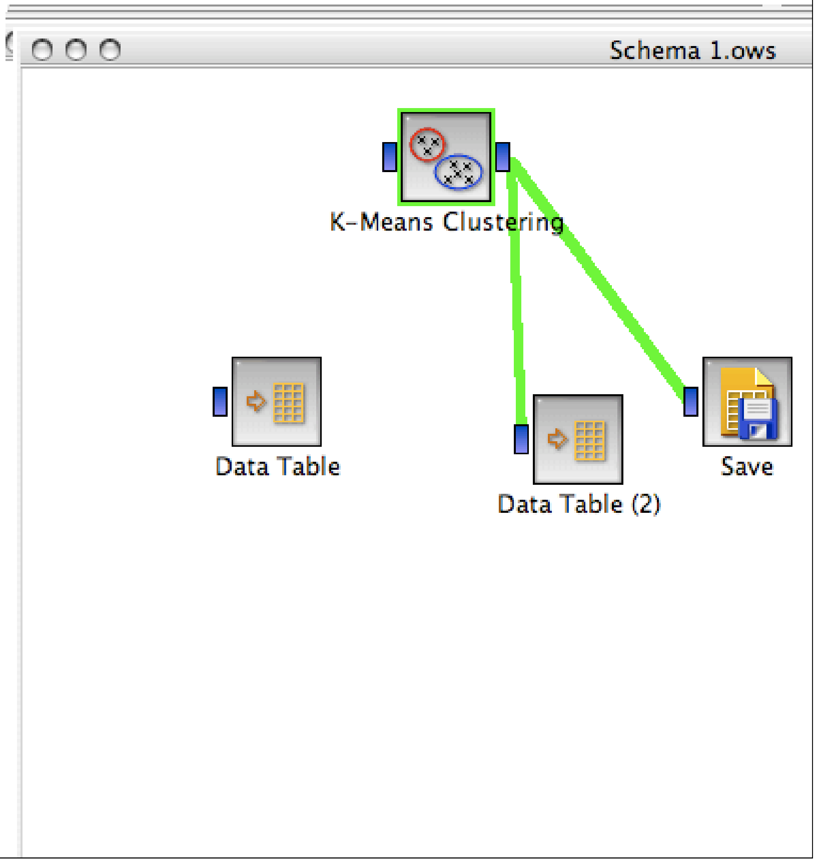
Training data

Original idea: Unsupervised clustering for category exploration (n =~ 1800)

Used Orange for data exploration and clustering experimentation

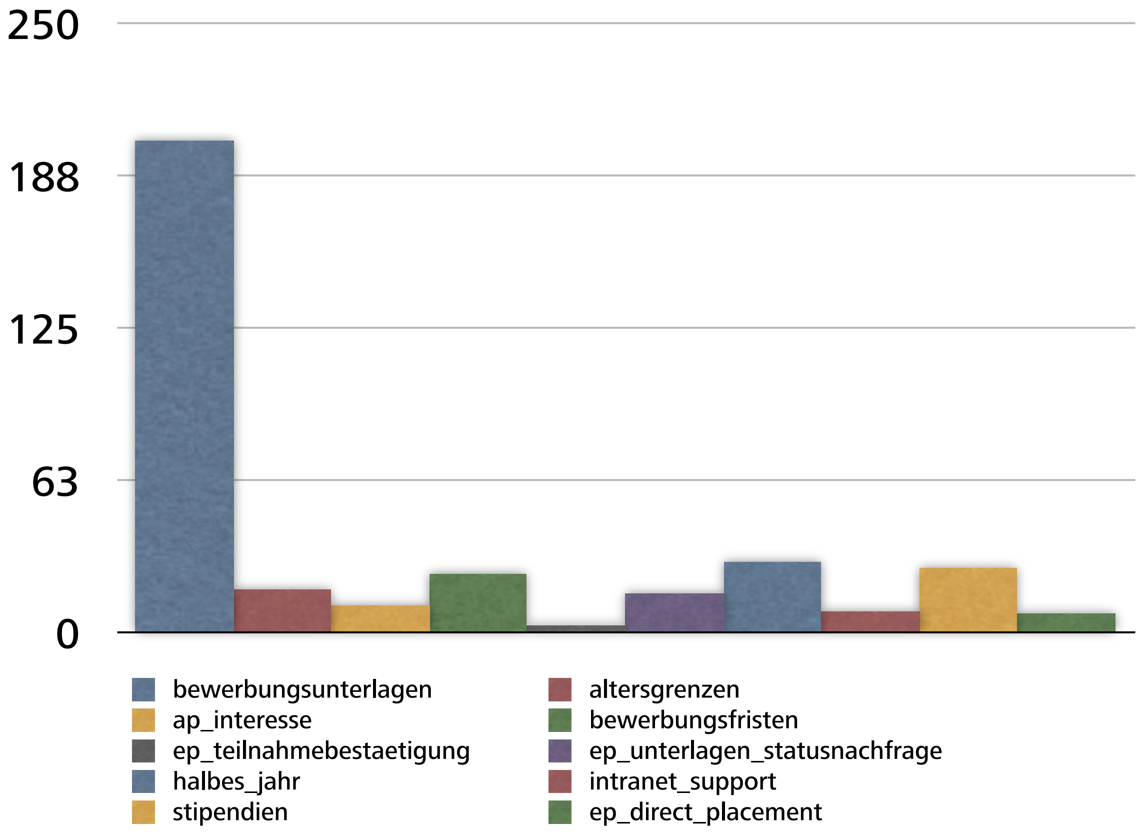
Manual classification of training data (n=350)

Initially: 26 categories, reduced to 10



Distribution of classes

Training data



Feature selection

1. Stopword removal

2. Initial selection of 500 most frequent terms

3. Supervised feature reduction to 16 features, using Weka

Both stemming and feature reduction improved accuracy.

Classification

Algorithms and results

Experimented with Naive Bayes, multinomial Naive Bayes and Support Vector Machine.

Naive Bayes yielded best results.

Precision for single classes ranged from 0% - 84%.

Overall correctly classified: ~77%

Classification accuracy

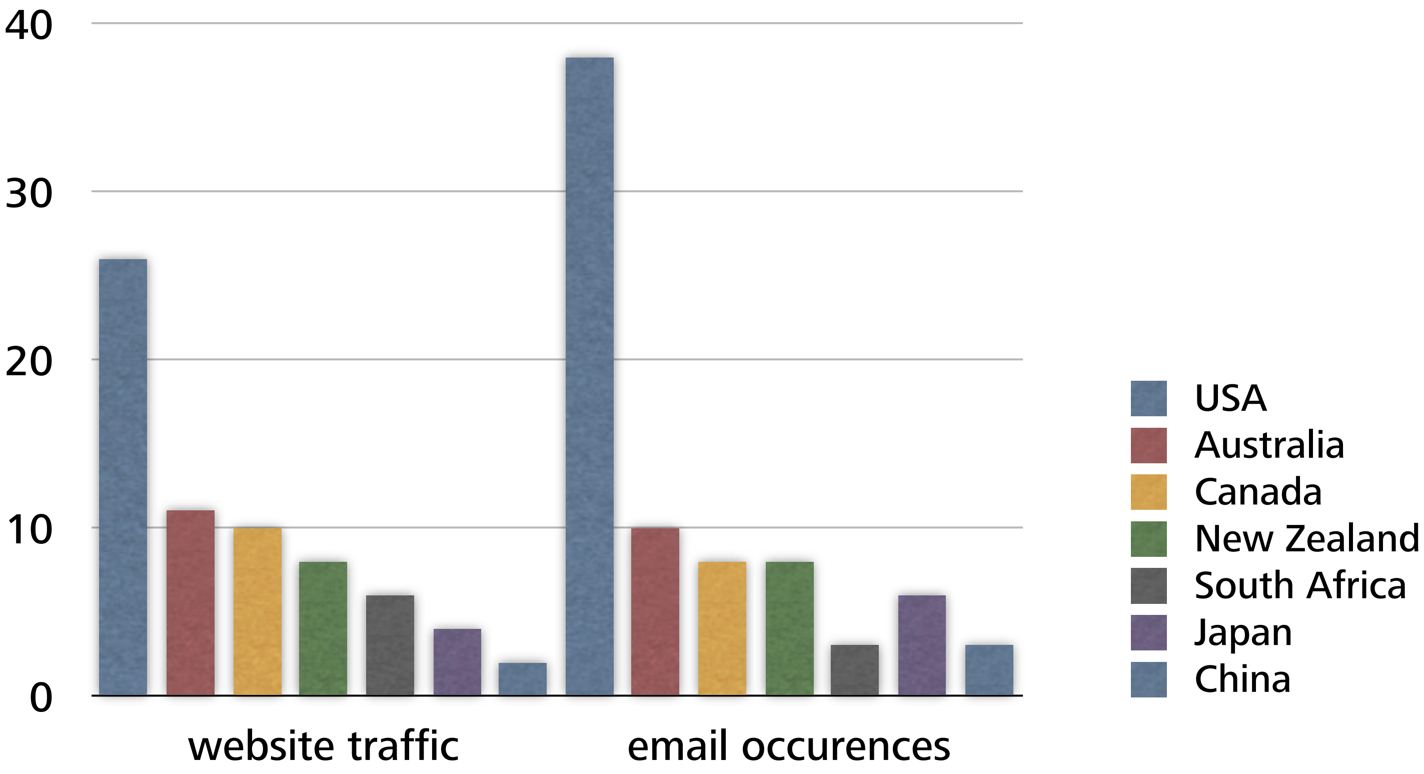
Confusion matrix

a	b	c	d	e	f	g	h	i	j	classified as
190	1	3	0	0	0	4	0	6	0	_BEWERBUNGSUNTERLAGEN
6	6	0	2	0	1	1	0	2	0	_altersgrenzen
5	0	4	0	0	0	0	0	0	2	_ap_interesse
6	0	0	15	0	1	1	0	1	0	_bewerbungsfristen
2	0	0	1	0	0	0	0	0	0	_ep_teilnahmebestaetigung
3	0	0	2	0	10	0	0	1	0	_ep_unterlagen_statusnachfrage
1	0	1	1	0	0	25	1	0	0	_halbesjahr
2	0	0	0	0	0	1	6	0	0	_intranet-support
10	4	0	1	0	2	3	0	8	0	_stipendien
0	0	2	0	0	0	1	0	1	4	_ep_direct_placement

Kappa = .61

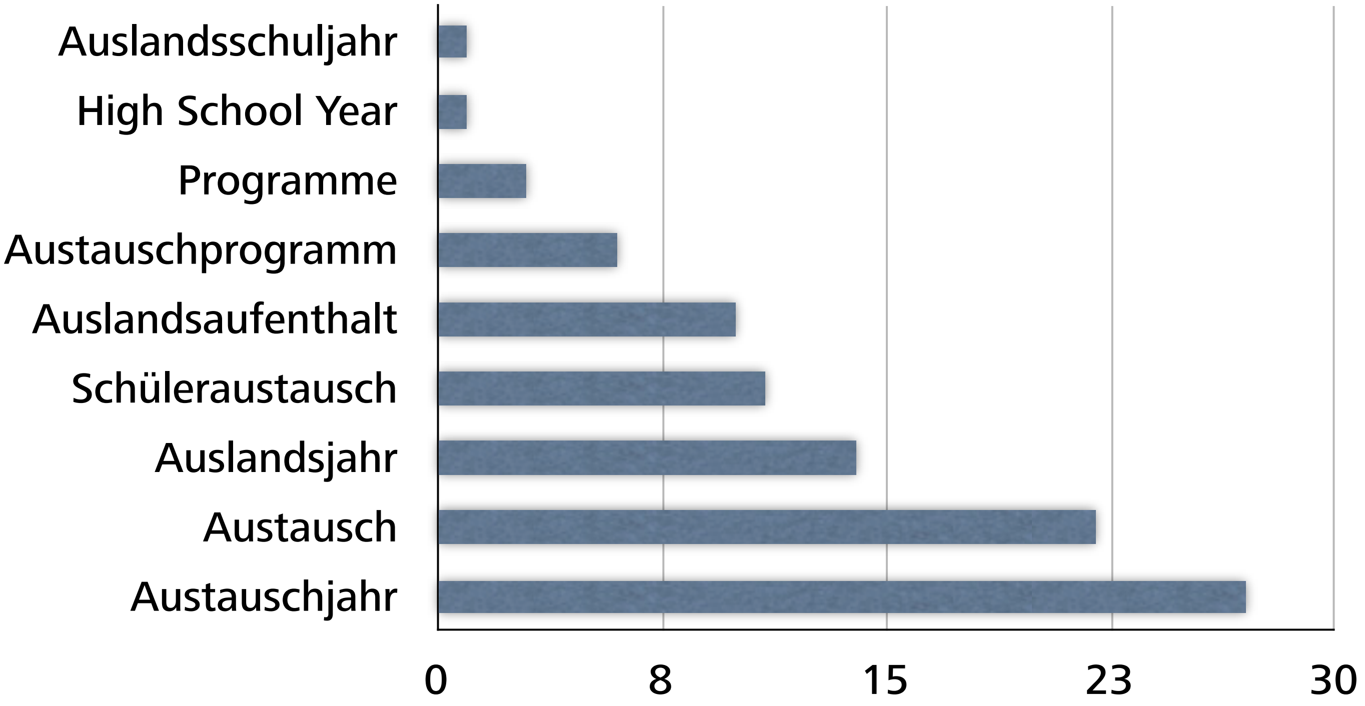
Further analysis

What countries are people interested in?



Further analysis

Wording: What do people call the product?



Further analysis

Recommendations for site usability refinement

People don’t find order form for application material

Perhaps site does not communicate details about programs clearly enough

Results and Outlook

Some emails are easy to handle, others need human attention

For from reliable auto-reply generation

Larger training set needed

Examine correlation between incoming and actual replies

Implement some email routing