Multilingual Search - How to build & improve

Paresh Paradkar

About me

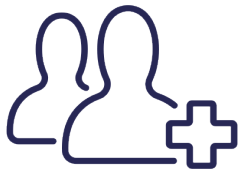
- Senior Software Engineer at Mimecast Services Ltd, London, UK.
- Apache Lucene and Elasticsearch enthusiast
- M.Sc. Applied Computer Sci, Universität Freiburg, Germany
- https://www.linkedin.com/in/paradkarparesh/

Mimecast at a Glance

mimecast®

2003
Founded

LONDON, UK - NASDAQ:MIME



32,200+
Customers

110%+
Retention Rate

NPS
Cxi Score

12
Data Centers

6 GLOBAL LOCATIONS

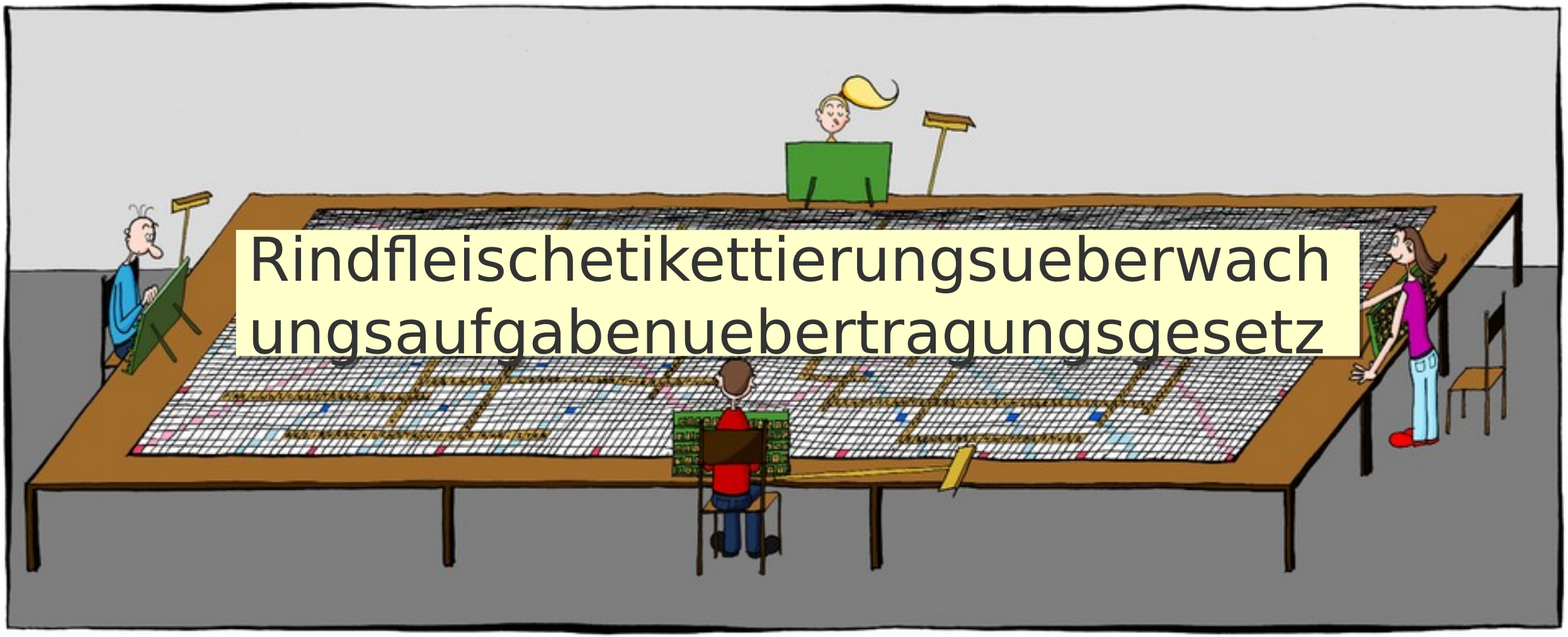
287B+
Messages

UNDER MANAGEMENT

465M+
Emails

PROCESSED EACH DAY

3M+
Searches per week



img: redit.com

Compound words from real users...

- Leasingvertragsunterlage
- Fassadenwerbeanlage
- Veranstaltungsbilder
- Passwortänderung
- Genehmigungsanfrage
- Versandbestätigung
- Modernisierungsstaatsvertrag

- Terminbestätigung
- Systemhauskongress
- Druckersystemhaus
- Ausweisregistrierung
- Registrierungsbestätigungen
- Bewertungsaufforderungen
- Kündigungsschutzprozess

Umlauts & Eszett

- Support for umlauts equivalents and eszett characters
- Ä ä = ae
- Ö ö = oe
- Ü ü = ue
- ß = ss

Custom decompounding analyzer

Custom decompounding analyzer

Custom decompounding analyzer


Thanks I can find sub-words now but can I search my both emails written in German and English together now?

Language Detection Model

- Logistic Regression model based on the Wikipedia dataset
- Detects languages with high precision
- High recall for English language

Test 4		Predicted					
		de	en	es	fr	it	pt
Real	de	29799	64	6	35	15	12
	en	51	29880	5	27	19	5
	es	11	49	29739	48	27	60
	fr	32	57	21	29879	24	7
	it	18	32	6	18	29805	9
	pt	7	15	24	18	27	29996

Separate Index per language

Separate field per language

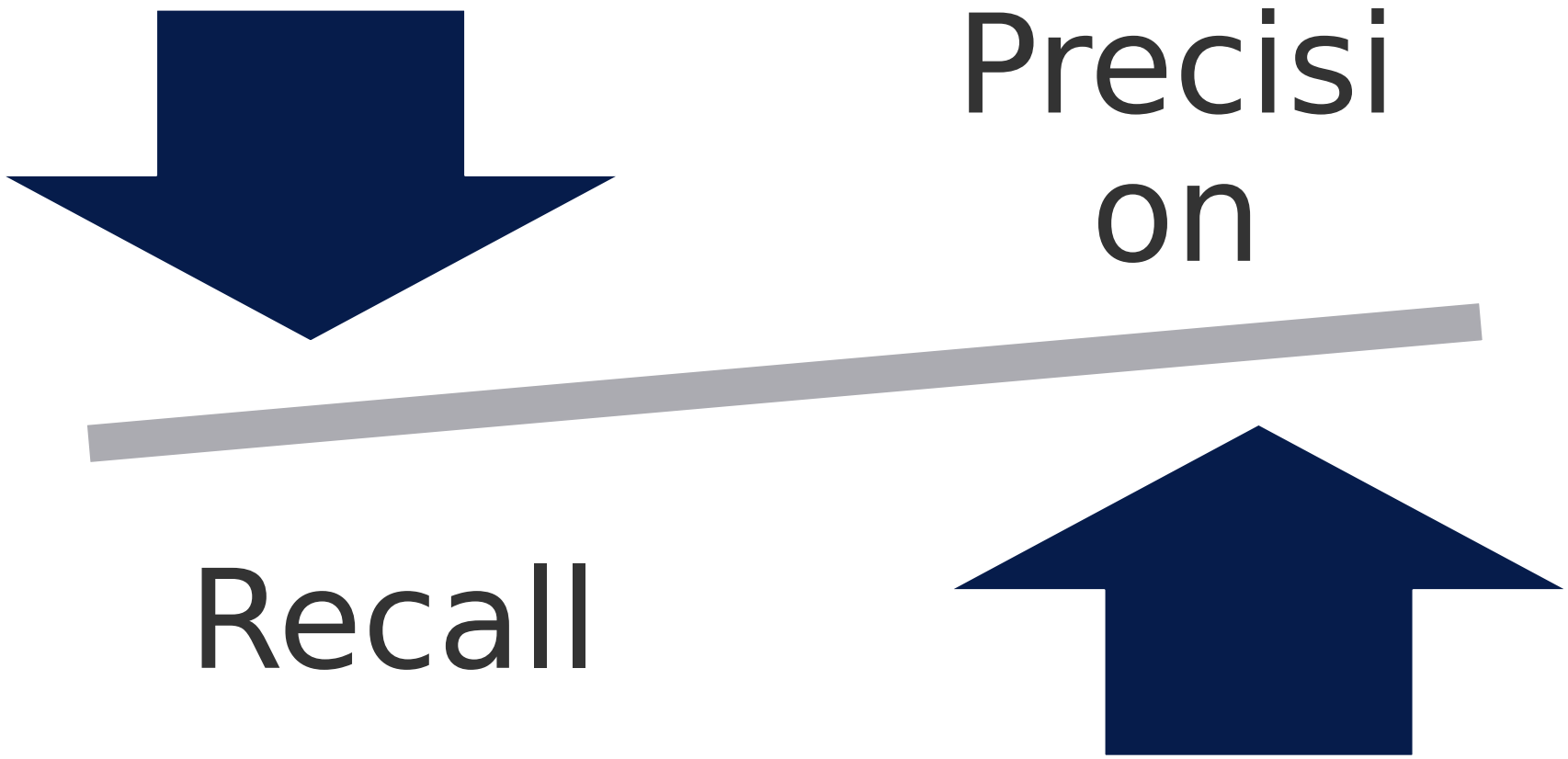
Separate document per language

Separate Index per language

Separate field per language

Separate document per language


Cool!! I can find my German & English emails as well.
But it returns email containing word “bestätigung” as top result when I



Query log analysis

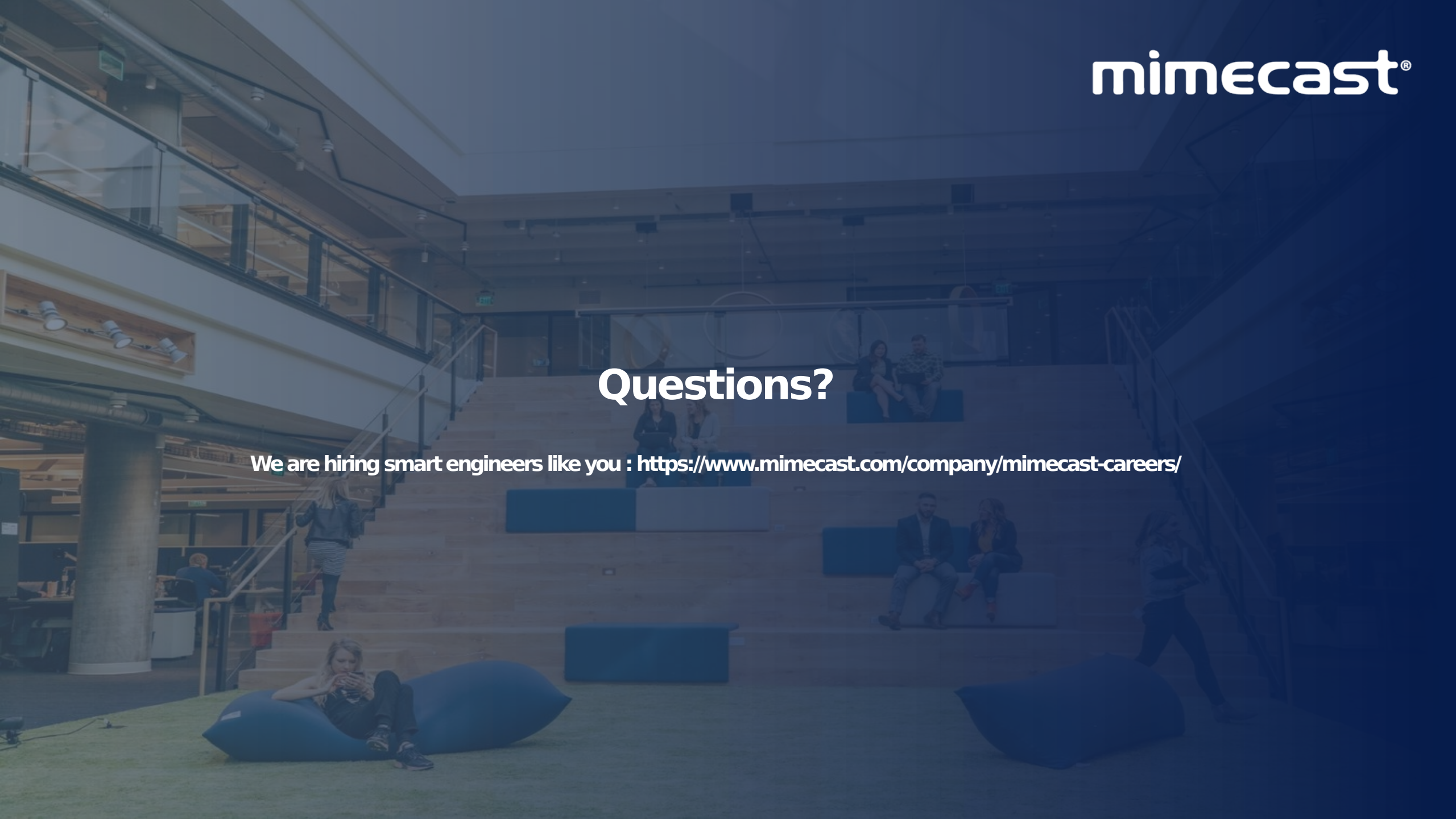
- Analyze query logs to understand what people are searching
- Feedback to the analyzer to fine tune the analyzer
- Adjust the minimum size of the sub-words while decompounding

Query Rewriting

- Store sub-words in a separate field
- Rewrite queries by boosting original field over the sub-words field.
- content: Bestätigung => (content:+Best^5 or subwords:Best)
- Docs containing TerminBestätigung will be ranked higher than docs with only sub-words like Bestätigung

Summary

- We built a language detection model to identify the language of the document
- We decided the structure of the emails – separate index for every language
- We wrote our custom analyzer to decompound words
- We used query rewriting to rank the documents with higher relevancy to increase our precision.


mimecast®
Questions?
We are hiring smart engineers like you : https://www.mimecast.com/company/mimecast-careers/


mimecast®
Thank you