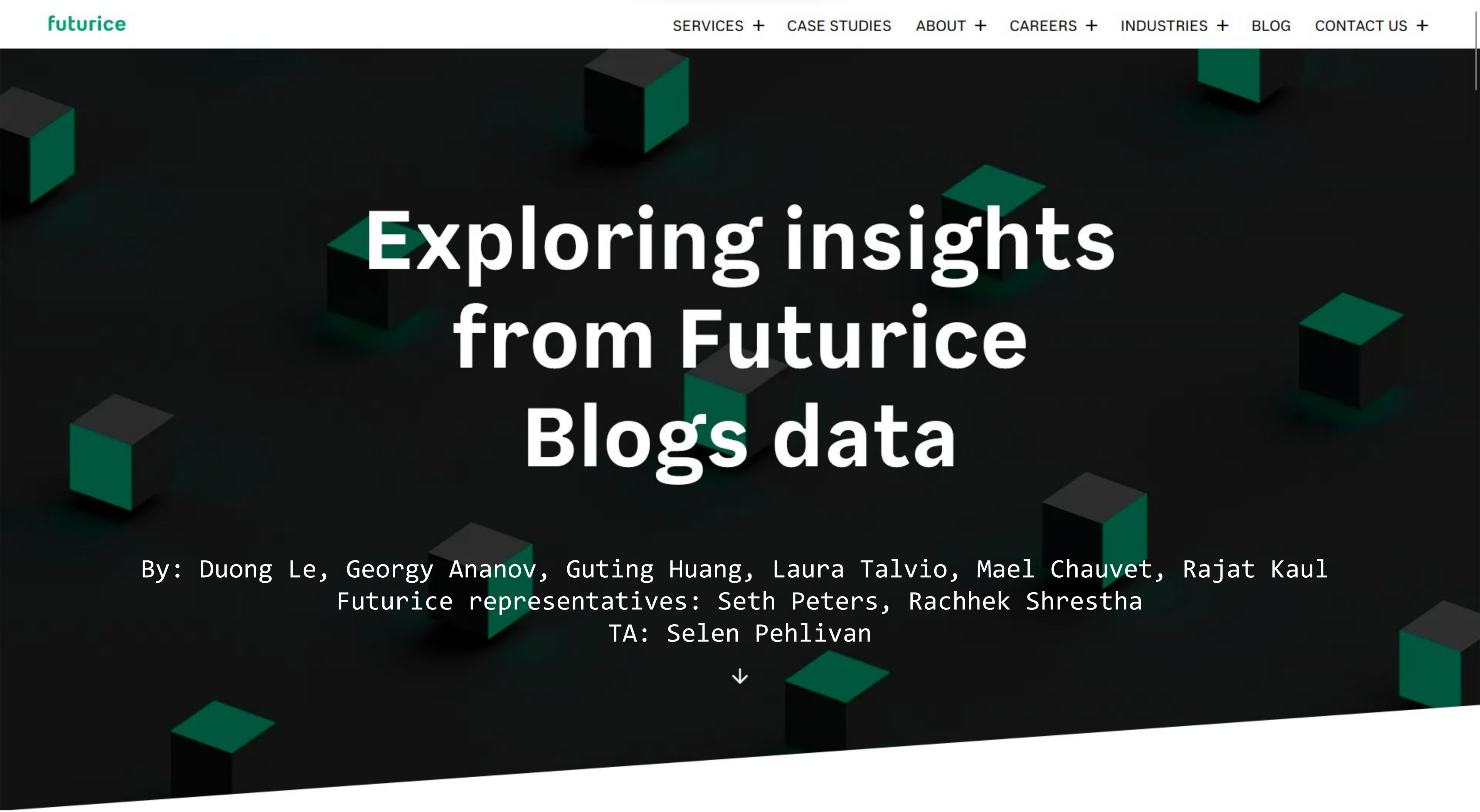


Exploring insights from Futurice Blogs data

By: Duong Le, Georgy Ananov, Guting Huang, Laura Talvio, Mael Chauvet, Rajat Kaul
Futurice representatives: Seth Peters, Rachhek Shrestha
TA: Selen Pehlivan

Overview

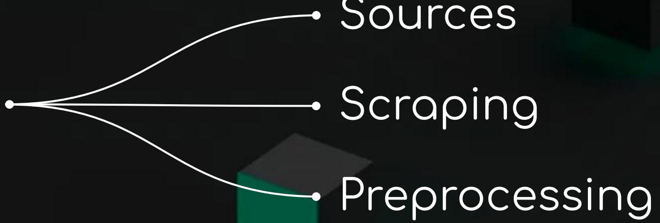
1. Problem definition
    - Current state of blog writing
    - Research questions
2. Data
3. Features
4. Outputs
5. Future plans
6. Questions

Overview

1. Problem definition
2. Data
3. Features
4. Outputs
5. Future plans
6. Questions

Overview

1. Problem definition
2. Data
3. Features
4. Outputs
5. Future plans
6. Questions



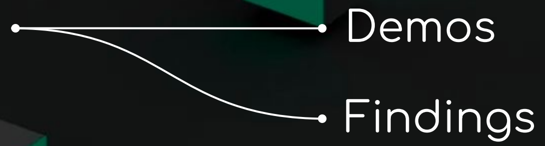
- Sources
- Scraping
- Preprocessing

Overview

1. Problem definition
2. Data
3. Features
   - Choice of features
   - Extraction methods
4. Outputs
5. Future plans
6. Questions

Overview

1. Problem definition
2. Data
3. Features
4. Outputs
5. Future plans
6. Questions



Demos

Findings

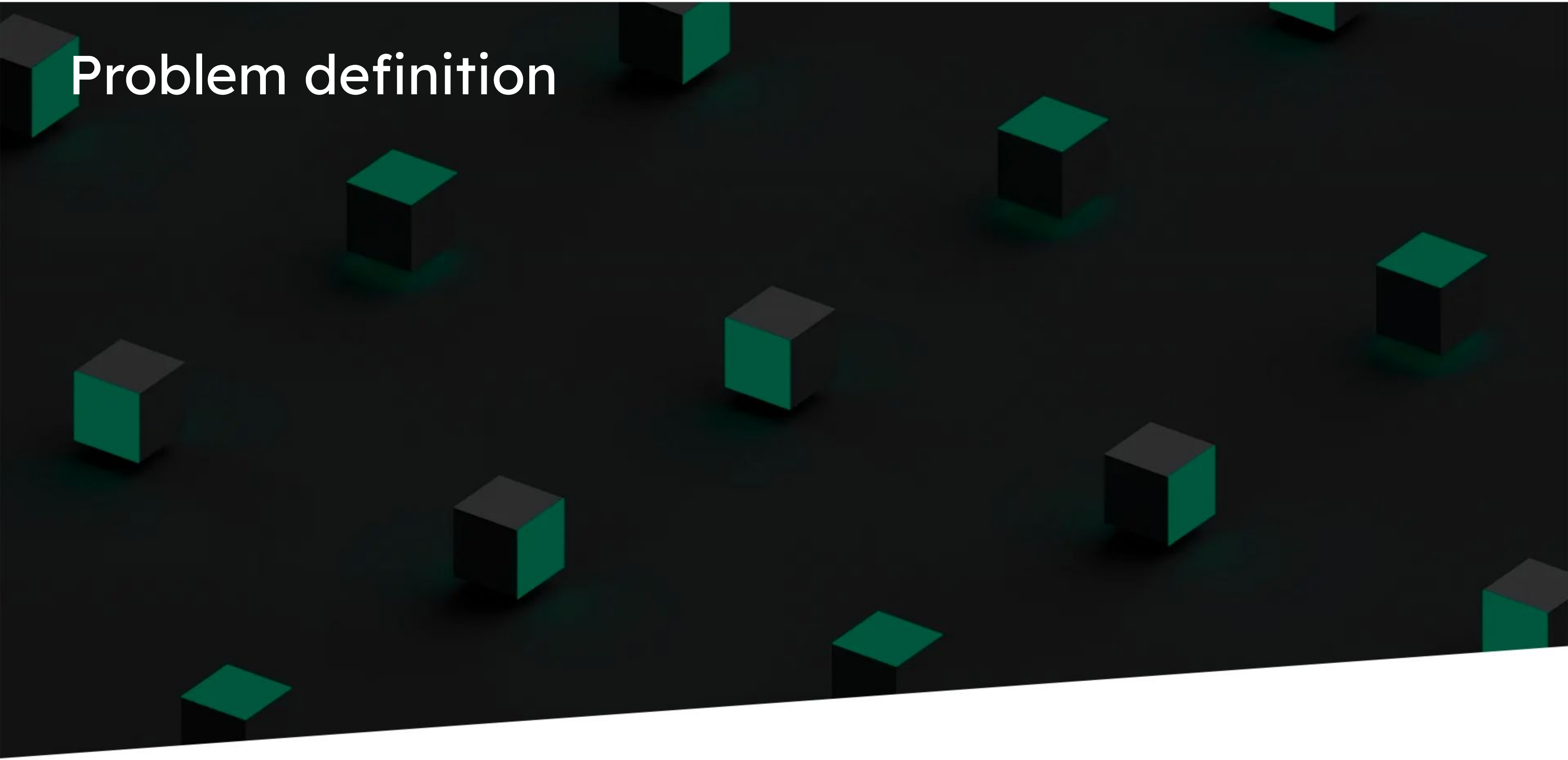
Overview

1. Problem definition
2. Data
3. Features
4. Outputs
5. Future plans
   - For the upcoming week
   - Further opportunities
6. Questions

Overview

1. Problem definition
2. Data
3. Features
4. Outputs
5. Future plans
6. Questions

futurice



Problem definition

Problem definition

Current state

- Blog writing decisions are based on external SEO analysis and interviews with the sales team
- Information gathered about existing blogs is difficult to leverage

Problem definition

Current state

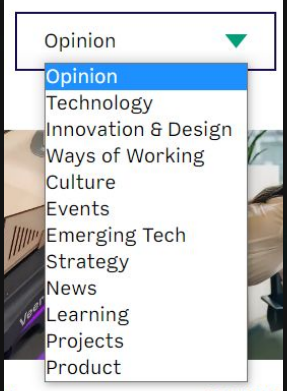
- Blog writing decisions are based on external SEO analysis and interviews with the sales team
- Information gathered about existing blogs is difficult to leverage

Our goals

- Create a tool that helps writers to understand the structure of existing blogs at a glance
- Find out which aspect of the blog have an impact on the blog's success
- Make a simple recommendation system on how the blog can be opimised

Research questions

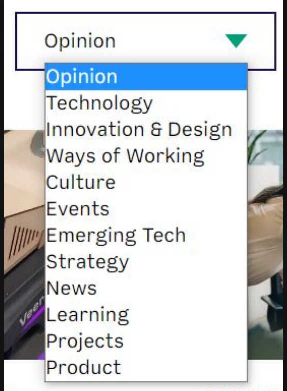
- Does the current topics represent the structure of the blogs?
  - Currently, there are over 800 blogs with 12 different topics
  - Some topics are similar, while some other are ambiguous
  - Is there a better way to categorize blogs?



Research questions

- Does the current topics represent the structure of the blogs?
    - Currently, there are over 800 blogs with 12 different topics
    - Some topics are similar, while some other are ambiguous
    - Is there a better way to categorize blogs?

- What are the most popular words over a period of time?
    - A question that was asked by the researchers at Futurice
    - It can help see how the writers react to global events or trends



Research questions

- What features of a blog affect its success?
  - Which features can we extract from the blog text?
  - Which ones are the most important?

Research questions

- What features of a blog affect its success?
    - Which features can we extract from the blog text?
    - Which ones are the most important?
    - How can these findings help writers evaluate and imrove their work?

Research questions

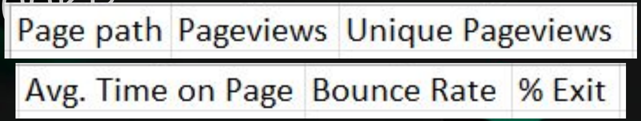
- What features of a blog affect its success?
  - Which features can we extract from the blog text?
  - Which ones are the most important?
  - How can these findings help writers evaluate and imrove their work?

- How to leverage data when making small blog-writing decisions?
  - In which situations can past data help writers in a meaningful way?
  - How can we deliver the data to the writers?

Data Sources

- Futurice website, around 800 blog post, 5 000KB
- Google Analytics data, 1 000KB
- Google Trends data, pytrends

Google Analytics



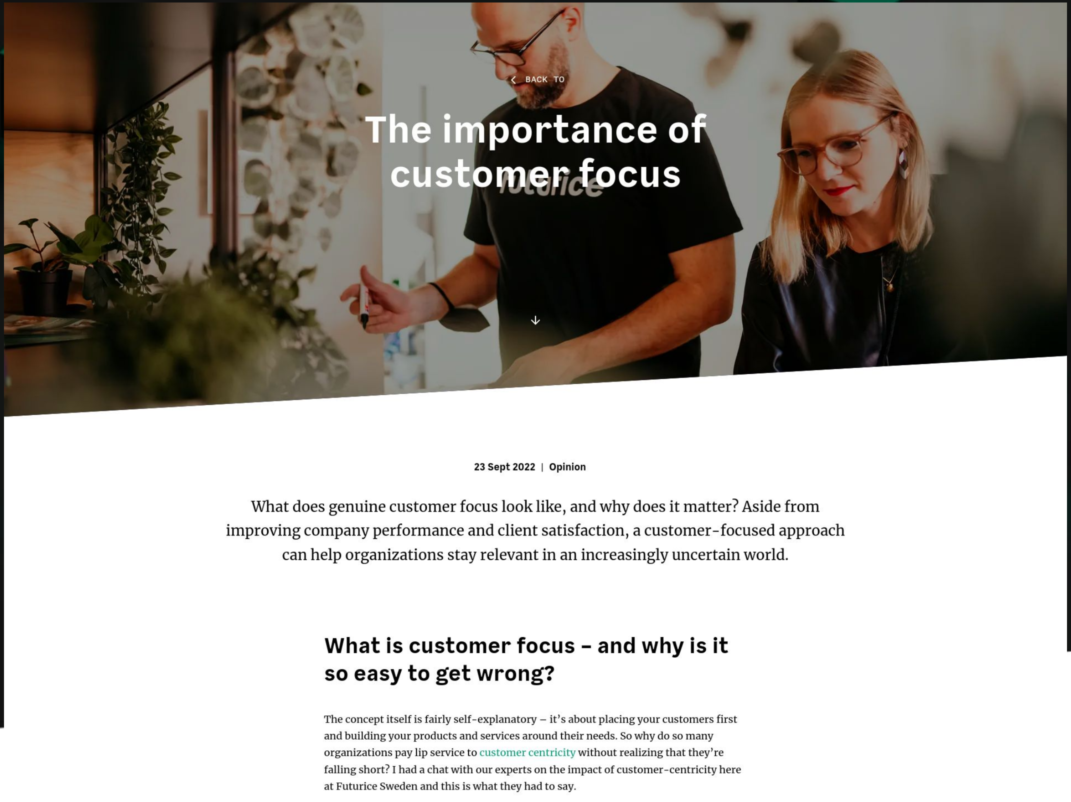
Page path	Pageviews	Unique Pageviews

Avg. Time on Page	Bounce Rate	% Exit

futurice
HTML
Google Analytics
CSV
{...}
JSON

futurice

Structure of the Blogs



‹ BACK TO

The importance of customer focus

23 Sept 2022 | Opinion

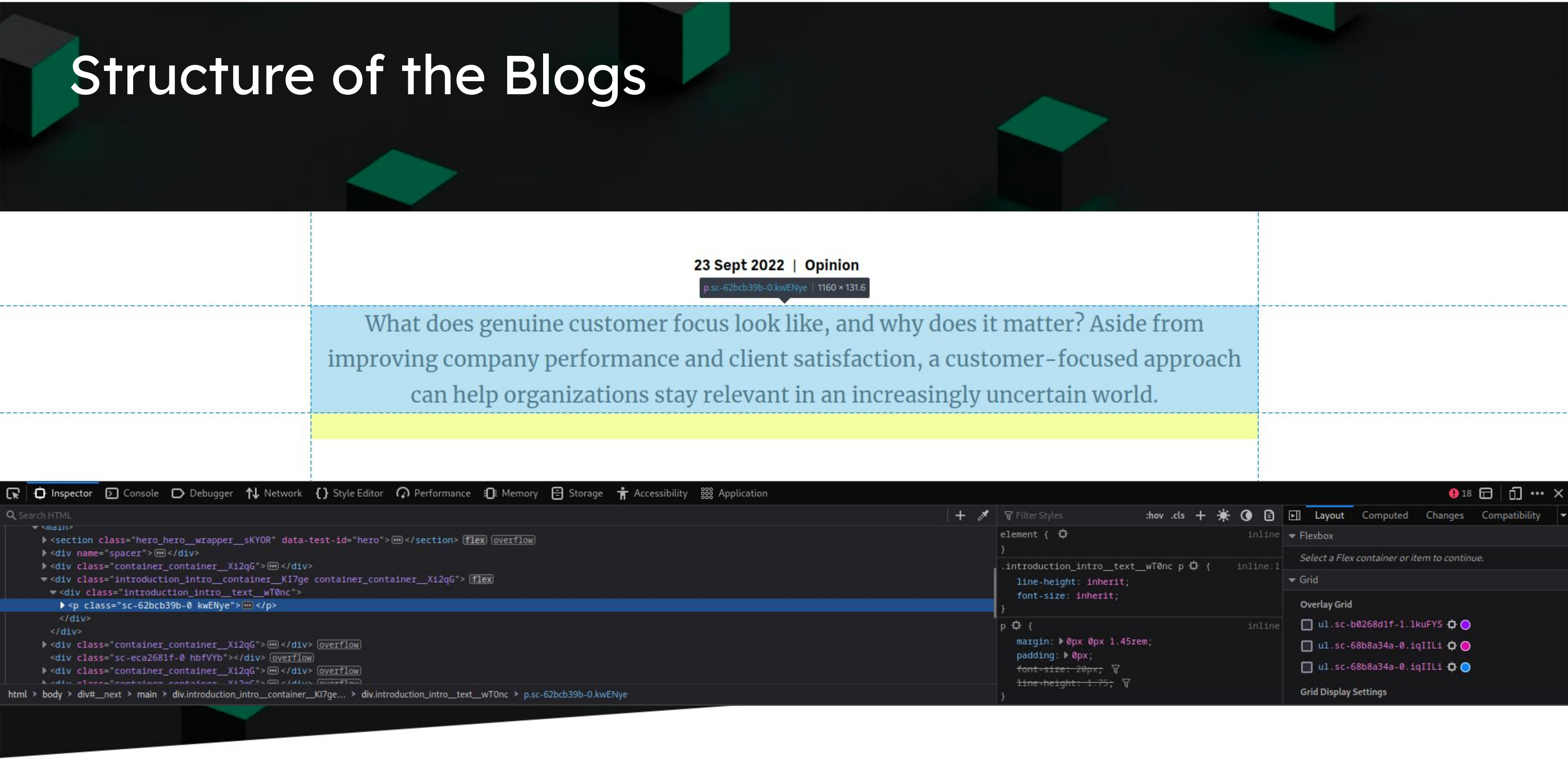
What does genuine customer focus look like, and why does it matter? Aside from improving company performance and client satisfaction, a customer-focused approach can help organizations stay relevant in an increasingly uncertain world.

What is customer focus – and why is it so easy to get wrong?

The concept itself is fairly self-explanatory – it's about placing your customers first and building your products and services around their needs. So why do so many organizations pay lip service to customer centricity without realizing that they're falling short? I had a chat with our experts on the impact of customer-centricity here at Futurice Sweden and this is what they had to say.

futurice

Structure of the Blogs



23 Sept 2022 | Opinion

What does genuine customer focus look like, and why does it matter? Aside from improving company performance and client satisfaction, a customer-focused approach can help organizations stay relevant in an increasingly uncertain world.

```
▾<main>
  ▸<section class="hero_hero__wrapper__sKYOR" data-test-id="hero">⋯</section> flex overflow
  ▸<div name="spacer">⋯</div>
  ▸<div class="container_container__Xi2qG">⋯</div>
  ▾<div class="introduction_intro__container__KI7ge container_container__Xi2qG"> flex
    ▾<div class="introduction_intro__text__wT0nc">
      ▸<p class="sc-62bcb39b-0 kwENye">⋯</p>
     </div>
   </div>
  ▸<div class="container_container__Xi2qG">⋯</div> overflow
   <div class="sc-eca2681f-0 hbfVYb"></div> overflow
  ▸<div class="container_container__Xi2qG">⋯</div> overflow
```

html › body › div#__next › main › div.introduction_intro__container__KI7ge... › div.introduction_intro__text__wT0nc › p.sc-62bcb39b-0.kwENye

```
element {                                        inline
}
.introduction_intro__text__wT0nc p {             inline:1
  line-height: inherit;
  font-size: inherit;
}
p {                                              inline
  margin: ▸0px 0px 1.45rem;
  padding: ▸0px;
  ~~font-size: 20px;~~
  ~~line-height: 1.75;~~
}
```

▾ Flexbox

Select a Flex container or item to continue.

▾ Grid

Overlay Grid

□ ul.sc-b0268d1f-1.lkuFYS

□ ul.sc-68b8a34a-0.iqIILi

□ ul.sc-68b8a34a-0.iqIILi

Grid Display Settings

Scraping

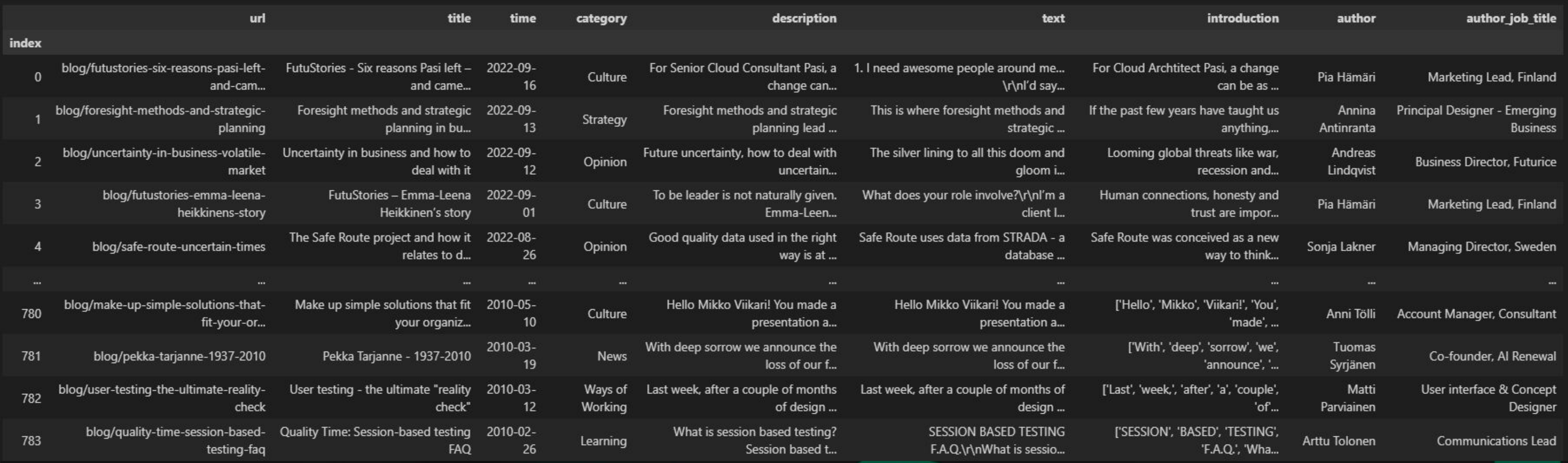
Requests and **BeautifulSoup** were used to get the links from the Futurice blog parent page.

- Initially, we used the above libraries for downloading the webpage as well.
- But we ran into an issue where the webpage being downloaded didn’t have any Javascript (JS) loaded, because **Requests** downloads the HTML of the website without waiting for all the JS to load.
- This caused us to download incomplete websites.

So we switched to **Selenium** for scraping purposes, and it yielded many more benefits.

In addition, we used **pandas** for data storage and processing, and **time** for preventing DDOS protections from kicking in.

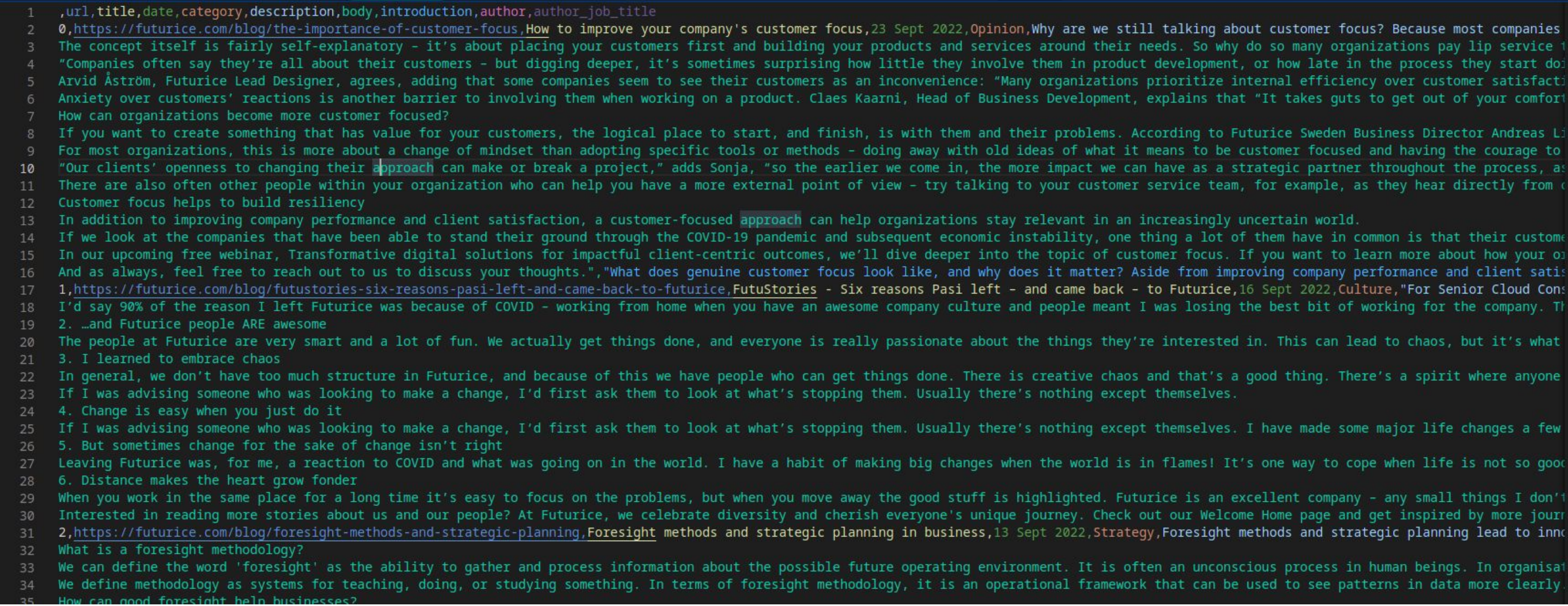
DataFrame



	url	title	time	category	description	text	introduction	author	author_job_title
index									
0	blog/futustories-six-reasons-pasi-left-and-cam...	FutuStories - Six reasons Pasi left – and came...	2022-09-16	Culture	For Senior Cloud Consultant Pasi, a change can...	1. I need awesome people around me... \r\nI'd say...	For Cloud Archtitect Pasi, a change can be as ...	Pia Hämäri	Marketing Lead, Finland
1	blog/foresight-methods-and-strategic-planning	Foresight methods and strategic planning in bu...	2022-09-13	Strategy	Foresight methods and strategic planning lead ...	This is where foresight methods and strategic ...	If the past few years have taught us anything,...	Annina Antinranta	Principal Designer - Emerging Business
2	blog/uncertainty-in-business-volatile-market	Uncertainty in business and how to deal with it	2022-09-12	Opinion	Future uncertainty, how to deal with uncertain...	The silver lining to all this doom and gloom i...	Looming global threats like war, recession and...	Andreas Lindqvist	Business Director, Futurice
3	blog/futustories-emma-leena-heikkinens-story	FutuStories – Emma-Leena Heikkinen's story	2022-09-01	Culture	To be leader is not naturally given. Emma-Leen...	What does your role involve?\r\nI'm a client I...	Human connections, honesty and trust are impor...	Pia Hämäri	Marketing Lead, Finland
4	blog/safe-route-uncertain-times	The Safe Route project and how it relates to d...	2022-08-26	Opinion	Good quality data used in the right way is at ...	Safe Route uses data from STRADA - a database ...	Safe Route was conceived as a new way to think...	Sonja Lakner	Managing Director, Sweden
...	...	...	...	...	...	...	...	...	...
780	blog/make-up-simple-solutions-that-fit-your-or...	Make up simple solutions that fit your organiz...	2010-05-10	Culture	Hello Mikko Viikari! You made a presentation a...	Hello Mikko Viikari! You made a presentation a...	['Hello', 'Mikko', 'Viikari!', 'You', 'made', ...	Anni Tölli	Account Manager, Consultant
781	blog/pekka-tarjanne-1937-2010	Pekka Tarjanne - 1937-2010	2010-03-19	News	With deep sorrow we announce the loss of our f...	With deep sorrow we announce the loss of our f...	['With', 'deep', 'sorrow', 'we', 'announce', '...	Tuomas Syrjänen	Co-founder, AI Renewal
782	blog/user-testing-the-ultimate-reality-check	User testing - the ultimate "reality check"	2010-03-12	Ways of Working	Last week, after a couple of months of design ...	Last week, after a couple of months of design ...	['Last', 'week,', 'after', 'a', 'couple', 'of'...	Matti Parviainen	User interface & Concept Designer
783	blog/quality-time-session-based-testing-faq	Quality Time: Session-based testing FAQ	2010-02-26	Learning	What is session based testing? Session based t...	SESSION BASED TESTING F.A.Q.\r\nWhat is sessio...	['SESSION', 'BASED', 'TESTING', 'F.A.Q.', 'Wha...	Arttu Tolonen	Communications Lead

futurice

CSV



```
,url,title,date,category,description,body,introduction,author,author_job_title
0,https://futurice.com/blog/the-importance-of-customer-focus,How to improve your company's customer focus,23 Sept 2022,Opinion,Why are we still talking about customer focus? Because most companies
The concept itself is fairly self-explanatory – it’s about placing your customers first and building your products and services around their needs. So why do so many organizations pay lip service t
“Companies often say they’re all about their customers – but digging deeper, it’s sometimes surprising how little they involve them in product development, or how late in the process they start doi
Arvid Åström, Futurice Lead Designer, agrees, adding that some companies seem to see their customers as an inconvenience: “Many organizations prioritize internal efficiency over customer satisfacti
Anxiety over customers’ reactions is another barrier to involving them when working on a product. Claes Kaarni, Head of Business Development, explains that “It takes guts to get out of your comfort
How can organizations become more customer focused?
If you want to create something that has value for your customers, the logical place to start, and finish, is with them and their problems. According to Futurice Sweden Business Director Andreas Li
For most organizations, this is more about a change of mindset than adopting specific tools or methods – doing away with old ideas of what it means to be customer focused and having the courage to
“Our clients’ openness to changing their approach can make or break a project,” adds Sonja, “so the earlier we come in, the more impact we can have as a strategic partner throughout the process, as
There are also often other people within your organization who can help you have a more external point of view – try talking to your customer service team, for example, as they hear directly from c
Customer focus helps to build resiliency
In addition to improving company performance and client satisfaction, a customer-focused approach can help organizations stay relevant in an increasingly uncertain world.
If we look at the companies that have been able to stand their ground through the COVID-19 pandemic and subsequent economic instability, one thing a lot of them have in common is that their custome
In our upcoming free webinar, Transformative digital solutions for impactful client-centric outcomes, we’ll dive deeper into the topic of customer focus. If you want to learn more about how your or
And as always, feel free to reach out to us to discuss your thoughts.","What does genuine customer focus look like, and why does it matter? Aside from improving company performance and client satis
1,https://futurice.com/blog/futustories-six-reasons-pasi-left-and-came-back-to-futurice,FutuStories - Six reasons Pasi left - and came back - to Futurice,16 Sept 2022,Culture,"For Senior Cloud Cons
I’d say 90% of the reason I left Futurice was because of COVID – working from home when you have an awesome company culture and people meant I was losing the best bit of working for the company. Th
2. …and Futurice people ARE awesome
The people at Futurice are very smart and a lot of fun. We actually get things done, and everyone is really passionate about the things they’re interested in. This can lead to chaos, but it’s what
3. I learned to embrace chaos
In general, we don’t have too much structure in Futurice, and because of this we have people who can get things done. There is creative chaos and that’s a good thing. There’s a spirit where anyone
If I was advising someone who was looking to make a change, I’d first ask them to look at what’s stopping them. Usually there’s nothing except themselves.
4. Change is easy when you just do it
If I was advising someone who was looking to make a change, I’d first ask them to look at what’s stopping them. Usually there’s nothing except themselves. I have made some major life changes a few
5. But sometimes change for the sake of change isn’t right
Leaving Futurice was, for me, a reaction to COVID and what was going on in the world. I have a habit of making big changes when the world is in flames! It’s one way to cope when life is not so good
6. Distance makes the heart grow fonder
When you work in the same place for a long time it’s easy to focus on the problems, but when you move away the good stuff is highlighted. Futurice is an excellent company – any small things I don’t
Interested in reading more stories about us and our people? At Futurice, we celebrate diversity and cherish everyone's unique journey. Check out our Welcome Home page and get inspired by more journ
2,https://futurice.com/blog/foresight-methods-and-strategic-planning,Foresight methods and strategic planning in business,13 Sept 2022,Strategy,Foresight methods and strategic planning lead to inno
What is a foresight methodology?
We can define the word 'foresight' as the ability to gather and process information about the possible future operating environment. It is often an unconscious process in human beings. In organisat
We define methodology as systems for teaching, doing, or studying something. In terms of foresight methodology, it is an operational framework that can be used to see patterns in data more clearly.
How can good foresight help businesses?
```

Data preprocessing

- Clean up html tags from the blog data

Data preprocessing

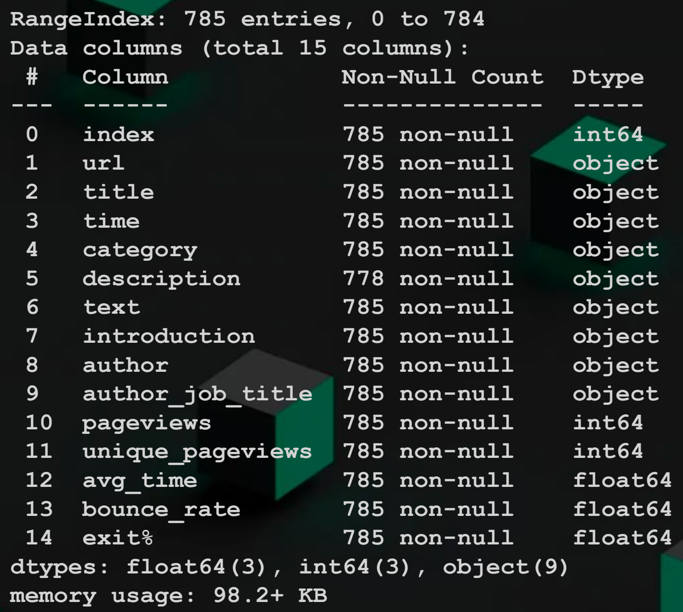
- Clean up html tags from the blog data
- Prepare Google analytics data
    - Consolidate data from several excel sheets
    - Aggregate entries that are related to the same blog article
    - Clean up and format the URLs to match the formatting of the scraped data

Data preprocessing

- Clean up html tags from the blog data
- Prepare Google analytics data
    - Consolidate data from several excel sheets
    - Aggregate entries that are related to the same blog article
    - Clean up and format the URLs to match the formatting of the scraped data
- Match the scraped blog entries against the analytics entries

Data preprocessing

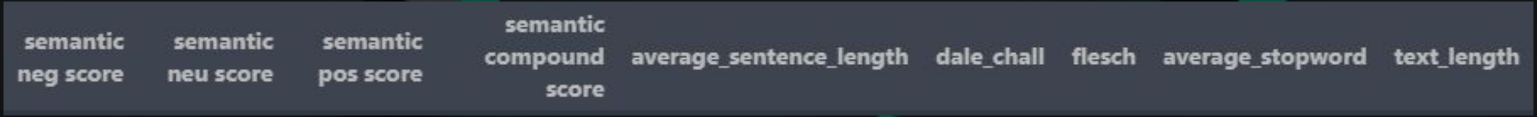
- Clean up html tags from the blog data
- Prepare Google analytics data
    - Consolidate data from several excel sheets
    - Aggregate entries that are related to the same blog article
    - Clean up and format the URLs to match the formatting of the scraped data
- Match the scraped blog entries against the analytics entries



```
RangeIndex: 785 entries, 0 to 784
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   index              785 non-null    int64
 1   url                785 non-null    object
 2   title              785 non-null    object
 3   time               785 non-null    object
 4   category           785 non-null    object
 5   description        778 non-null    object
 6   text               785 non-null    object
 7   introduction       785 non-null    object
 8   author             785 non-null    object
 9   author_job_title   785 non-null    object
 10  pageviews          785 non-null    int64
 11  unique_pageviews   785 non-null    int64
 12  avg_time           785 non-null    float64
 13  bounce_rate        785 non-null    float64
 14  exit%              785 non-null    float64
dtypes: float64(3), int64(3), object(9)
memory usage: 98.2+ KB
```

Feature extraction

- The following features were extracted from the blogs:
  - Semantic scores (Positive, Negative, Neutral, and Compound)
  - Readability scores (Dale-Chall, Flesch)
  - Text length and average sentence length
  - Average stopwords length (NLTK's stopwords package)
- Most of the features are extracted with the help of external libraries (e.g. textstat, nltk.sentiment)



semantic neg score	semantic neu score	semantic pos score	semantic compound score	average_sentence_length	dale_chall	flesch	average_stopword	text_length

Topic Modeling

- Are the current topics representative of the structure of the blogs?
- Provides a general summary of past blogs
- Helps future writers to learn about topics covered by existing blogs from Futurice

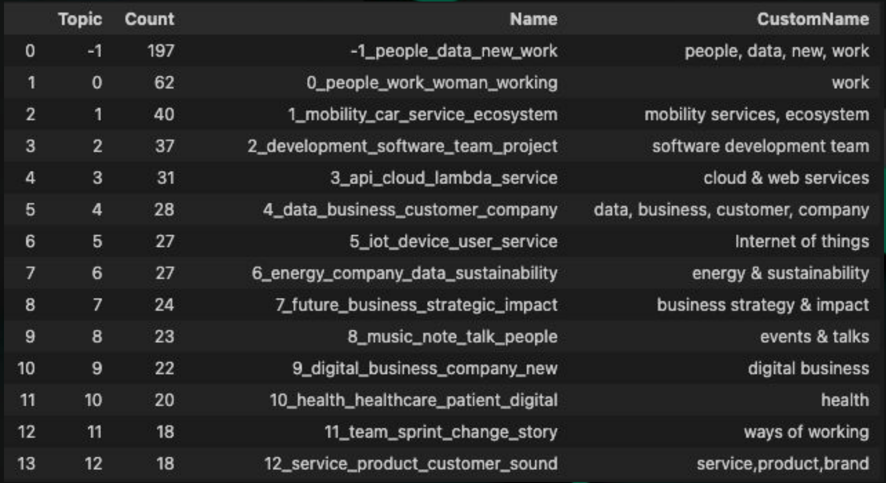
Topic Modeling (BERTopic)

- Outputs a list of topics and a list of documents associated with each other
- First some preprocessing: lowercase text, expand contractions, remove punctuations, numbers, stopwords, and lemmatize
- Steps of the BERTopic algorithm:
  - Embed Documents (SentenceTransformer)
  - Clustering embeddings (UMAP, HDBSCAN)
  - Bag-of-Words
  - Topic-Representation (c-TF-IDF)

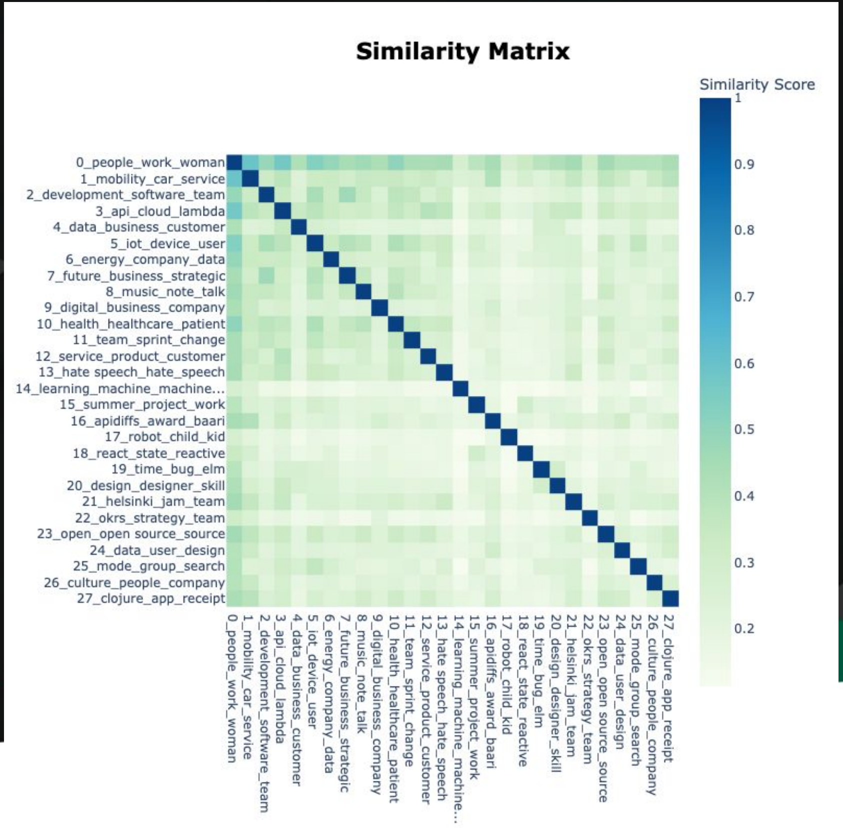
BERTopic

NLTK

Topic Modeling (BERTopic)



	Topic	Count	Name	CustomName
0	-1	197	-1_people_data_new_work	people, data, new, work
1	0	62	0_people_work_woman_working	work
2	1	40	1_mobility_car_service_ecosystem	mobility services, ecosystem
3	2	37	2_development_software_team_project	software development team
4	3	31	3_api_cloud_lambda_service	cloud & web services
5	4	28	4_data_business_customer_company	data, business, customer, company
6	5	27	5_iot_device_user_service	Internet of things
7	6	27	6_energy_company_data_sustainability	energy & sustainability
8	7	24	7_future_business_strategic_impact	business strategy & impact
9	8	23	8_music_note_talk_people	events & talks
10	9	22	9_digital_business_company_new	digital business
11	10	20	10_health_healthcare_patient_digital	health
12	11	18	11_team_sprint_change_story	ways of working
13	12	18	12_service_product_customer_sound	service,product,brand



Topic Modeling (BERTopic)

- Outputs a list of topics and a list of documents associated with each other
- First some preprocessing: lowercase text, expand contractions, remove punctuations, numbers, stopwords, and lemmatize
- Steps of the BERTopic algorithm:
  - Embed Documents (SentenceTransformer)
  - Clustering embeddings (UMAP, HDBSCAN)
  - Bag-of-Words
  - Topic-Representation (c-TF-IDF)
- Dynamic Topic Modeling
  - shows the evolution/trend of topics of the blogs over time

BERTopic

NLTK

Category Modeling using Clustering Models

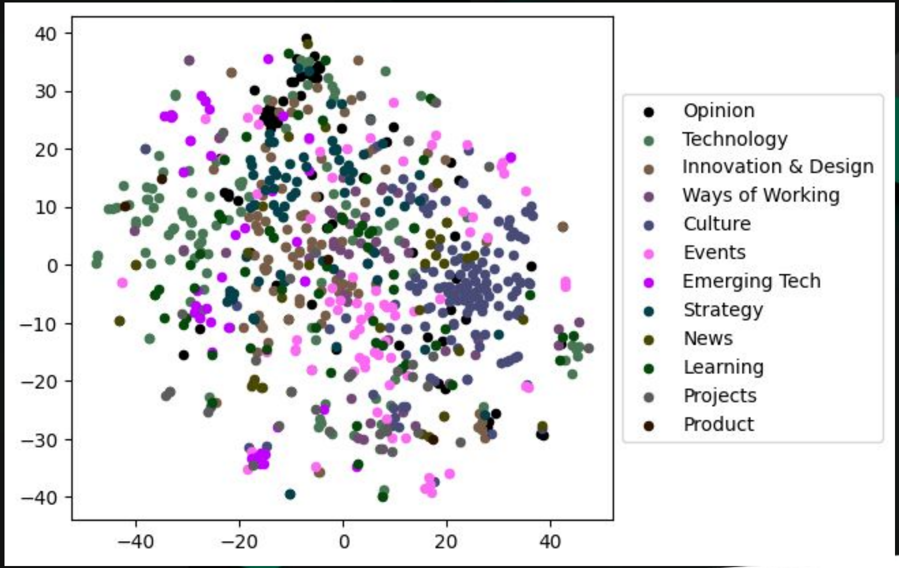
- Outputs clusters grouping similar documents together
- Steps of the Clustering process:
    - Run documents through TF-IDF Vectorizer
    - Cluster Documents using K-Means Clustering
    - Find relevant clusters
- Information we can get out of the clusters

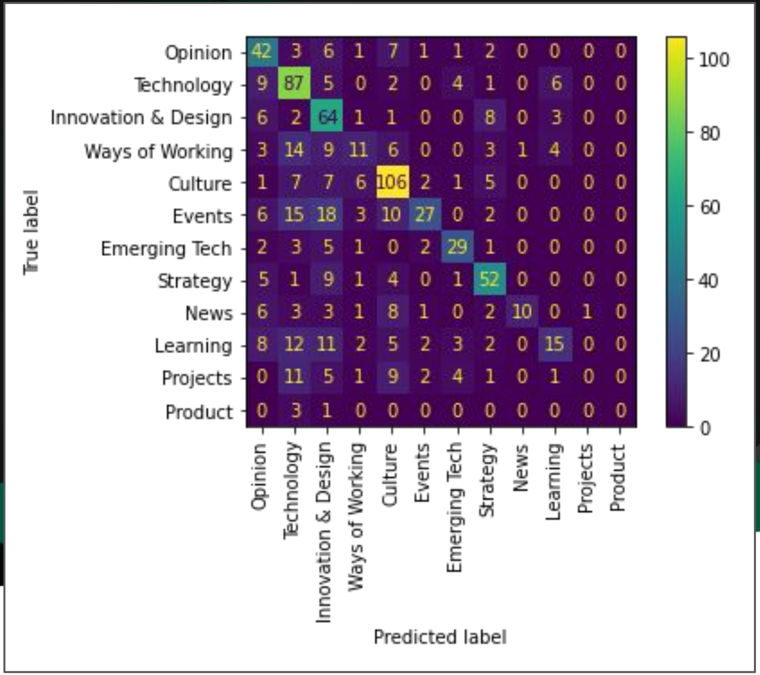
Category Modeling using Clustering Models

- Turning the documents into features using TF-IDF:
    - Preprocess the documents the same way as in BERTopic
    - Train a TfIdfVectorizer model using inverse document frequency and normalizing the vectors using an L2 norm.
    - Use TSNE to represent those features on a 2D plane and try to find noticeable patterns within the plot

Category Modeling using Clustering Models

- 2D representation of documents after using TSNE:



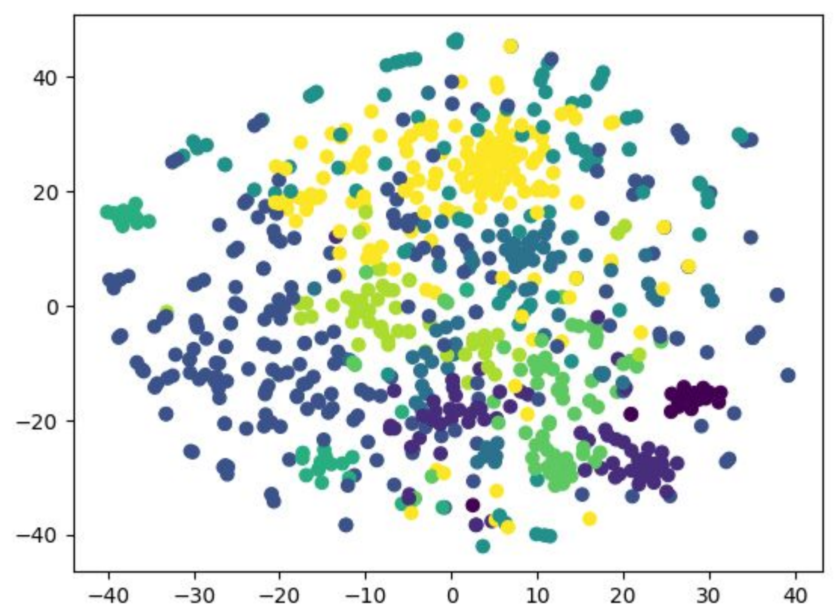


	Opinion	Technology	Innovation & Design	Ways of Working	Culture	Events	Emerging Tech	Strategy	News	Learning	Projects	Product
Opinion	42	3	6	1	7	1	1	2	0	0	0	0
Technology	9	87	5	0	2	0	4	1	0	6	0	0
Innovation & Design	6	2	64	1	1	0	0	8	0	3	0	0
Ways of Working	3	14	9	11	6	0	0	3	1	4	0	0
Culture	1	7	7	6	106	2	1	5	0	0	0	0
Events	6	15	18	3	10	27	0	2	0	0	0	0
Emerging Tech	2	3	5	1	0	2	29	1	0	0	0	0
Strategy	5	1	9	1	4	0	1	52	0	0	0	0
News	6	3	3	1	8	1	0	2	10	0	1	0
Learning	8	12	11	2	5	2	3	2	0	15	0	0
Projects	0	11	5	1	9	2	4	1	0	1	0	0
Product	0	3	1	0	0	0	0	0	0	0	0	0

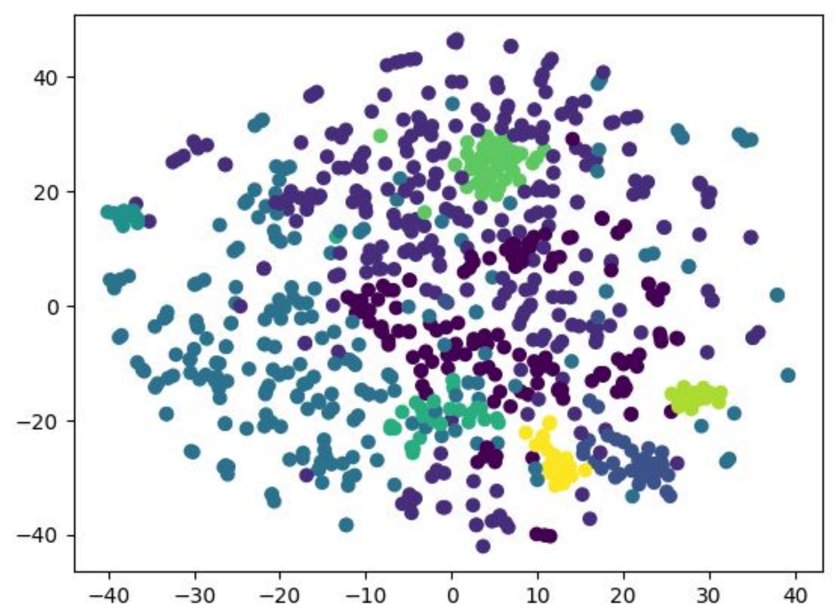
Category Modeling using Clustering Models

- Train Clustering models on those features

KMeans:



Agglomerative Clustering:

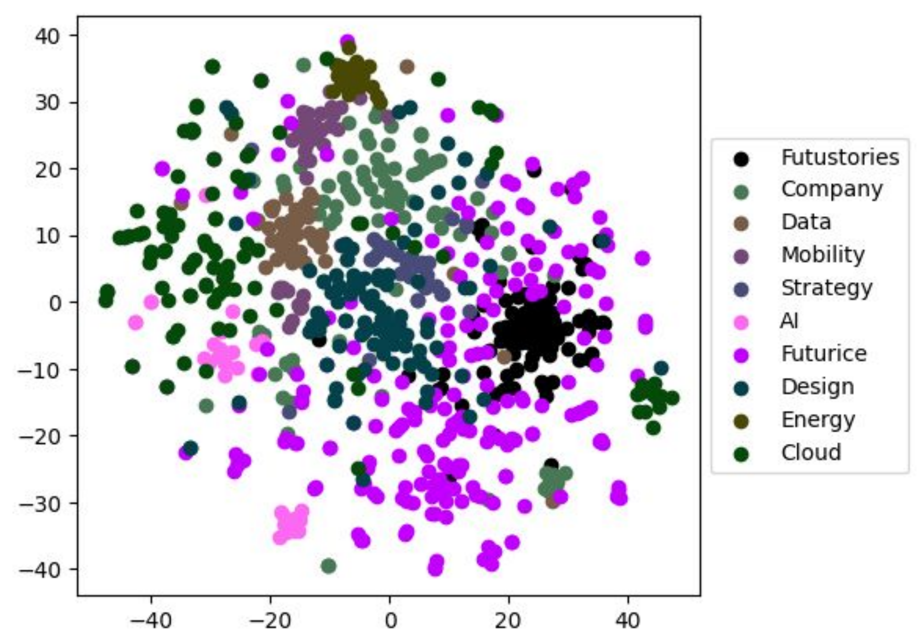


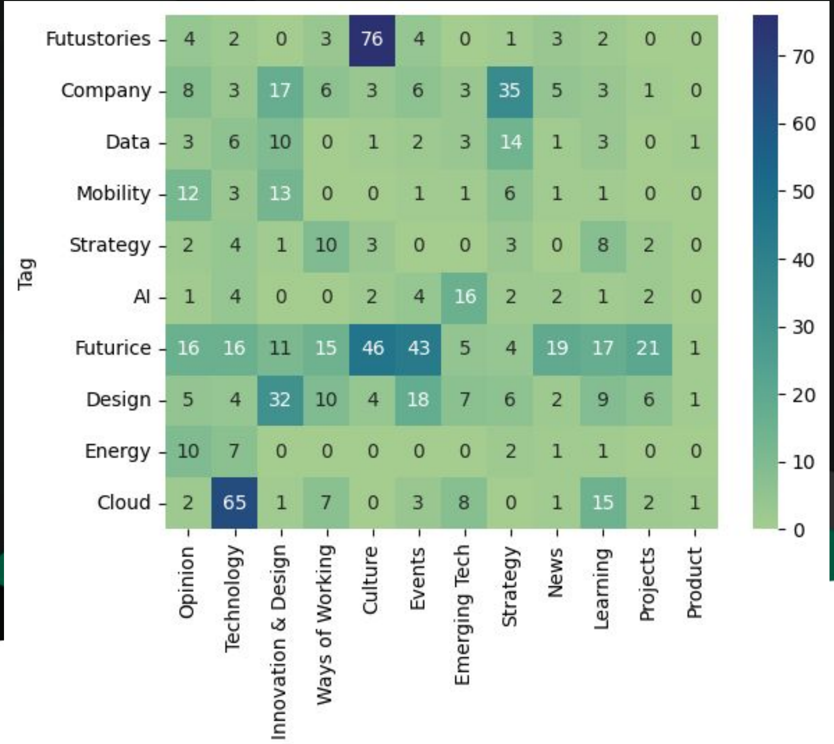
Category Modeling using Clustering Models

- Finding relevant clusters:
  - Checked both models from 4 to 13 clusters each
  - Manually checked which documents each cluster contained to assess how well the model was performing
  - Determined KMeans was clustering better than Agglomerative Clustering
  - Finished by choosing a KMeans model with relevant clusters

Category Modeling using Clustering Models

- Visualization of clusters obtained from KMeans:





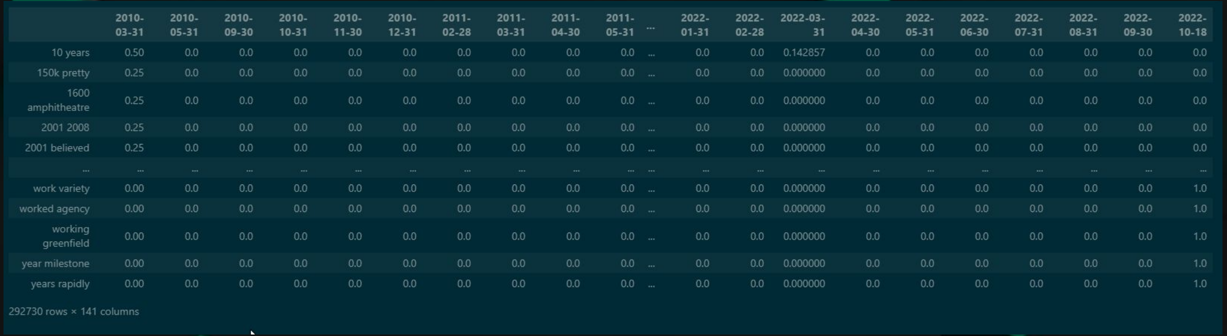
Tag	Opinion	Technology	Innovation & Design	Ways of Working	Culture	Events	Emerging Tech	Strategy	News	Learning	Projects	Product
Futustories	4	2	0	3	76	4	0	1	3	2	0	0
Company	8	3	17	6	3	6	3	35	5	3	1	0
Data	3	6	10	0	1	2	3	14	1	3	0	1
Mobility	12	3	13	0	0	1	1	6	1	1	0	0
Strategy	2	4	1	10	3	0	0	3	0	8	2	0
AI	1	4	0	0	2	4	16	2	2	1	2	0
Futurice	16	16	11	15	46	43	5	4	19	17	21	1
Design	5	4	32	10	4	18	7	6	2	9	6	1
Energy	10	7	0	0	0	0	0	2	1	1	0	0
Cloud	2	65	1	7	0	3	8	0	1	15	2	1

Words popularity

- The blogs are divided into different periods
- Use **`CountVectorizer`** to tokenize the blogs into bigrams and calculate the frequency of those bigrams

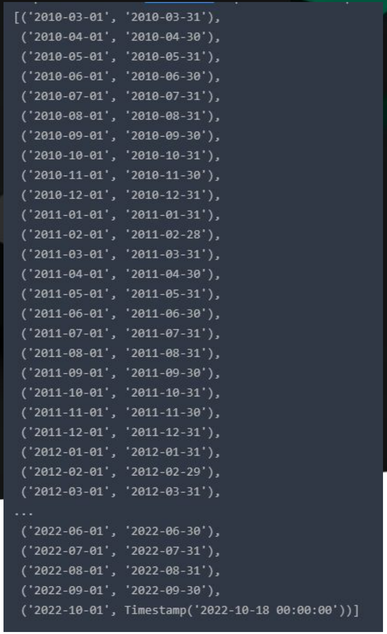
Words popularity

- The blogs are divided into different periods
- Use `CountVectorizer` to tokenize the blogs into bigrams and calculate the frequency of those bigrams



	2010-03-31	2010-05-31	2010-09-30	2010-10-31	2010-11-30	2010-12-31	2011-02-28	2011-03-31	2011-04-30	2011-05-31	...	2022-01-31	2022-02-28	2022-03-31	2022-04-30	2022-05-31	2022-06-30	2022-07-31	2022-08-31	2022-09-30	2022-10-18
10 years	0.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.142857	0.0	0.0	0.0	0.0	0.0	0.0	0.0
150k pretty	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1600 amphitheatre	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2001 2008	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2001 believed	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
work variety	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	1.0
worked agency	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	1.0
working greenfield	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	1.0
year milestone	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	1.0
years rapidly	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	1.0

292730 rows × 141 columns



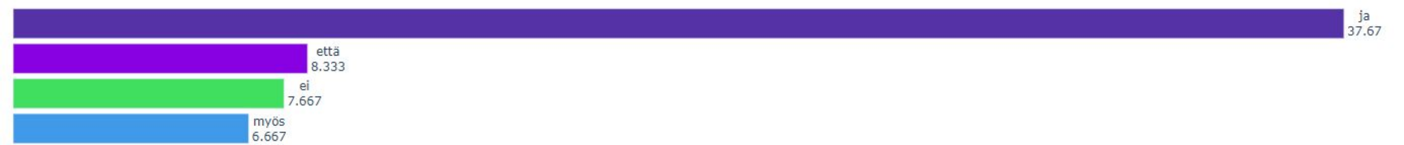
```
[('2010-03-01', '2010-03-31'),
 ('2010-04-01', '2010-04-30'),
 ('2010-05-01', '2010-05-31'),
 ('2010-06-01', '2010-06-30'),
 ('2010-07-01', '2010-07-31'),
 ('2010-08-01', '2010-08-31'),
 ('2010-09-01', '2010-09-30'),
 ('2010-10-01', '2010-10-31'),
 ('2010-11-01', '2010-11-30'),
 ('2010-12-01', '2010-12-31'),
 ('2011-01-01', '2011-01-31'),
 ('2011-02-01', '2011-02-28'),
 ('2011-03-01', '2011-03-31'),
 ('2011-04-01', '2011-04-30'),
 ('2011-05-01', '2011-05-31'),
 ('2011-06-01', '2011-06-30'),
 ('2011-07-01', '2011-07-31'),
 ('2011-08-01', '2011-08-31'),
 ('2011-09-01', '2011-09-30'),
 ('2011-10-01', '2011-10-31'),
 ('2011-11-01', '2011-11-30'),
 ('2011-12-01', '2011-12-31'),
 ('2012-01-01', '2012-01-31'),
 ('2012-02-01', '2012-02-29'),
 ('2012-03-01', '2012-03-31'),
...
 ('2022-06-01', '2022-06-30'),
 ('2022-07-01', '2022-07-31'),
 ('2022-08-01', '2022-08-31'),
 ('2022-09-01', '2022-09-30'),
 ('2022-10-01', Timestamp('2022-10-18 00:00:00'))]
```

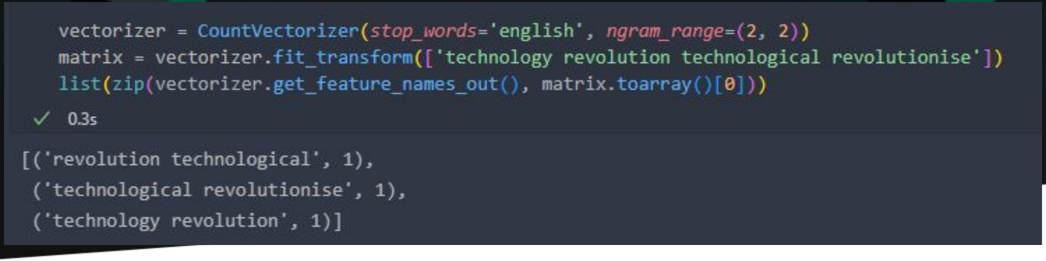
Words popularity

- Current weakness:
    - Cannot filter out foreign language words (e.g., Finnish)
    - Words that share the same stems will be counted as separate (e.g. technology revolution vs technology revolutionise)

Words popularity

- Current weakness:
  - Cannot filter out foreign language words (e.g., Finnish)
  - Words that share the same stems will be counted as separate (e.g. technology revolution vs technology revolutionise)



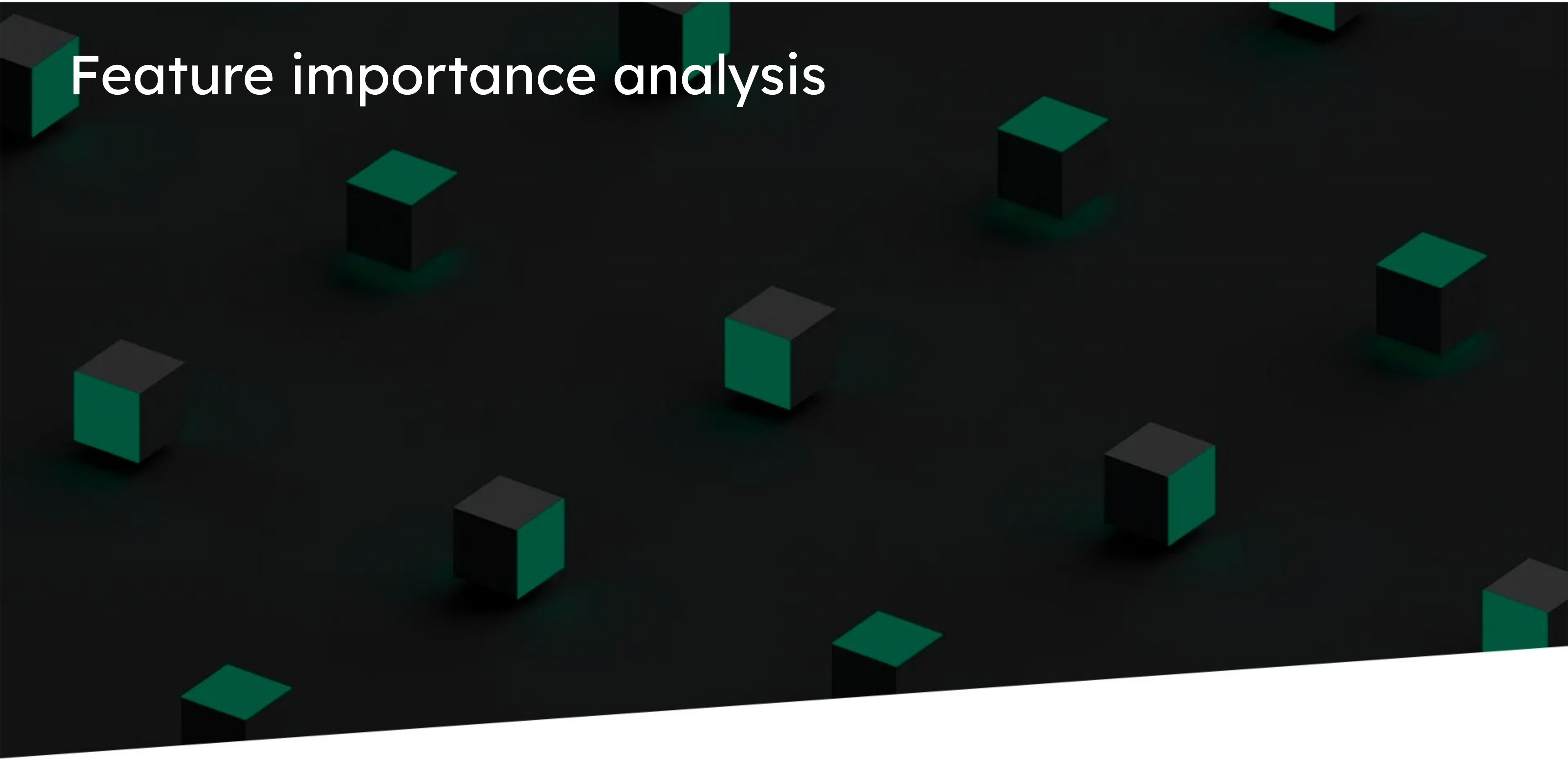


```
vectorizer = CountVectorizer(stop_words='english', ngram_range=(2, 2))
matrix = vectorizer.fit_transform(['technology revolution technological revolutionise'])
list(zip(vectorizer.get_feature_names_out(), matrix.toarray()[0]))
```

✓ 0.3s

```
[('revolution technological', 1),
 ('technological revolutionise', 1),
 ('technology revolution', 1)]
```

Feature importance analysis

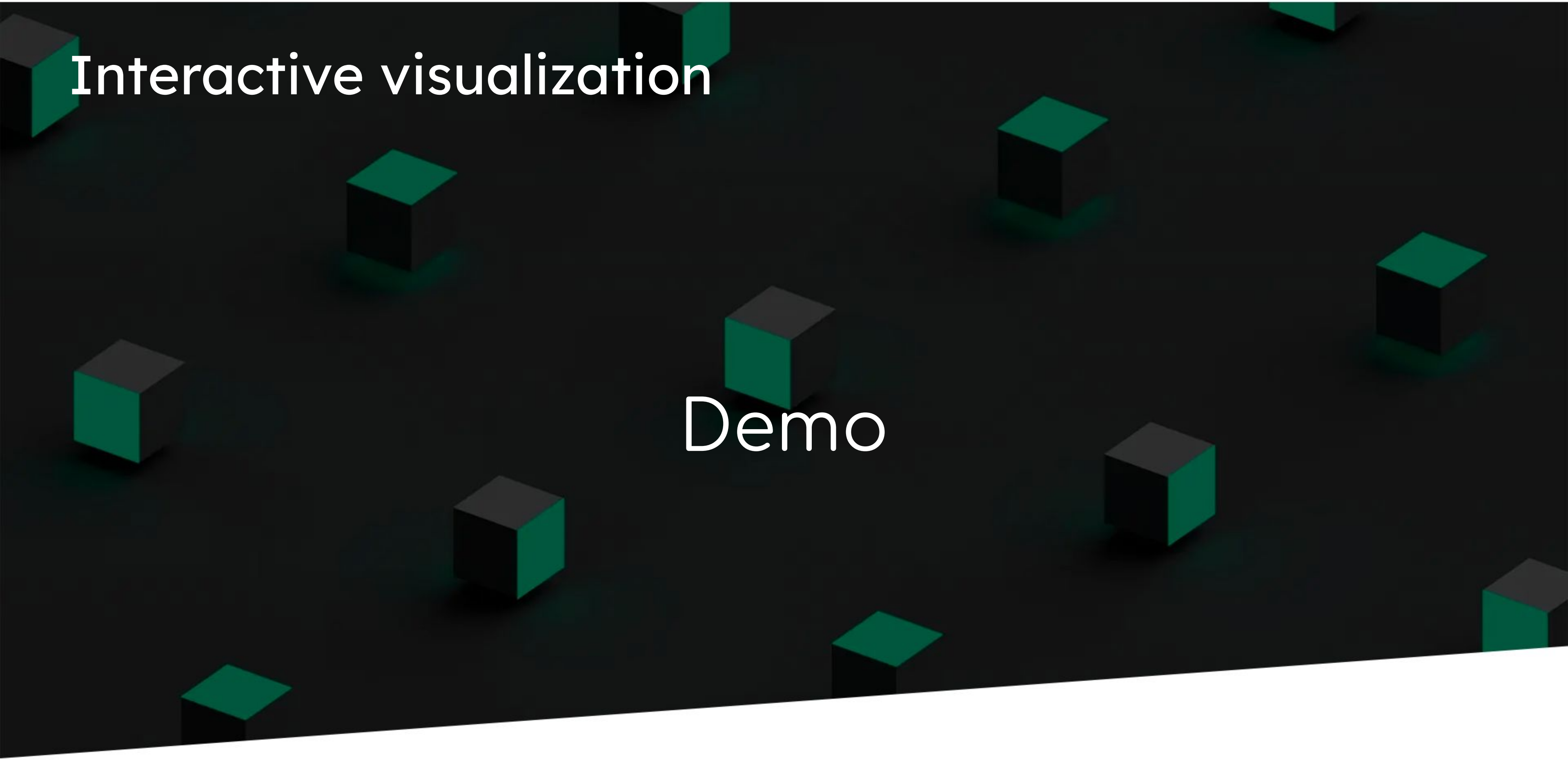


Feature importance analysis

- **Idea:** use coefficients of a linear regressor to estimate importance of the features
- **Augmentation:** apply shrinkage method to get a more meaningful estimate
- **Another augmentation:** replace linear regressor with random forest regresstion

Blog Doctor

Demo



Interactive visualization

Demo

Future Prospects

For the next week:

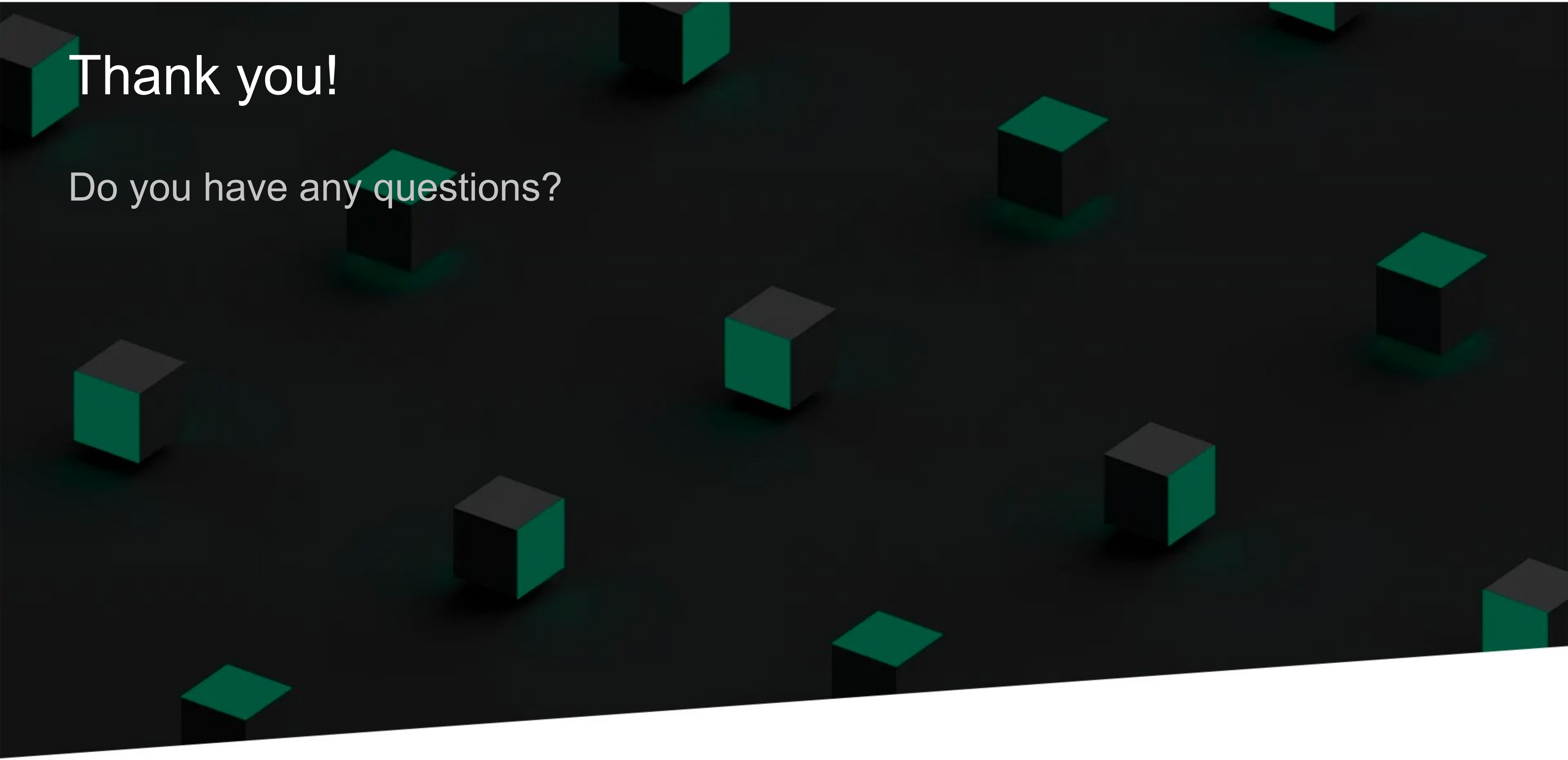
- Clean up the repository
- Automate database updates
- Compile the final report

Further opportunities:

- Incorporate external datasets
- Compare topics of articles from Futurice with that of other companies
- Explore connectivity-based features, features based on titles/previews

cluster_temp_save.csv
clustering_with_text.ipynb
document_clusters.html
experimenting_with_text.i...
experimenting_with_titles....
GloVe.ipynb
Keywords Extraction.ipynb
README.md
regression_analysis.ipynb
statistic_graph_stuff.ipynb
README.md
requirements.txt
.gitattributes

blog_corpus.txt
blog_text.csv
blog_urls_all.csv
blogs_with_analytics.csv
google_analytics.csv
linked.json
README.md
vectors.txt
data_processing



Thank you!

Do you have any questions?