Apprenticeship report and Memoire
Topic extraction from customer call logs using word embeddings and clustering techniques

Khang Duy LAI
Apprenticeship supervisor **Romain DADEN**
Academic tutor **Kim NGUYEN**

Université Paris-Saclay

4/9/2023

Company introduction

Sector of activity

- SFR stands for Société française du radiotéléphone.
- A French telecommunications company.
- The second oldest mobile network and telecommunications company in France.

Sector of activity

Main business sectors

- Mobile services
- Fixed-line services
- Internet services
- Television services

Sector of activity

Main business sectors

- Mobile services
- Fixed-line services
- Internet services
- Television services

- Business to consumer
- Business to business

Application of data in business operation

Data is the most valuable asset for businesses, help to improve operational efficiency, marketing and advertising, optimize pricing, etc.

-> SFR Analytics department is responsible to collect, analyze and draw company's vision.

Organization

- Data science team
- Data analyst team
- Data governance team

Application of data in business operation

Data is the most valuable asset for businesses, help to improve operational efficiency, marketing and advertising, optimize pricing, etc.

-> SFR Analytics department is responsible to collect, analyze and draw company's vision.

Organization

- Data science team
- Data analyst team
- Data governance team

Team's main missions

- Reducing churn rate.
- Image processing for infrastructure optimization.
- Customer's feedback analysis.
- Predictive analytics for business management.

Missions

Roles

- One of the member of the **data science** team.
- Contribute to the company's datadriven decision-making processes and strategic initiatives.
- Extracting valuable insights from the vast volumes of data generated within the telecommunications landscape.
- Forcast future situation using machine learning and deep learning techniques.
- Automate some manual tasks.

Roles

- One of the member of the **data science** team.
- Contribute to the company's datadriven decision-making processes and strategic initiatives.
- Extracting valuable insights from the vast volumes of data generated within the telecommunications landscape.
- Forcast future situation using machine learning and deep learning techniques.
- Automate some manual tasks.

Some of the projects in charge

- Fiber distribution point fake and near-duplicated images detection.
- Research of moving team's infrastructure to Google Cloud.
- Topic extraction from customer call logs using word embeddings and clustering techniques.

Fiber distribution point fake and near-duplicated images detection

Fiber distribution point fake and near-duplicated images detection

- A fiber distribution point (PM) is a technical cabinet.
- Operators have a network of PMs.

Fiber distribution point fake and near-duplicated images detection

- A fiber distribution point (PM) is a technical cabinet.
- Operators have a network of PMs.
- The system is shared between operators to reduce cost.

Fiber distribution point fake and near-duplicated images detection

- A fiber distribution point (PM) is a technical cabinet.
- Operators have a network of PMs.
- The system is shared between operators to reduce cost.
- Outsourcing the maintenant works to other company.

Fiber distribution point fake and near-duplicated images detection

- A fiber distribution point (PM) is a technical cabinet.
- Operators have a network of PMs.
- The system is shared between operators to reduce cost.
- Outsourcing the maintenant works to other company.
  - Required to capture the image of the PM after finish works.
  - Some forgery images have been sent to the system.

Fiber distribution point fake and near-duplicated images detection

- A fiber distribution point (PM) is a technical cabinet.
- Operators have a network of PMs.
- The system is shared between operators to reduce cost.
- Outsourcing the maintenant works to other company.
  - Required to capture the image of the PM after finish works.
  - Some forgery images have been sent to the system.
    - Slightly rotate the image
    - Add into the image a color bar or put it inside a color box
    - Cropped
    - Color shift

Fiber distribution point fake and near-duplicated images detection

Fiber distribution point fake and near-duplicated images detection

The system right now using image fingerprint (Perceptual hashing) to detect forgery image.

The accuracy using this system is more than 90%.

Fiber distribution point fake and near-duplicated images detection

The system right now using image fingerprint (Perceptual hashing) to detect forgery image.

The accuracy using this system is more than 90%.

Objective

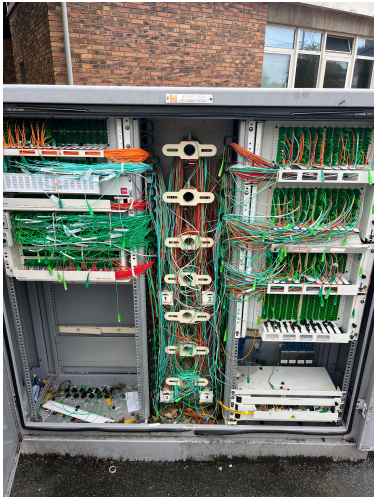
- Deploy a supplement forgery detection system for that 10% inaccuracy.

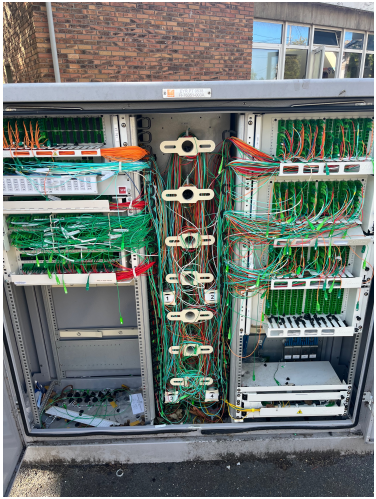
Forgery case example

Figure 1: Original image

Figure 2: Forgery image

False positive case example





Methods

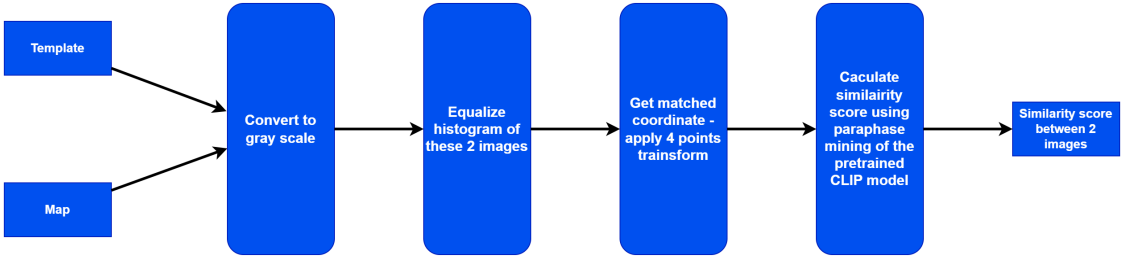
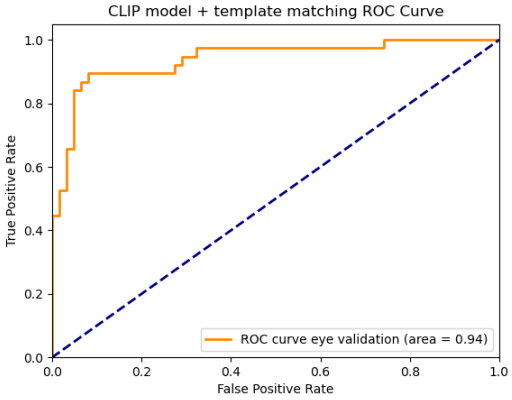


Figure 3: Overview of the used system

Result



seuil	nb_FP	nb_FN	nb_err_tot
90%	38	1	39
90,50%	37	1	38
91%	35	1	36
91,50%	33	1	34
92%	30	1	31
92,50%	28	1	29
93%	27	1	28
93,50%	25	1	26
94%	23	1	24
94,50%	23	1	24
95%	17	1	18
95,50%	14	3	17
96%	13	4	17
96,50%	7	4	11
97%	3	4	7
97,50%	2	6	8
98%	0	15	15
98,50%	0	22	22
99%	0	26	26
99,50%	0	28	28
100%	0	42	42

Result

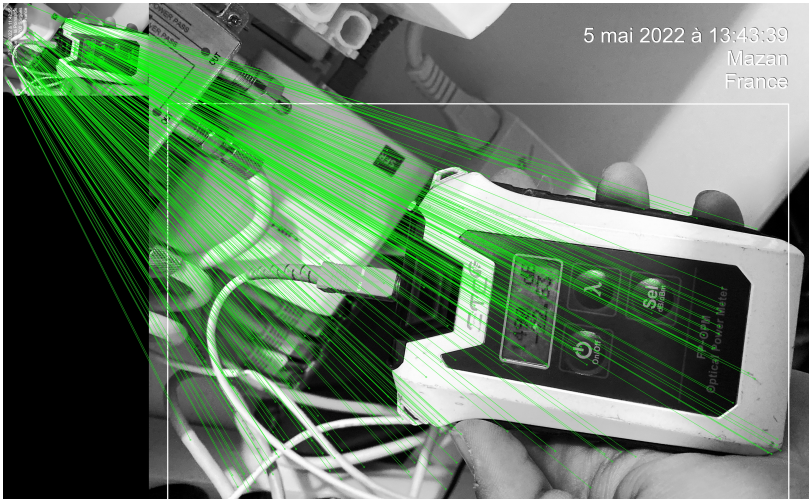


Figure 4:

Cloud migration in SFR Analytics

Cloud migration in SFR Analytics

Current system of the team is based on self hosted server.

Cloud migration in SFR Analytics

Current system of the team is based on self hosted server.

Have many drawbacks

- Need to maintain ourself and regularly update.
  - Maintainance complexity is high.
- Hard to scale.
- High cost.

Cloud migration in SFR Analytics

Current system of the team is based on self hosted server.

Have many drawbacks

- Need to maintain ourself and regularly update.
  - Maintainance complexity is high.
- Hard to scale.
- High cost.

Solution

Research to migrate to the cloud.

- The company has 6 months trial contract with Google.

Cloud migration in SFR Analytics

Role: Research usage of VertexAI, a platform designed to help organizations build, deploy, and manage machine learning models at scale.

Current system

- R script automation using Airflow.
- The system has shown not so reliable.

Proposed system

- Apply MLOps on models management using VertexAI.

Cloud migration in SFR Analytics

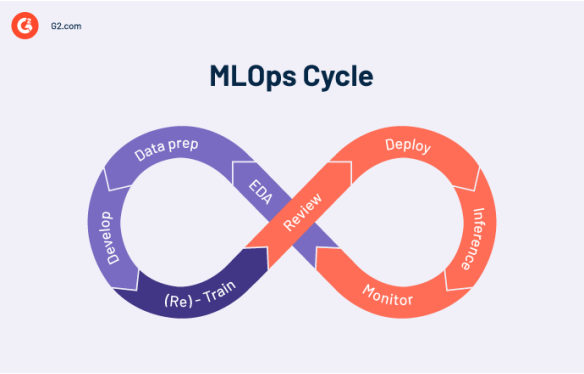


Figure 5: MLOps Cycle

Cloud migration in SFR Analytics

Researchs

- Vertex AI workbench to work direcly on the cloud.
- Run custom models.
- Work with other GCP services.
  - For example, create dataset from Google Big Querry
- Create custom docker images for training.
- Run custom evaluation.

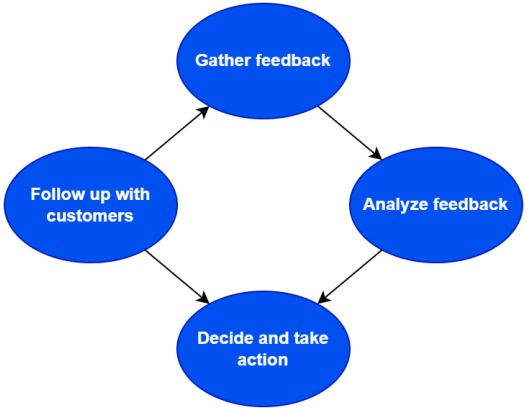
Cloud migration in SFR Analytics

Propose accepted.

- The migration will begin and starting from November 2023.
- Sad not to be the member who participate in the real migration.

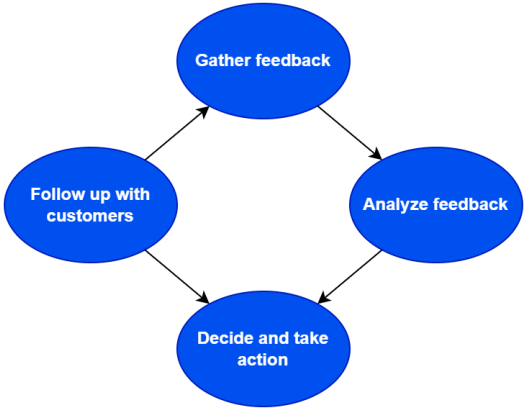
Topic extraction from customer call logs using word embeddings and clustering techniques

Introduction



- Ensuring customer satisfaction stands as one of the most important goals for enterprises across industries.
- The collection and analysis of invaluable data derived from customer interactions is the key to improving companies' competitive advantage.
- Call logs is one of the valuable data to improve customer sastifaction.

Introduction



- Ensuring customer satisfaction stands as one of the most important goals for enterprises across industries.
- The collection and analysis of invaluable data derived from customer interactions is the key to improving companies' competitive advantage.
- Call logs is one of the valuable data to improve customer sastifaction.

Objective

The necessary to find topics, which is the significant problems that customers are facing to improve overall quality of the service.

Dataset

	contact_id	transcript_speaker	transcript_word	transcript_starttime
1	25579914	2	ça	00:00:03.5000000
2	25579914	2	va	00:00:03.6000000
3	25579914	2	pas	00:00:03.8000000
4	25579914	1	on	00:00:03.9000000
5	25579914	1	peut	00:00:04.1000000
6	25579914	1	faire	00:00:04.2000000
7	25579914	1	bonjour	00:00:04.4000000
8	25579914	1	laura	00:00:04.8000000
9	25579914	1	à	00:00:05
10	25579914	1	votre	00:00:05.1000000
11	25579914	1	écoute	00:00:05.2000000
12	25579914	2	bonjour	00:00:07.1000000
13	25579914	2	c`est	00:00:07.7000000
14	25579914	2	vraiment	00:00:07.9000000
15	25579914	2	le	00:00:08.2000000
16	25579914	2	numéro	00:00:08.3000000

Dataset

- Need to sign a special permission to have access.
- Final result is to have the acquired conversation in order.
- SQL query orderby `transcript_starttime`, `contact-id` and `transcript_speaker`.

Architechture

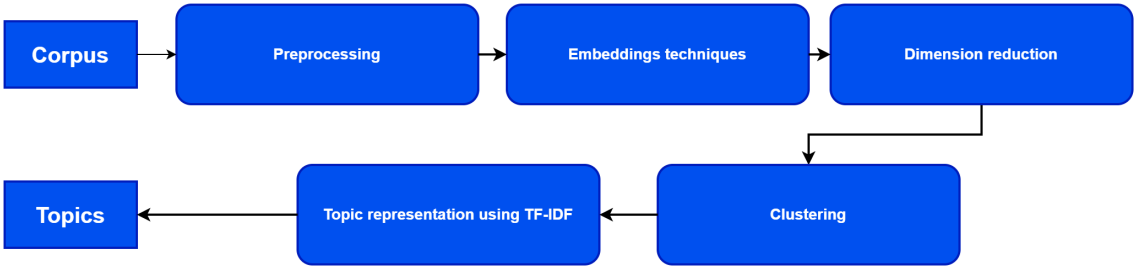


Figure 6: Overal architechture of the system

Data preprocessing

- It is not recommended to clean the corpus before apply embeddings method.
- The dataset acquired is the transcription of conversations.
- Have many filler words.

Data preprocessing

- It is not recommended to clean the corpus before apply embeddings method.
- The dataset acquired is the transcription of conversations.
- Have many filler words.

```
filler_word = ['redacted', 'bonjour', 'oui', 'monsieur', 'madame',
'ben', 'quelque', 'moment', 'justement', 'euh', 'non', 'hein',
'collègue', 'hein', 'allô', 'ah', 'alors', 'voilà', 'bah']
```

Data preprocessing

- **Original:** oui monsieur bonjour j'ai un problème avec ma ligne fixe euh je le note oui oui oui oui oui c'est-à-dire que je je n'arrive pas à téléphoner que vous avez un numéro mais ça sera pas ça donne rien du tout sais plus comment j'ai pas compris monsieur je vais vous dire depuis au moins 3 ou 4 jours alors j'avais une personne chez vous qui m'a dit que y'avait un problème sur le réseau que ça allait être solutionné mais en fin de compte j'avais pas du tout oui à londres. . .
- **Hard clean:** j'ai problème ligne fixe note c'est-à-dire n'arrive téléphoner numéro ça ça donne rien tout sais plus comment j'ai compris vais dire depuis moins jours j'avais personne chez m'a dit y'avait problème réseau ça allait être solutionné fin compte j'avais tout londres. . .
- **Soft clean:** j'ai un problème avec ma ligne fixe je le note c'est-à-dire que je je n'arrive pas téléphoner que vous avez un numéro mais ça sera pas ça donne rien du tout sais plus comment j'ai pas compris je vais vous dire depuis au moins ou jours j'avais une personne chez vous qui m'a dit que y'avait un problème sur le réseau que ça allait être solutionné mais en fin de compte j'avais pas du tout londres. . .

Documents embeddings

2 methods of sentence transformer:

- CamemBERT (768 dimensions)
- miniLMv2 multilangual (384 dimensions)

Dimension reduction

2 methods:

- PCA with 80% of variance
  - Primarily used for linear dimensionality reduction
- UMAP - enhanced method from T-SNE
  - A nonlinear dimensionality reduction technique that focuses on preserving the pairwise similarities between data points
  - UMAP is faster than T-SNE

Clustering

3 methods:

- Kmeans
- HDBSCAN
- Hierarchy clustering

Word extraction

We need to use a modified version of TF-IDF.

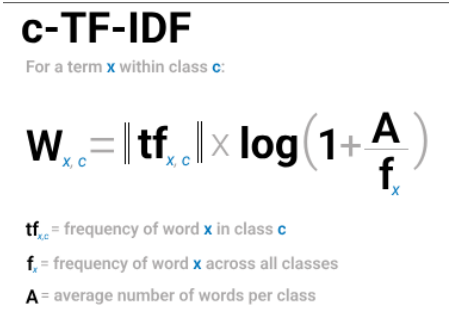


Figure 7: Modified version of TF-IDF

Results - CamemBERT

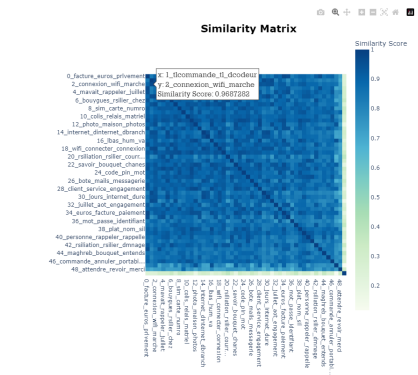


Figure 8: Hierarchy + PCA

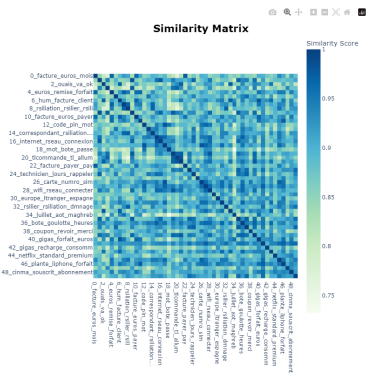


Figure 9: Hierarchy + UMAP

Results - MiniLMv2

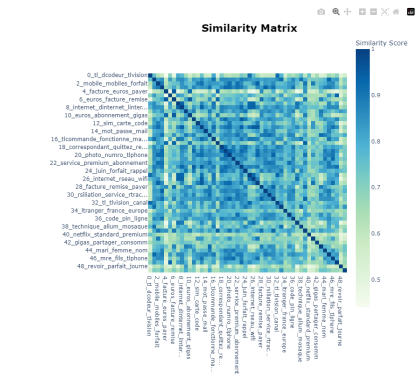


Figure 10: Hierarchy + PCA

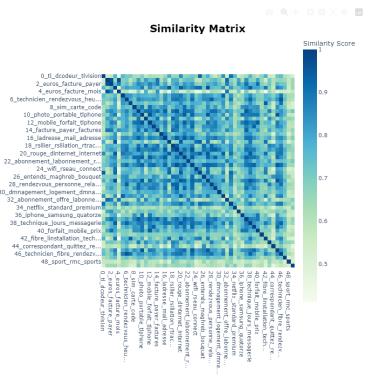


Figure 11: Hierarchy + UMAP

Results

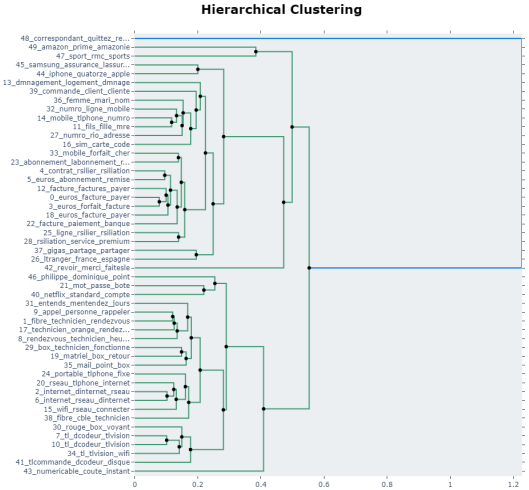


Figure 12:

Future works

We can see the topics of the call in most of techniques, however, there are still works to do.

- Hyperparameter tuning.
- Build an API to return suggested words.

Conclusion

Memoire

- Can see the needed topics with the result match the prediction before.
- MiniLM tends to have better performance than CamemBERT, and it is also much faster to inference.
- HDBSCAN works not very good with PCA, presume to be some pairwise similaries between data points have been lost.

Appenticeship duration

The apprenticeship offers a valuable opportunity for me to improve my skills, and a chance to apply what I have learned over the courses at the university on real projects.