Démonstration : exploration sémantique de données texte de la relation client

Guillaume Dubuisson Duplessis, François Bullier, Anne-Laure Guénet

EDF Commerce, Direction des Systèmes d'Information et du Numérique (DSIN),
CSC datascience & IA, 420 rue Estienne d'Orves, 92700 Colombes

{guillaume.dubuisson-duplessis, anne-laure.guenet}@edf.fr

Résumé

L'exploration des données texte est une tâche-clé non-supervisée préliminaire à l'exploitation de ces données dans les cas d'usage opérationnels. Cet article décrit l'outil `nemo` *(« Neural sEMantic explOration ») récemment développé en interne à EDF Commerce pour servir nos besoins opérationnels et qui a déjà été utilisé dans une dizaine de cas d'usage.* `nemo` *facilite grandement l'exploration sémantique, rendant la tâche beaucoup plus aisée.*

Mots-clés

TALN, exploration de données texte, plongement vectoriel de documents

Abstract

Text data exploration is a crucial unsupervised task that needs to be performed prior to utilizing the data for operational purposes. This article describes `nemo` *(« Neural sEMantic explOration »), which has recently been developed internally at EDF Commerce to meet our operational needs and has already been used in around ten use cases.* `nemo` *enables semantic text data exploration, greatly facilitating the exploration task.*

Keywords

NLP, text data exploration, sentence embedding

1 Introduction

La relation client au sein d'EDF Commerce génère chaque mois des millions de données texte aussi bien de la part des clients (e.g., e-mails, réponses libres à des questionnaires de satisfaction) que de la part des conseillers (e.g., commentaires de contact). Ces données majoritairement en français sont riches. Elles offrent un large panel de structures allant d'expressions libres et spontanées à des formes contraintes comme des formulaires. Elles présentent également une grande diversité en ce qui concerne le respect de l'orthographe, de la syntaxe et du niveau de langue. Ces données sont utilisées pour répondre au mieux aux attentes de nos clients en suivant le cadre réglementaire du « règlement général sur la protection des données » (RGPD) [3]. En outre, elles sont exploitées dans de nombreux cas d'usage visant à optimiser la relation client (e.g., sur le canal e-mail [4]).

L'exploration des données texte est une tâche-clé non-supervisée préliminaire à l'exploitation de ces données dans les cas d'usage opérationnels. Cette exploration vise à mieux connaître les données en faisant ressortir les principaux sujets abordés dans un corpus. Elle permet la création de plans d'annotation pertinents en alignant le contenu des données texte aux besoins « métier » et le développement de solutions techniques basées sur l'apprentissage supervisé. Les techniques exploratoires font également partie intégrante de la boîte à outils de l'ingénieur de la donnée pour mener à bien ses projets. Par exemple, des techniques d'extraction de paraphrases se révèlent particulièrement utiles pour augmenter des données dans un contexte où l'annotation manuelle est coûteuse, ou pour aider à la modélisation linguistique. Cet article décrit l'outil `nemo` (« Neural sEMantic explOration ») récemment développé en interne à EDF Commerce pour servir nos besoins opérationnels et qui a déjà été utilisé dans une dizaine de cas d'usage. `nemo` permet une exploration sémantique des données texte plus efficace, améliorant considérablement la qualité des résultats obtenus. Il intègre des fonctionnalités utiles pour un cadre opérationnel et des modèles spécialisés sur les données de la relation client d'EDF Commerce.

La Section 2 positionne `nemo` par rapport aux travaux existants. La Section 3 expose les fonctionnalités centrales de `nemo` ainsi que ses avantages et limites. La Section 4 aborde la création de modèles spécialisés au domaine de la relation client d'EDF et souligne l'intérêt de ces modèles pour l'exploration de données dans un cadre opérationnel. Enfin, la Section 5 conclut cet article et indique les principales perspectives de ces travaux.

2 Travaux connexes

La découverte de structures et de thèmes dans des données texte est une tâche traitée de longue date par la communauté du traitement automatique de la langue naturelle. A cette fin, certaines approches sont bien établies parmi lesquelles la « Latent Dirichlet Allocation » (LDA) [1] et la méthode de Reinert popularisée par l'outil Iramuteq [12]. Ces approches sont fondamentalement limitées par leur dépendance à une représentation du texte sous forme de « sac de mots » qui échoue à capturer la richesse sémantique derrière les mots et *in fine* à représenter fidèlement un docu-

ment texte. Pour faire face à cette limite, de nouvelles représentations du texte sémantiquement plus riches ont vu le jour sous la forme de plongements vectoriels de mots non-contextualisés (e.g., word2vec [10], GloVe [11]) et contextualisés fondés sur l'architecture *transformer* [2, 8, 7]. Sentence-BERT (SBERT) utilise ingénieusement ces derniers pour générer des plongements vectoriels de bonne qualité au niveau du document [13].

Notre outil `nemo` revisite l'exploration de texte à l'heure de ces représentations sémantiquement plus riches. Il exploite les modèles pré-entraînés fournis par SBERT et met à disposition de nos utilisateurs des modèles spécialisés sur le domaine de la relation client EDF. A l'heure où nous écrivons cet article, BERTopic [5] poursuit un objectif similaire à celui de `nemo` et évolue rapidement. `nemo` se distingue par ses fonctionnalités de descriptions d'un cluster (hiérarchie, extraction de mots-clés, extraction de prototypes), ses modèles spécialisés et son utilisation éprouvée au sein de nombreux cas d'usage opérationnels.

3 Présentation de l'outil `nemo`

`nemo` (« Neural sEMantic explOration ») est une librairie Python utilisable dans des notebooks Jupyter facilitant l'exploration sémantique des données texte à EDF. `nemo` repose sur une représentation dense des documents basée sur les sentence-embeddings [13] et permet d'encoder simplement un corpus via un appel de fonction. Par défaut, `nemo` propose l'utilisation du modèle paraphrase-multilingual-mpnet-base-v2 [1].

3.1 Fonctionnalités proposées par `nemo`

3.1.1 Clustering

`nemo` fournit plusieurs approches de clustering à ses utilisateurs. Deux approches sont plus fréquemment utilisées. La première est HDBSCAN [9] qui fonde le clustering sur la densité des points. Cette approche a l'avantage d'extraire les clusters les plus denses mais elle exclut de nombreux documents. Lorsque l'exhaustivité est requise pour un cas d'usage, les utilisateurs se tournent le plus souvent vers un clustering hiérarchique plus classique. Cette approche a le double avantage d'inclure l'ensemble des documents et d'offrir une vision hiérachique des clusters via un dendrogramme.

3.1.2 Visualisation et exploration de données

`nemo` simplifie la découverte d'un corpus de données texte en offrant une visualisation interactive de documents regroupés par similarité sémantique. La Figure 1 montre une impression écran de la visualisation du clustering d'un corpus de données et de la possibilité de l'explorer interactivement. Chaque point représente un document tandis que chaque couleur représente un cluster préalablement calculé.

1. La carte de ce modèle est disponible à cet URL : `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

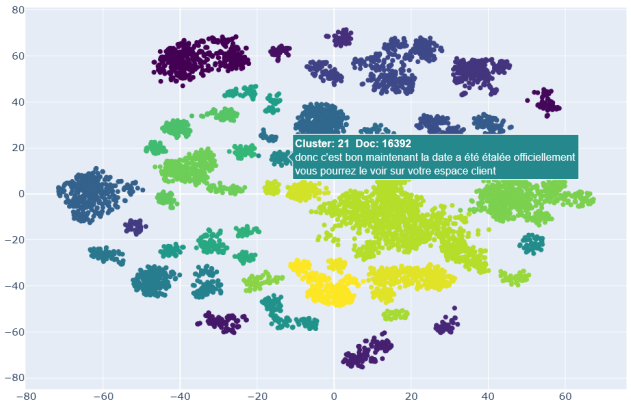

FIGURE 1 – Visualisation d'un corpus de données texte avec `nemo`. Le survol d'un point permet d'afficher le texte du document. Les couleurs représentent des clusters calculés avec HDBSCAN.

3.1.3 Descriptions des clusters

`nemo` facilite la description des clusters sur deux niveaux. Le premier niveau est celui des mots-clés. `nemo` permet l'extraction des mots-clés pertinents d'un cluster via plusieurs approches au choix basées sur le TF-IDF, le cTF-IDF [5], le chi2 [12] ou encore via des graphes de mots proposés par la librairie gowpy [2]. Le second niveau est celui des documents. `nemo` permet l'extraction de parangons – les documents représentatifs – pour chaque cluster via l'algorithme ProtoDash [6].

3.1.4 Recherche sémantique

Enfin, `nemo` permet de rechercher des documents sémantiquement proches dans l'ensemble du corpus. Cette fonctionnalité est particulièrement pratique pour l'extraction de paraphrases afin d'augmenter des données ou pour aider la modélisation linguistique. Par exemple, la recherche de documents similaires à l'énoncé « je n'ai plus de courant » retourne des paraphrases telles que « on est en panne d'électricite » ou « j'ai tout sauf l'électricite ».

3.2 Discussion sur les fonctionnalités

`nemo` a l'avantage de fournir des résultats pertinents rapidement. Un gain de temps appréciable provient du prétraitement minimal des textes comparativement aux approches classiques. La pertinence des résultats est nettement améliorée dans nos cas d'usage grâce aux représentations sémantiques fournies par les sentence-embeddings.

Une limite concerne la taille des documents. Plus un document est grand, plus il y a de thèmes qui peuvent y être abordés. Il est souvent nécessaire de découper ces documents pour améliorer les résultats de l'analyse. Une autre contrainte concerne le fait que les modèles librement disponibles atteignent leur limite sur des domaines spécifiques comme la relation client d'EDF. Cela motive un travail de spécialisation de modèles.

2. La librairie gowpy est disponible à cet URL : `https://github.com/GuillaumeDD/gowpy`

4 Spécialisation au domaine de la relation client d'EDF

4.1 Objectif et démarche

La spécialisation des modèles de sentence-embeddings vise à encoder la sémantique du domaine de la relation client EDF qui manque aux modèles pré-entraînés librement disponibles (sans pour autant oublier la sémantique « générale »). Un modèle spécialisé doit, par exemple, capturer la relation de synonymie entre les expressions « PDL », « PRM » et « point référence mesure » qui font référence au point de livraison de l'énergie au client dans les données texte d'EDF. La démarche de spécialisation est la suivante. Premièrement, nous avons construit semi-automatiquement un jeu de données de paraphrases spécialisées pour la relation client d'EDF Commerce (Section 4.2). Ce jeu de données est ensuite utilisé pour spécialiser le modèle de paraphrases librement disponible paraphrase-multilingual-mpnet-base-v2. Enfin, les modèles de paraphrases spécialisés sont validés sur plusieurs tâches *proxy* pour vérifier leur qualité (cf. Section 4.3).

4.2 Construction du jeu de données de paraphrases

La construction d'un jeu de données de paraphrases spécifique à EDF Commerce n'a pas été faite de zéro pour des raisons de coûts d'annotation manuelle, mais via l'utilisation de jeux de données existants anonymisés. En vertu du règlement général sur la protection des données (RGPD), les données texte sont désidentifiées et ne comportent aucune donnée à caractère personnel telle que des noms, prénoms, adresses, numéros [3]. Les données à caractère personnel sont substituées par le type de l'entité (e.g., "Je suis *[person]* et je vous appelle pour obtenir la facture de mon logement situé à *[localisation]*"). Nous avons utilisé des jeux de données annotés en étiquettes thématiques couvrant plusieurs types de données comme des emails ou des énoncés adressés à des chatbots. Ces types représentent des documents de taille variée allant de quelques mots à plusieurs paragraphes. Les étiquettes sont de granularité diverse allant d'étiquettes fines (e.g., intention dans le cadre d'un NLU de chatbot), moyennes (e.g., « contrat »), à grosses (e.g., « réclamation »). Nous avons spécifié manuellement la similarité en fonction des étiquettes attribuées aux documents. Schématiquement, une paire positive est formée par des documents ayant les mêmes étiquettes (similarité strictement supérieure à 0 et inférieure ou égale à 1), et une paire négative est formée par des documents dans des étiquettes opposées (similarité à 0). Nous avons assimilés les documents avec les mêmes étiquettes fines à des paraphrases parfaites (similarité à 1). La similarité est diminuée pour des étiquettes à grain plus large. Pour chaque étiquette, nous avons équilibré le nombre de paires positives et négatives.

Lors de nos différentes expérimentations, nous avons noté l'importance des paires fortement opposées (appelées « hard negatives » en anglais). Il s'agit de documents ayant les mêmes mots mais des sens différents comme « je n'ai plus de courant » et « je ne suis pas au courant ». Nous avons spécifié manuellement les étiquettes fortement opposées et nous avons utilisé une technique de minage basée sur BM25 [14] pour créer des paires négatives de documents fortement opposés.

4.3 Entraînement et sélection des modèles

La spécialisation des modèles a nécessité plusieurs itérations en faisant varier les données d'entraînement (plusieurs shémas de similarité sémantique entre étiquettes, différentes façons de miner des « hard negatives ») et le nombre de paires utilisées pour adapter le modèle. Le finetuning a utilisé des hyperparamètres standards, à savoir un taux apprentissage de $1e-5$, un nombre d'epochs de 2, et une taille de batch de 4. Nous avons validé la qualité des modèles sur trois tâches *proxy* : clustering (via le score de silhouette), classification (via la métrique F1) et en similarité sémantique (via la corrélation). Les meilleurs modèles obtenus sont ceux basés sur un schéma de pondération prenant en compte le grain des étiquettes dans les paires d'apprentissage (i.e. utilisant l'ensemble de l'échelle de 0 à 1) et en intégrant des « hard negatives » minés avec BM25. Nous avons conservé plusieurs modèles allant d'une spécialisation légère (100 paires par étiquette) à une spécialisation prononcée (1500 paires par étiquette). L'utilisateur de `nemo` a la main pour choisir le degré de spécialisation en fonction des données de son cas d'usage.

4.4 Apport des modèles spécialisés

L'apport attendu des modèles spécialisés est l'amélioration de l'exploration de données texte sur le domaine de la relation client. La Figure 2 illustre l'impact positif des modèles de sentence-embeddings spécialisés sur la tâche de clustering et de visualisation des données en 2D. Sur cet exemple d'utilisation des modèles spécialisés sur un nouveau jeu de données non-utilisé pour la spécialisation, on voit apparaître un nuage de documents avec des clusters plus marqués et plus homogènes d'un point de vue sémantique. Cet apport se confirme sur d'autres cas d'usage.

5 Conclusion et perspectives

Cet article a présenté `nemo`, un outil d'exploration sémantique de données texte de la relation client d'EDF qui simplifie et accélère considérablement la phase exploratoire de nos projets. Les perspectives incluent l'ajout d'indicateurs pour qualifier la pertinence sémantique des clusters, une meilleure prise en compte des documents longs et ultimement le rapprochement des phases d'exploration de texte et de modélisation.

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Sofiane Kerroua, Mathilde Poulain, Mathilde Jeuland, Kamar Koraibi, Irène Truche, Aurore Hamimi, Philippe Very, Mourad Miled, Subhi Issa, Oualid Akhsass, Marwen Touzi, Florent Mely, Maëlle Voisinet, Amandine Bessou, Laetitia Leroux, Laura Rouhier, Sonia Audheon, Marie Hervé,

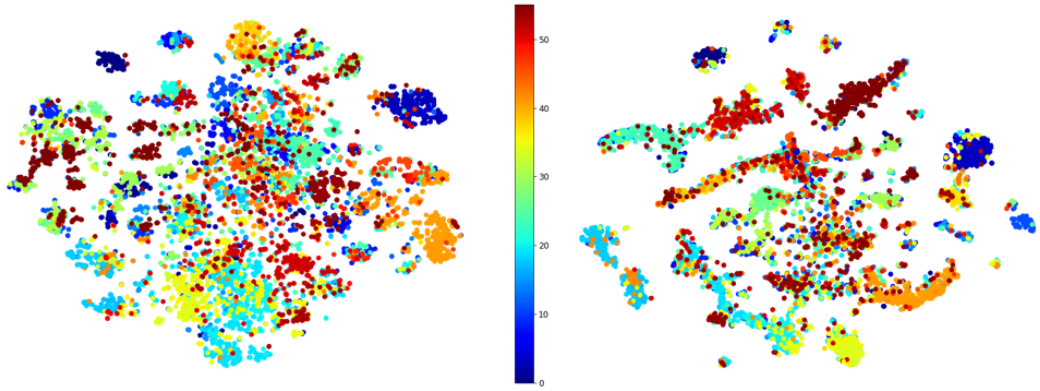

FIGURE 2 – Comparaison du modèle open-source paraphrase-multilingual-mpnet-base-v2 (à gauche) et d'un modèle spécialisé pour les données EDF Commerce (à droite) sur un jeu de données interne de retranscriptions d'un serveur vocal interactif (SVI) en langue naturelle. Aucune donnée du SVI n'a été utilisée pour la spécialisation. La couleur des points représente des classes attribuées manuellement (vérité terrain).

François Raynaud.

Références

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4171–4186, 2019.

[3] Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, and Anne-Laure Guénet. Désidentification de données texte produites dans un cadre de relation client. In *Actes de la 27eme conférence Traitement Automatique des Langues Naturelles (TALN) – démonstrations*, pages 10–13, 2020.

[4] Guillaume Dubuisson Duplessis, Sofiane Kerroua, Ludivine Kuznik, and Anne-Laure Guénet. Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume IV : Démonstrations*, pages 623–626, 2019.

[5] Maarten Grootendorst. Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*, 2022.

[6] Karthik S Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269. IEEE, 2019.

[7] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert : Unsupervised language model pre-training for french. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[8] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[9] Leland McInnes, John Healy, and Steve Astels. hdbscan : Hierarchical density based clustering. *J. Open Source Softw.*, 2(11) :205, 2017.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[12] Pierre Ratinaud and Pascal Marchand. Application de la méthode alceste à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "cablegate" avec iramuteq. *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles*, pages 835–844, 2012.

[13] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[14] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4) :333–389, 2009.