

TP2 : Extraction d'information Traitement automatique de corpus

Taleb Ahmed 000575310 <https://github.com/00atal/tac.git>

1. Introduction

Pour réaliser ce TP, j'ai choisi de travailler sur l'année 1939 du corpus CAMille. Cette année, marquée par le déclenchement de la Seconde Guerre mondiale, présente un intérêt particulier sur le plan historique et discursif. L'analyse des articles publiés à cette période permettra de mieux comprendre le traitement médiatique des événements majeurs ainsi que les tendances lexicales et émotionnelles propres à ce contexte.

2. Mots clés

	Terme	Poids	Interprétation
1	mère	0.0953	Mot très fréquent ou saillant, lié à un thème familial ou affectif.
2	flllo	0.0945	Probablement une faute de frappe ou une variante de « fille ».
3	jeune	0.0944	Évoque la jeunesse ; renforce le thème familial ou publicitaire (« jeune fille »).
4	Laruns	0.0895	Nom propre — probablement un lieu ou un nom de famille.
5	BRILLANTS	0.0887	Terme commercial, souvent associé aux bijoux.
6	BIJOUX	0.0881	Confirme le champ lexical du luxe et de la joaillerie.
7	INCASSABLE	0.0871	Terme technique ou publicitaire — suggère la solidité d'un produit.
8	INOXVDAm.F	0.0861	Expression typique d'un contexte commercial (inoxydable, version française abrégée).
9	Écrire	0.0821	Peut indiquer une mention d'annonce (« Écrire à... »), fréquente dans les petites annonces.
10	bon	0.0820	Mot courant, peu thématique, mais fréquent dans le langage publicitaire (« bon état », « bon prix »).

3. Stopwords

Après une première exploration, des erreurs d'OCR ont été détectées. Il a donc fallu effectuer un prétraitement pour corriger et homogénéiser les données textuelles.

Exemple Avant: '■ 2, et 3 JANyiÉR 1 9 •JJ iy-^ tJT» v'\.^- -..\. *;rV| y#>- **••*■' »

Exemple Aprés: JANYIÉR TJT EQOLE CIFTLLEBÉO ONNU VMTIR IESOIR FFÀYL



J'ai dû adapter et enrichir la liste de stopwords (notamment en ajoutant des mots comme et, ou, de) afin d'éliminer les termes sans valeur sémantique. Cette étape a permis d'obtenir une visualisation plus claire, centrée sur les mots réellement significatifs comme prix, maison ou place. Le nuage de mots met en évidence la forte présence de termes liés à la vie quotidienne (maison, heures, prix, place) et à la publicité (vente, commerce).

On y retrouve également des références géographiques comme Paris ou Anvers, suggérant une dimension urbaine et économique marquée. Ces résultats traduisent un corpus où dominent les annonces et les informations pratiques, typiques de la presse de 1939, ce qui étonne à la veille de la guerre.

4 Entités nommées

Les entités les plus mentionnées dans le corpus montrent une forte présence de **marques et organisations** (Ag, Rossel, Libby's, CEI) ainsi que de **personnes** (Jacques, Henri de Laruns, Clara, Arnaud). Les **lieux** évoqués sont surtout **Bruxelles, Tours et Rotterdam**, reflétant un contexte à la fois local et international. On note également la présence de termes liés à des **produits ou services** (Garantie, Compte chèque, Horlogerie, Bracelet), ce qui confirme l'importance du discours commercial et publicitaire.

Globalement, le corpus illustre un mélange de **vie quotidienne, activités commerciales et références individuelles**, offrant un aperçu de la société et de l'économie pré-guerre.

Si vous voulez, je peux aussi faire une **version ultra-courte** en 2 phrases maximum pour une lecture rapide.

	Top 10 Personnes	Occurrences	Top 10 Organisations	Occurrences	Top 10 Lieux	Occurrences
1	Ag	16	Téléphone	2	Bruxelles	5
2	Rossel	13	U	2	fr	3
3	Libby'	6	Garantie	1	Tours	3
4	Jacques	4	CEI	1	Brux	3
5	Ecrire Ag	3	Compte chèque	1	B	2
6	Henri de Laruns	3	Postal	1	A	2
7	Clara	3	Horlogerie A. BONNET	1	Ec	2
8	Arnaud	3	MONDE	1	occ	2
9	Juliette	2	ROTTERDAM	1	F	1
10	Maman	2	DPTIC	1	Bracelet	1

Le modèle **fr_core_news_md** détecte mal les entités dans ce type de texte : beaucoup de faux positifs et de confusions entre personnes, organisations et lieux. Il est peu robuste au bruit, aux fautes et aux données OCR. Mieux vaut utiliser **fr_core_news_lg** ou un **modèle NER spécialisé**, après nettoyage du texte.

5 Polarité et subjectivité

Pour cette étape, j'ai sélectionné dix phrases issues du corpus de 1939 et les ai analysées à l'aide du notebook s4_sentiment.ipynb. Afin de pouvoir délimiter correctement les phrases, je suis repartie du corpus non nettoyé, car certaines opérations de prétraitement (comme la suppression de ponctuation) rendaient la segmentation moins fiable. Les résultats montrent une majorité de phrases neutres ou objectives, avec quelques occurrences légèrement positives (ex. : messages publicitaires ou informatifs) et négatives (ex. : annonces de décès, ton grave). Globalement, la tonalité du corpus reste mesurée et factuelle, ce qui correspond au style journalistique de la presse de l'époque.

Text	Polarity Value	Polarity %	Polarity	Subjectivity Value	Subjectivity %	Subjectivity
Louise Rainer et Alan Marshall sont les protagonistes convaincus de ce film émouvant, mis en scène avec talent par Robert Sinclair.	0.267	27%	positive	0.283	28%	objective
Il est prudent pour un critique de cinéma de ne pas abuser du mot chef-d'œuvre.	-0.05	5%	negative	0.525	52%	subjective
Agence Rossel sous le n°19868V — PARQUETS CI RftS Rem.	0.0	0%	neutral	0.0	0%	objective
V* 1 " ^0.- SOLDÉ CS/365 GRAND CHOIX DE FLEURS pour corsages.	0.3	30%	positive	0.2	20%	objective
3 15 Ind,Cbimiques 8150 Ind.Chlm.p.l 420 Laeken 1800 La Métal.Chlm.p.	0.0	0%	neutral	0.0	0%	objective
cas où le pire viendrait à se produire, dès renoncer momentanément à nos concessions et à nos intérêts en Chine, et de supporter toutes les pertes que cela entraînerait plutôt que de faillir & notre devoir d'aider, et de soutenir les Chinois dans leur lutte atroce, mais qui n'a jamais été si riche en espérances, pour le droit et la liberté.	-0.002	0%	negative	0.45	45%	objective
Le comte Ciano a notamment déclaré: La pacte que nous venons de signer fixé sans équivoque la solidarité politique et militaire complète de l'Allemagne et de l'Italie..	0.09	9%	positive	0.05	5%	objective
L'avion Congo —'Belgique a quitté Fort-Archambadd à 12 h.	0.0	0%	neutral	0.0	0%	objective
Ces révélations auxquelles "il ne s'attendait pas et les menaces de Jeanne Sajotte l'auraient troublé à .	-0.1	10%	negative	0.8	80%	subjective
des Bouleaux, Waterm.20721 Porte Louise, app raeubié 5 p.	0.0	0%	neutral	0.0	0%	objective