Pour en finir avec les pré-traitements (au moins sur les données textuelles)

Gaël Lejeune

17 Octobre 2024

Sorbonne Université

Le dogme

Understanding Preprocessing

Preprocessing is a critical step in NLP that involves cleaning and preparing text data for analysis. It includes several tasks such as tokenization, removing stop words, stemming, lemmatization, and more. These tasks help in reducing the noise in the data, making it more manageable and meaningful for analysis.

Text preprocessing is an essential step in **natural language processing** (NLP) that involves cleaning and transforming unstructured text data to prepare it for analysis. It includes **tokenization**, **stemming**, lemmatization, stop-word removal, and part-of-speech tagging. In this article, we will introduce the basics of text preprocessing and provide **Python** code examples to illustrate how to implement these tasks using the **NLTK library**. By the end of the article, readers will better understand how to prepare text data for NLP tasks.

Machine Learning heavily relies on the quality of the data fed into it, and thus, data preprocessing plays a crucial role in ensuring the accuracy and efficiency of the model. In this article, we will discuss the main text preprocessing techniques used in NLP.

1. Text Cleaning

In this step, we will perform fundamental actions to clean the text. These actions involve transforming all the text to lowercase, eliminating characters that do not qualify as words or whitespace, as well as removing any numerical digits present.

I. Converting to lowercase

Here is a comprehensive list of common techn text preprocessing:

1. Text lowercasing
2. Tokenization
3. Stop-word removal
4. Handling Numerical values
5. Handling Special characters
6. Whitespace stripping
7. Lemmatization/Stemming

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?
- Lesquels sont documentés et justifiés

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?
- Lesquels sont documentés et justifiés

Les pré-traitements sont en fait des traitements à part entière puisqu'ils ont un impact non nul sur les opérations subséquentes réalisées [Millour, 2020]

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?
- Lesquels sont documentés et justifiés

Les pré-traitements sont en fait des traitements à part entière puisqu'ils ont un impact non nul sur les opérations subséquentes réalisées [Millour, 2020]

D'un point de vue de conception :

- Ils prennent du temps
- Focalisent-il l'attention sur les bons problèmes ?

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?
- Lesquels sont documentés et justifiés

Les pré-traitements sont en fait des traitements à part entière puisqu'ils ont un impact non nul sur les opérations subséquentes réalisées [Millour, 2020]

D'un point de vue de conception :

- Ils prennent du temps
- Focalisent-il l'attention sur les bons problèmes ?
- Améliorent-ils les résultats ?

Traitements ou pré-traitements

Quelle est la différence

- Les pré-traitements sont anodins (mais obligatoires ?)
- Les "traitements" sont plus nobles ?
- Lesquels sont documentés et justifiés

Les pré-traitements sont en fait des traitements à part entière puisqu'ils ont un impact non nul sur les opérations subséquentes réalisées [Millour, 2020]

D'un point de vue de conception :

- Ils prennent du temps
- Focalisent-il l'attention sur les bons problèmes ?
- Améliorent-ils les résultats ?

De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.
- Removing HTML tags/ URLs.

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.
- Removing HTML tags/ URLs.
- Removing punctuation.
- Removing stop words

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.
- Removing HTML tags/ URLs.
- Removing punctuation.
- Removing stop words
- Removing Emojies

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.
- Removing HTML tags/ URLs.
- Removing punctuation.
- Removing stop words
- Removing Emojies
- Tokenization
- Stemming
- Lemmatization

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

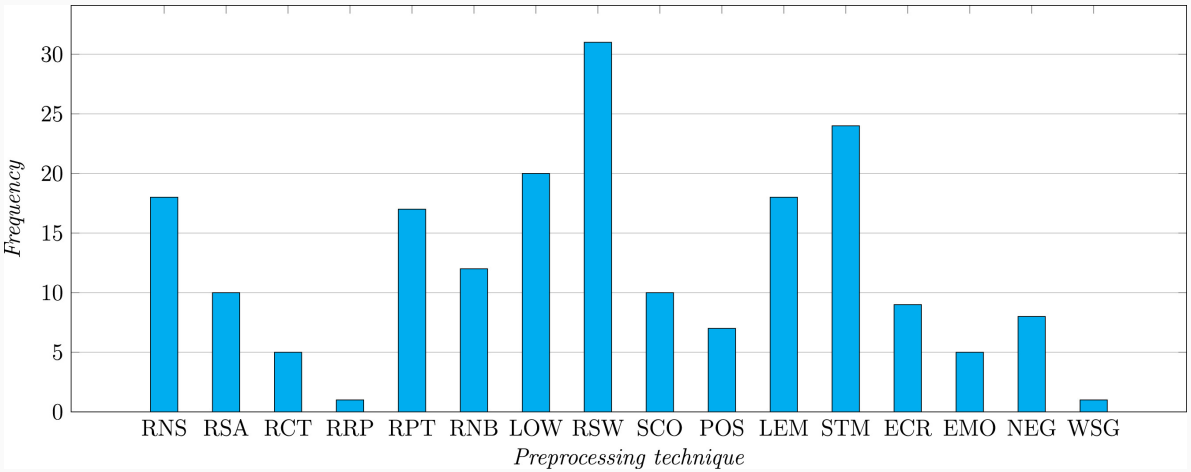
De quels pré-traitements parle-t-on ?

In the literature there is no convention adopted, and each work tests some preprocessing techniques rather than others. [1]

- Lowercase letters.
- Spelling Correction.
- Removing HTML tags/ URLs.
- Removing punctuation.
- Removing stop words
- Removing Emojies
- Tokenization
- Stemming
- Lemmatization

1. Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods . . . [Siino et al., 2024]

Etat des lieux plus précis (Siino et al.)



DON	Do Nothing
RNS	Replace Noise
RSA	Replace Slang/Abbreviations
RCT	Replace Contraction
RRP	Remove Repeated Punctuation
RPT	Removing Punctuation
RNB	Remove Numbers
LOW	Lowercasing
RSW	Remove Stop Words
SCO	Spelling Correction
POS	Part-of-Speech Tagging
LEM	Lemmatization
STM	Stemming
ECR	Remove Elongation
EMO	Emoticon Handling
NEG	Negation Handling
WSG	Word Segmentation (sometrendingtopic)

Polarité Critiques en (Siino et al.)

	IMDB					
Preprocessing	**RoBERTa**	**XLNet**	**ELECTRA**	**ANN**	**CNN**	**BiLSTM**
DON (D)	0.884 ± 0.00	0.885 ± 0.00	0.888 ± 0.00	0.835 ± 0.01	0.856 ± 0.00	0.847 ± 0.00
LOW (L)	0.877 ± 0.00	0.881 ± 0.01	**0.895 ± 0.04**	0.842 ± 0.01	**0.857 ± 0.00**	0.843 ± 0.01
RSW (R)	**0.885 ± 0.00**	**0.886 ± 0.00**	0.890 ± 0.07	0.840 ± 0.01	0.855 ± 0.00	0.843 ± 0.01
STM (S)	0.853 ± 0.00	0.852 ± 0.03	0.857 ± 0.05	0.834 ± 0.01	0.856 ± 0.00	0.837 ± 0.02
(L)→(R)	0.875 ± 0.04	0.878 ± 0.01	0.888 ± 0.01	0.840 ± 0.01	0.854 ± 0.00	0.844 ± 0.01
(L)→(S)	0.849 ± 0.00	0.847 ± 0.01	0.860 ± 0.03	**0.845 ± 0.00**	0.855 ± 0.00	0.845 ± 0.02
(R)→(L)	0.876 ± 0.04	0.874 ± 0.00	0.890 ± 0.01	0.844 ± 0.01	0.855 ± 0.00	0.847 ± 0.01
(R)→(S)	0.826 ± 0.02	0.823 ± 0.32	0.832 ± 0.02	0.839 ± 0.00	0.855 ± 0.00	0.844 ± 0.02
(S)→(L)	0.849 ± 0.00	0.845 ± 0.03	0.864 ± 0.01	0.839 ± 0.00	0.854 ± 0.00	0.840 ± 0.01
(S)→(R)	0.798 ± 0.07	0.817 ± 0.01	0.832 ± 0.01	0.843 ± 0.01	0.854 ± 0.00	0.843 ± 0.01
(L)→(S)→(R)	0.806 ± 0.04	0.782 ± 0.12	0.824 ± 0.01	0.837 ± 0.01	0.855 ± 0.00	0.839 ± 0.34
(L)→(R)→(S)	0.838 ± 0.34	0.820 ± 0.02	0.837 ± 0.04	0.842 ± 0.01	0.854 ± 0.00	0.845 ± 0.00
(S)→(L)→(R)	0.812 ± 0.01	0.645 ± 0.18	0.818 ± 0.02	0.840 ± 0.01	0.856 ± 0.00	0.845 ± 0.01
(S)→(R)→(L)	0.818 ± 0.02	0.820 ± 0.05	0.837 ± 0.01	0.843 ± 0.01	0.853 ± 0.00	0.839 ± 0.01
(R)→(L)→(S)	0.829 ± 0.03	0.837 ± 0.17	0.825 ± 0.05	0.838 ± 0.01	0.855 ± 0.00	**0.848 ± 0.01**
(R)→(S)→(L)	0.806 ± 0.03	0.822 ± 0.07	0.848 ± 0.01	0.838 ± 0.01	**0.857 ± 0.00**	0.838 ± 0.34

Figure 1 – Médiane de l'exactitude (*accuracy*) sur 5 runs + diffmax. Pour chaque modèle le meilleur résultat est en gras, le pire en rouge

Etat des lieux plus précis (en) (Siino et al.)

IMDB : Polarité Critiques, PCL : Langage Condescendant Presse

FNS : AA Fake News, 20N : Catégorisation Forums

	IMDB			PCL			FNS			20N		
Preprocessing	**NB**	**SVM**	**LR**	**NB**	**SVM**	**LR**	**NB**	**SVM**	**LR**	**NB**	**SVM**	**LR**
DON	0.767	**0.835**	0.798	0.726	**0.729**	**0.693**	0.685	0.630	0.640	0.040	**0.160**	**0.140**
LOW	0.771	0.831	0.801	**0.736**	0.696	0.668	0.695	0.665	0.650	0.040	0.140	0.100
RSW	0.787	0.831	**0.833**	0.719	0.651	0.686	0.705	**0.715**	0.660	0.020	0.100	0.060
STM	0.741	0.794	0.773	0.683	0.678	0.691	0.675	0.645	0.640	0.040	**0.160**	0.080
LOW → RSW	0.787	0.828	**0.833**	0.706	0.671	0.683	0.720	0.690	0.680	0.040	0.140	0.040
LOW → STM	0.725	0.803	0.770	0.678	0.668	0.688	0.700	0.665	0.615	0.040	0.120	0.100
RSW → LOW	**0.789**	**0.835**	0.820	0.721	0.663	0.691	**0.725**	0.690	0.675	0.040	0.120	0.020
RSW → STM	0.780	0.794	0.811	0.671	0.641	0.656	0.680	0.695	0.675	0.020	**0.160**	0.100
STM → LOW	0.725	0.803	0.800	0.678	0.668	0.673	0.700	0.665	0.635	0.040	0.120	0.060
STM → RSW	0.775	0.790	0.821	0.681	0.641	0.646	0.675	0.675	0.670	0.020	0.140	0.120
LOW → STM → RSW	0.750	0.799	0.820	0.678	0.623	0.648	0.695	0.680	0.645	0.040	0.140	0.080
LOW → RSW → STM	0.747	0.794	0.821	0.668	0.636	0.661	0.700	0.685	0.650	0.040	0.140	0.080
STM → LOW → RSW	0.749	0.797	0.814	0.678	0.623	0.661	0.690	0.675	0.645	0.040	0.140	0.080
STM → RSW → LOW	0.749	0.797	0.814	0.678	0.623	0.661	0.690	0.685	0.655	0.040	0.140	0.080
RSW → LOW → STM	0.757	0.797	0.807	0.673	0.623	0.678	0.720	0.670	0.655	0.040	0.140	0.120
RSW → STM	0.756	0.797	0.803	0.673	0.623	0.651	0.720	0.675	**0.685**	0.040	**0.160**	0.080

Langage Figuratif Tweets `fr` (Choi 2020)

	Logistic Regression		Decision Tree		MNB		KNN		Random Forest	
	Count	**Tfidf**	**Count**	**Tfidf**	**Count**	**Tfidf**	**Count**	**Tfidf**	**Count**	**Tfidf**
DON	**50.20**	**52.03**	**50.41**	**42.89**	**51.42**	**52.24**	**38.82**	**45.73**	**53.25**	**51.22**
RPT	50.41	52.64	48.37	44.72	50.81	51.63	38.21	45.53	**53.05**	**52.64**
RSW	**52.24**	53.86	45.93	44.11	51.22	52.24	37.40	44.31	50.00	50.20
ACC	49.59	52.64	49.39	43.29	51.02	52.03	35.16	45.53	52.44	52.03
URL	**47.56**	**47.36**	**39.43**	**39.43**	**50.20**	**50.61**	**34.35**	**41.46**	**45.53**	**44.51**
LEM	50.20	**54.07**	**49.19**	44.72	**52.24**	**53.25**	**39.02**	45.53	50.41	51.63
STM	51.63	53.86	48.37	**45.93**	52.03	52.44	38.41	**46.75**	52.44	51.42

Table 1 – Exactitude moyenne (En bleu, le meilleur résultat, en rouge le plus faible résultat pour chaque classifieur)

Langage Figuratif Tweets `fr` (Choi 2020)

Classifieur	Count Vectorizer	Macro f-mesure	Tfidf Vectorizer	Macro f-mesure
Logistic Regression	LEM, RSW	53.53	LEM, RSW, RAC	54.35
Decision Tree	RPT, accents, RAC RSW	49.59	RAC, RPT	48.58
MNB	LEM, RSW, RAC	54.59	LEM, RSW, RAC	**55.89**
KNN	RAC, RSW, RPT	**38.20**	RAC, RSW	47.35
Random Forest	LEM, RSW, accents, RAC	51.38	LEM, RSW, accents, RAC	53.25

Table 2 – Meilleurs résultats en macro f-mesure (En bleu, le meilleur résultat, en rouge le plus faible résultat. Meilleur résultat du DEFT2017 65%)

Les combinaisons de pré-traitement

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis [Symeonidis et al., 2018]

Les combinaisons de pré-traitement

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis [Symeonidis et al., 2018]

Influence des pré-traitements sur la classification de textes - Application à la classification de tweets selon leur polarité (Heesoo Choi 2020, mémoire de master)

Les combinaisons de pré-traitement

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis [Symeonidis et al., 2018]

Influence des pré-traitements sur la classification de textes - Application à la classification de tweets selon leur polarité (Heesoo Choi 2020, mémoire de master)

Qu'apprend-on

- Deux pré-traitements peuvent mal interagir
- Le gain de performance est asymptotique

Les combinaisons de pré-traitement

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis [Symeonidis et al., 2018]

Influence des pré-traitements sur la classification de textes - Application à la classification de tweets selon leur polarité (Heesoo Choi 2020, mémoire de master)

Qu'apprend-on

- Deux pré-traitements peuvent mal interagir
- Le gain de performance est asymptotique
- Le cocktail ne fonctionne pas quel que soit :
  - la tâche
  - le type de textes
  - le classifieur
  - le modèle de langue

A un grain supérieur : Des donnés non structurées ?

Non-structuré à computationnelement opaque[de Busser and Moens, 2006]

- La ponctuation
- Les tabulations
- Les marques de fin de ligne

A un grain supérieur : Des donnés non structurées ?

Non-structuré à computationnelement opaque[de Busser and Moens, 2006]

- La ponctuation
- Les tabulations
- Les marques de fin de ligne
- Le statut de la structure (Xml par exemple)

En conclusion

- **Se méfier** des idées reçues
  - `vin et fromage`

En conclusion

- **Se méfier** des idées reçues
  - `vin et fromage`
- **Évaluer** réellement ce qu'on fait
  - `faire goûter son plat`

En conclusion

- **Se méfier** des idées reçues
  - `vin et fromage`
- **Évaluer** réellement ce qu'on fait
  - `faire goûter son plat`
- **pré-traiter ssi** strictement nécessaire
  - `sel et sucre`

En conclusion

- **Se méfier** des idées reçues
  - `vin et fromage`
- **Évaluer** réellement ce qu'on fait
  - `faire goûter son plat`
- **pré-traiter ssi** strictement nécessaire
  - `sel et sucre`
- **ne pas écraser** les observables
  - `légumes et wok`

En conclusion

- **Se méfier** des idées reçues
  - `vin et fromage`
- **Évaluer** réellement ce qu'on fait
  - `faire goûter son plat`
- **pré-traiter ssi** strictement nécessaire
  - `sel et sucre`
- **ne pas écraser** les observables
  - `légumes et wok`
- **conserver** les indices laissés par l'émetteur du texte
  - `ingrédients et température`

de Busser, R. and Moens, M.-F. (2006).

Information extraction and information technology, pages 1–22.

Springer.

Millour, A. (2020).

Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées.

PhD thesis, Sorbonne Université, France.

Siino, M., Tinnirello, I., and La Cascia, M. (2024).

Is text preprocessing still worth the time ? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers.

Information Systems, 121 :102342.

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018).

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis.

Expert Systems with Applications, 110 :298–310.