



# Deep Learning, traitement de langues naturelle et Text Mining

Damien Douteaux

Vincent Hocquemiller

**Louis Redonnet** 

# **SOMMAIRE**

Contexte et objectifs Résultats Bilan Conclusion

- Contexte et objectifs
- Livrables
- Évaluation des résultats
- Conclusion

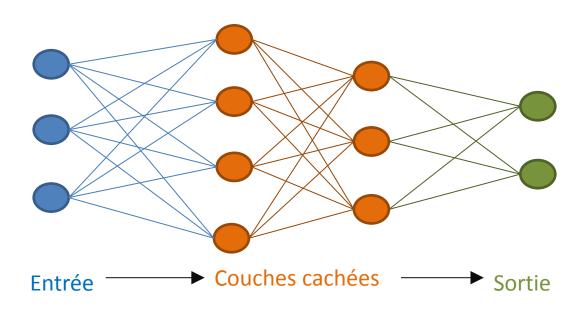


# LE DEEP LEARNING

Contexte et
objectifs
Résultats
Bilan
Conclusion

Deep Learning et traitement de langues naturelles et Text Mining.

- Méthode récente (1980 2000).
- Couches de neurones (unités de traitements).





### LE DEEP LEARNING

# Contexte et objectifs Résultats Bilan Conclusion

Deep Learning et traitement de langues naturelles et Text Mining.

- Méthode récente (1980 2000).
- Application pour :
  - Reconnaissance faciale/vocale
  - Vision par ordinateur
  - Intelligence artificielle
  - Traitement automatisé du langage

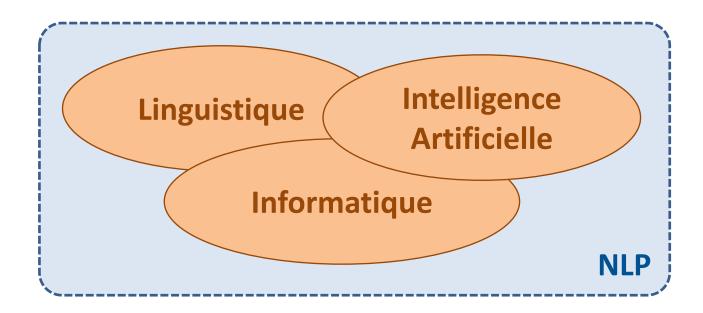


## TRAITEMENT DE LANGUES NATURELLES

Contexte et objectifs
Résultats
Bilan
Conclusion

Deep Learning et traitement de langues naturelles et Text Mining.

Le traitement (automatique) des langues naturelles est l'exploitation du langage humain par les outils informatiques.





### **TEXT MINING**

### Contexte et objectifs Résultats Bilan

Deep Learning et traitement de langues naturelles et Text Mining.

- Recenser, structurer des données textuelles.
- Approche globale et grossière du texte (sans s'attarder sur le sens).
- Historiquement différents modèles :
  - Bag of words;
  - N-Gram.
- Applications dans de nombreux domaines :
  - Page ranking;
  - Filtrage des communications ;
  - Intelligence économique (détection de sujets clés).



# **OBJECTIFS**

#### Contexte et objectifs Résultats Bilan Conclusion

- Fournir un état de l'art précis sur le contexte.
- Constituer une base de données de volume adapté au Deep Learning.
- Implémentation d'un réseau de neurones sur un cas dégagé par l'état de l'art.



# ÉTAT DE L'ART (1/2)

contexte et objectifs
Résultats
Bilan
Conclusion

#### **RECONNAISSANCE D'AUTEUR**

- Déterminer l'auteur d'un texte.
- Utilisation pour la détection de plagiat/similarité.

#### **TRADUCTION**

- Traduction de texte à l'échelle d'une phrase.
- Idée d'essai d'un même réseau pour plusieurs langues.

#### **INFÉRENCE**

- Étude de relations logiques entre les phrases d'un corpus.
- Recherche de contradictions ou préservation de la logique.

#### **QUESTIONS-RÉPONSES**

- Formuler des réponses automatiques à des questions simples.
- Utilisation pour des chatbox en ligne.

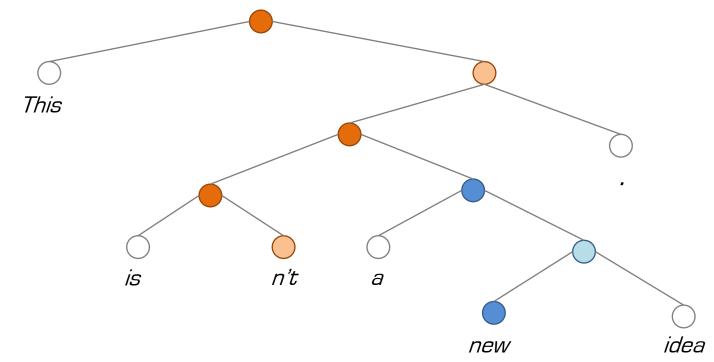


# ÉTAT DE L'ART (2/2)

Contexte el objectifs Résultats Bilan

#### **ANALYSE DE SENTIMENTS**

- Classifier les phrases par sentiment ou essayer de les prédire.
- Différentes structures de données et approches possibles.
- Utilisation en classification de mail, étude d'opinions,...



→ Sujet retenu pour les essais de la suite du projet.



# BASES DE DONNÉES MISES EN AVANT

Contexte et objectifs
Résultats
Bilan
Conclusion

	Nom	Quantité de données	Origine
	Large Movie Review Dataset	25000 x 2	Stanford
	Rottent Tomatoes Dataset	215 000	Kaggle
¥	Twitter Sentiment Corpus	5500	Niek Sanders
	Twitter Sentiment Analysis Corpus	1 578 627	?
Q	Sentiment Analyses Dataset	9645	Stanford
	UMICH S1650	40 000	Kaggle
<u>a</u> ,	Amazon reviews	> 6 millions <sup>1</sup>	Julian Mc Auley

<sup>1:</sup> la base complète (sur demande) fait 142,8 millions de critiques...(20 Gb).



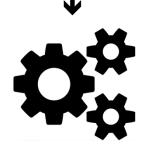
#### **Projet option**

# TRAITEMENT DES BASES DE DONNÉES

Contexte et objectifs
Résultats
Bilan
Conclusion

```
25 10.txt
                                                         Fichier TXT
                                                                           2 Ko
                                                                                    (4 (4 (4 (3 (2 Emerges) (3 (2 as) (3 (2 someths
"reviewerID": "A2IBPI20UZIR0U",
                                                                                    (3 (2 that) (3 (3 (2 's) (4 (3 (3 (2 so) (4 hor
                                          26 10.txt
                                                         Fichier TXT
                                                                           1 Ko
"asin": "1384719342",
                                                                                    (2 it) (2 (1 (2 does) (2 n't)) (2 (2 feel) (2
"reviewerName": "cassandra tu \"Yeah,
                                          27 9.txt
                                                         Fichier TXT
                                                                           1 Ko
                                                                                    (2 (2 (2 The) (2 film)) (3 (3 (3 provides)
"helpful": [0, 0],
                                          28 9.txt
                                                         Fichier TXT
                                                                           2 Ko
                                                                                    the) (2 (2 neurotic) (2 mindset))) (3 (2 of) (
"reviewText": "Not much to write about
                                                                                    (4 (2 who) (4 (2 have) (4 (2 reached) (4 (4 (2
"overall": 5.0,
                                          29_10.txt
                                                         Fichier TXT
                                                                           1 Ko
"summary": "good",
                                                                                    game))))))))))))) (2 .)))
                                          30 9.txt
                                                         Fichier TXT
                                                                           2 Ko
                                                                                    (4 (4 (2 Offers) (3 (3 (2 that) (3 (3 rare) (2
"unixReviewTime": 1393545600,
                                          31_10.txt
                                                         Fichier TXT
                                                                           3 Ko
"reviewTime": "02 28, 2014"
                                                                                    education))))) (2 .))
                                          32 10.txt
                                                         Fichier TXT
                                                                          3 Ko
                                                                                    (3 (2 Perhaps) (4 (2 (1 (1 no) (2 picture))
```

Parser C++



Export en TSV



Comment RatingDegrad Rating TextType BDDName
"Not much to write about here, but it does exactly what it's supposed to. filters out the pop sounds. now
my recordings are much more crisp. it is one of the lowest prices pop filters on amazon so might as well
buy it, they honestly work the same despite their pricing," 1 5 paragraph Amazon\_Musical\_Instruments

- Format uniforme entre les bases.
- Réduction de la taille.

 Pré-calcul sur la répartition des données.



# REPRÉSENTATION DES MOTS

Contexte et objectifs Résultats Bilan Conclusion

Le Deep Learning impose une représentation Sparse, Creuse ou Dense.

Le chat est couché 
$$\longrightarrow \begin{tabular}{ll} [le] & & & (1, 0, 0, 0) \\ [chat] & & (0, 1, 0, 0) \\ [est] & & (0, 0, 1, 0) \\ [couché] & & (0, 0, 0, 1) \\ \hline \end{tabular}$$

Longueur fixe (0.23, -1.40, ..., 2.98) (0.02, 2.03, ..., -0.99) (-0.62, 0.56, ..., 0.67) (2.10, 0.04, ..., 0.868) (2.2, 0.03, ..., 3.02) (0.08, 0.63, ..., 0.01) (-1.36, 1.22, ..., 0.67) (2.78, 0.40, ..., 0.57)

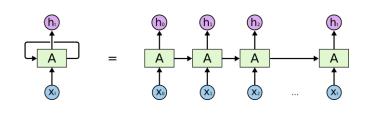


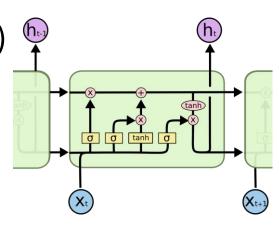
# MÉTHODES D'APPRENTISSAGE POSSIBLES

Contexte et objectifs Résultats Bilan Conclusion

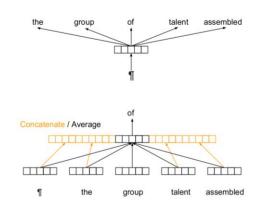
Trois façons de représenter une phrase :

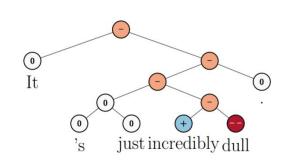
Une suite ordonnée de mots (LSTM)





- Un vecteur à apprendre.
- Un arbre descripteur.



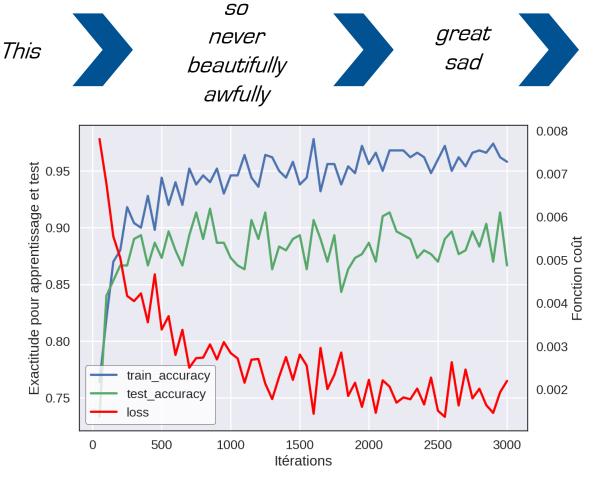




# UNE PREMIÈRE APPROCHE

Contexte et objectifs
Résultats
Bilan
Conclusion

#### **PROBLÈME « JOUET »**





### **BILAN DU PROJET**

Contexte et objectifs Résultats Bilan Conclusion

- ✓ État de l'art complet sur les applications du Deep Learning au NLP.
- ✓ Récupération d'une base de données adaptée au Deep Learning et prétraitement appliqué.
- ✓ Un *Toy Problem* qui fonctionne, et montre la validité des LSTM.
- Un début d'application sur nos données.

→ Les objectifs initiaux ont été réalisés.



# **CONCLUSION**

Contexte et objectifs Résultats Bilan Conclusion

#### **ENSEIGNEMENTS**

- Une occasion d'utiliser les librairies de Deep Learning de Python.
- Un projet enrichissant vis-à-vis des applications abordées.

#### **PERSPECTIVES**

- Ne pas s'enfermer que dans *tensorflow* et tester d'autres solutions.
- Pousser plus loin l'application sur les BDD traitées.



# **QUESTIONS**

Contexte et objectifs Résultats Bilan Conclusion

# Merci pour votre attention Et place aux question!

https://github.com/DDouteaux/Projet\_option\_info

