



## Deep Learning, traitement de langues naturelle et Text Mining

Damien Douteaux

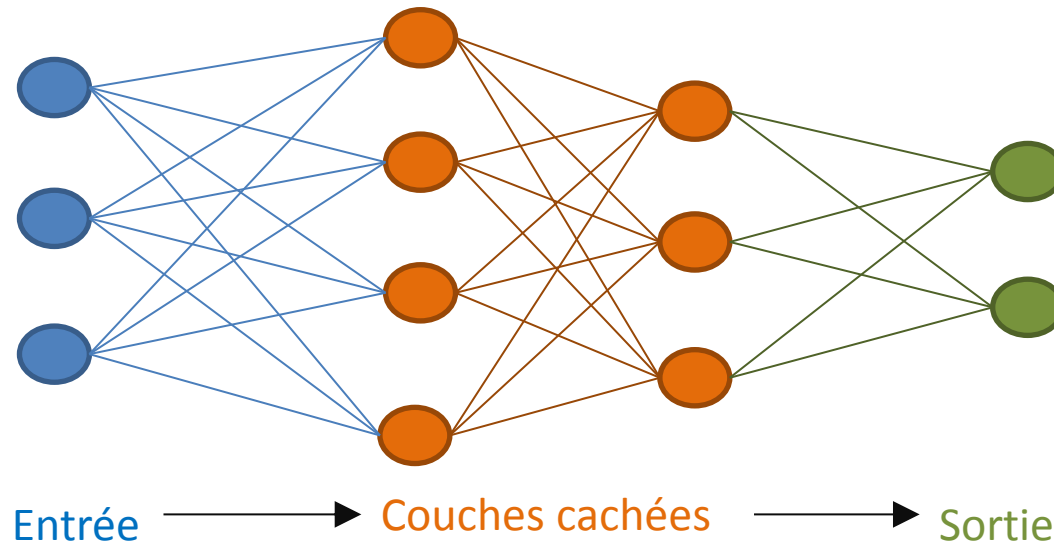
Vincent Hocquemiller

Louis Redonnet

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

**Deep Learning** et traitement de langues naturelles et Text Mining.

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



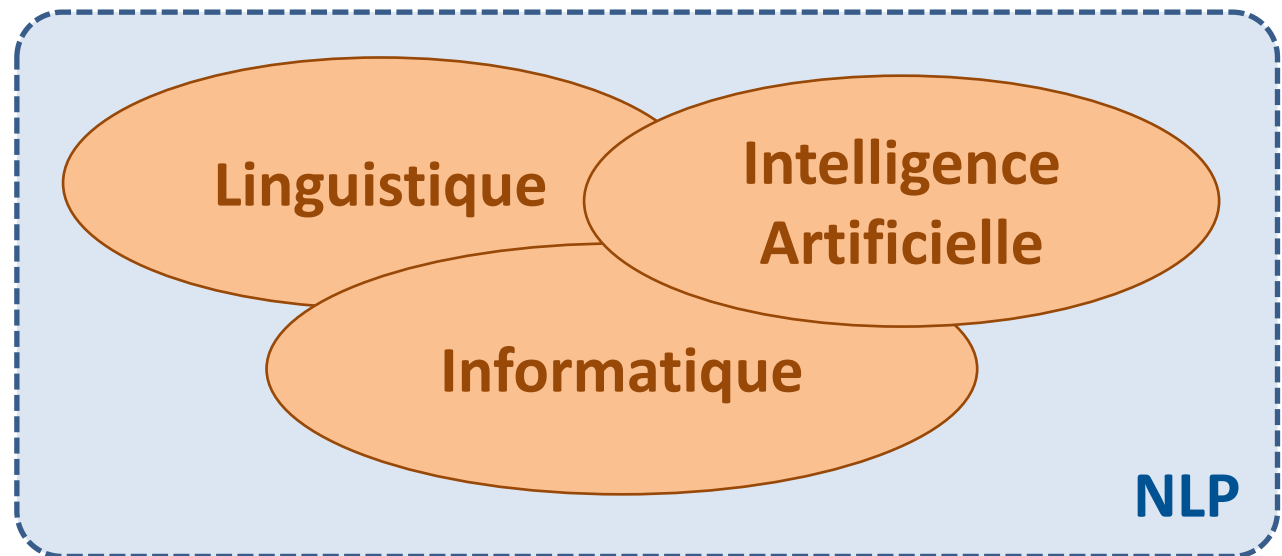
**Deep Learning** et traitement de langues naturelles et Text Mining.

---

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

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.



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).

- 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.

## 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

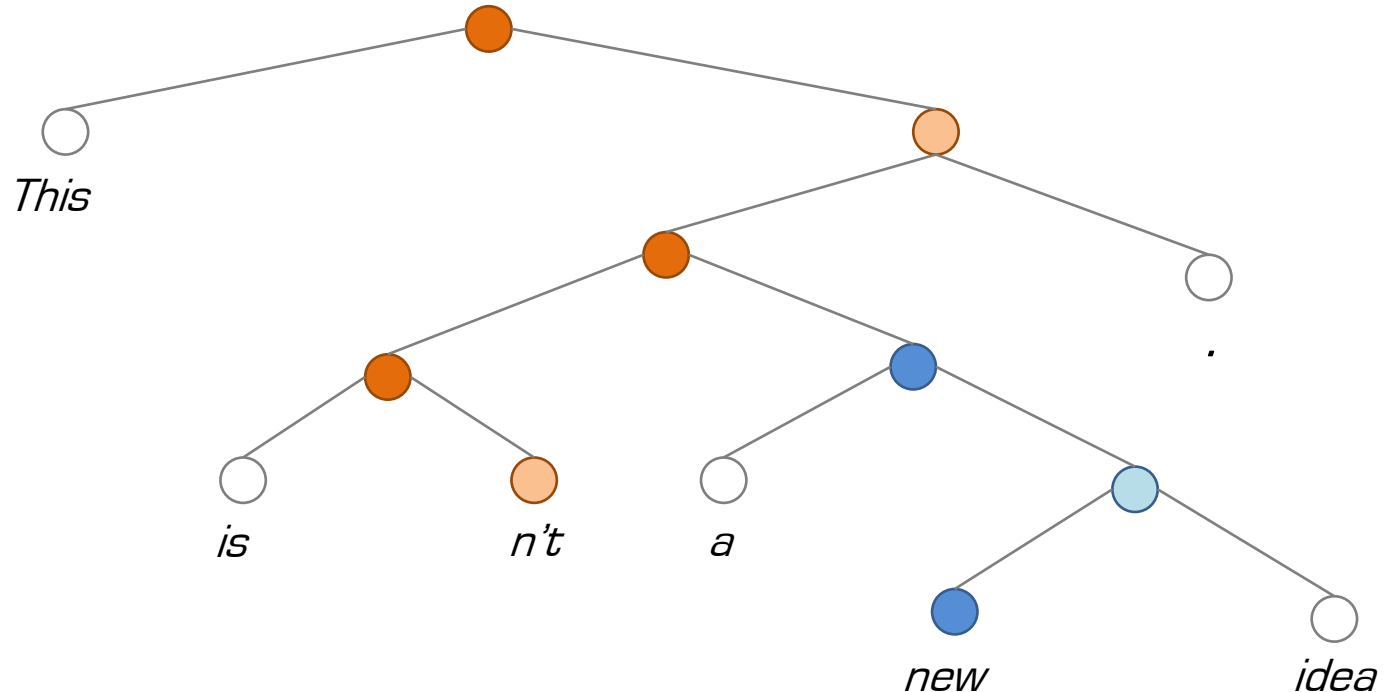
---

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







## ANALYSE DE SENTIMENTS

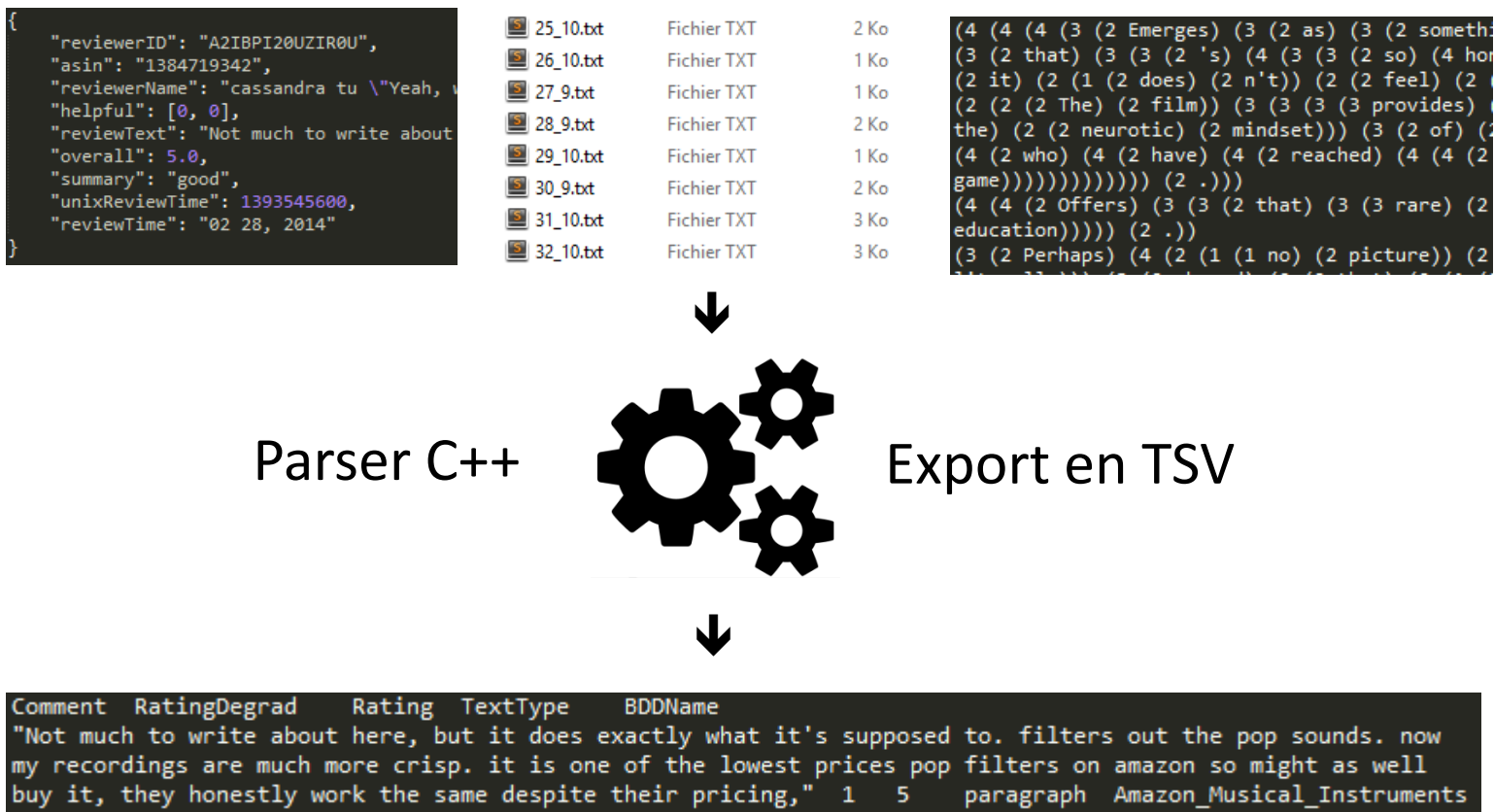
- Classifier les phrases par sentiment ou essayer de les prédire.
- Différentes structures de données et approches possibles.
- Utilisation en classification de mail, étude d'opinions,...



➔ **Sujet retenu pour les essais de la suite du projet.**

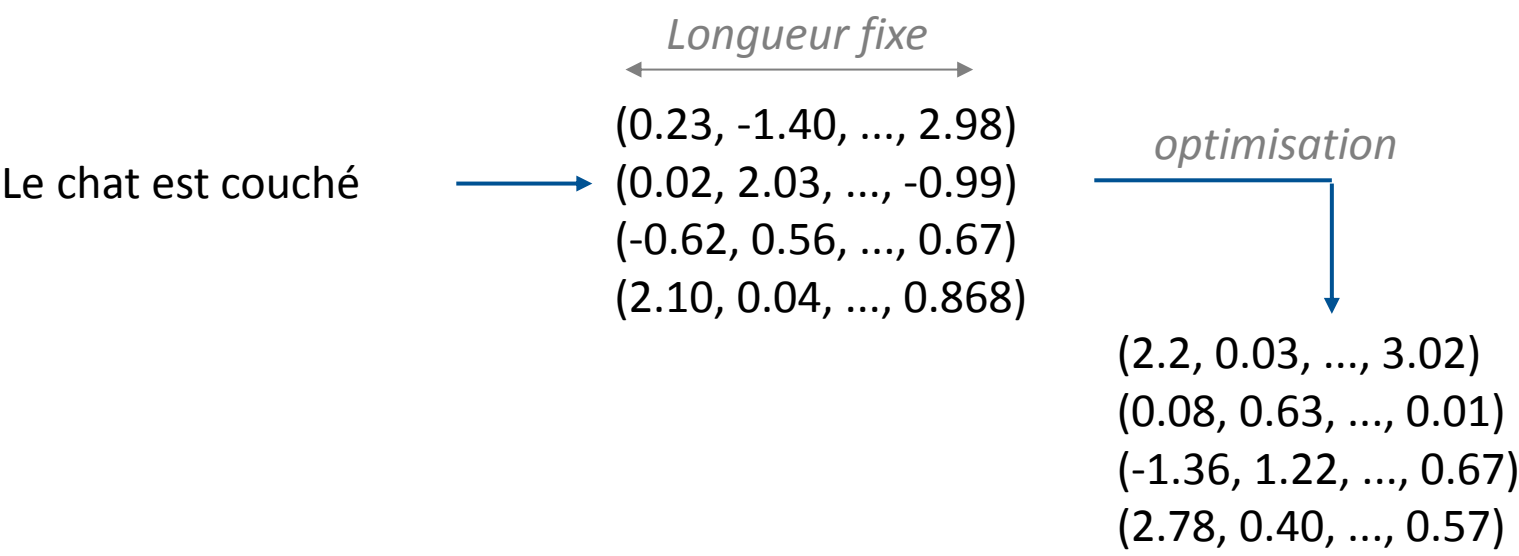
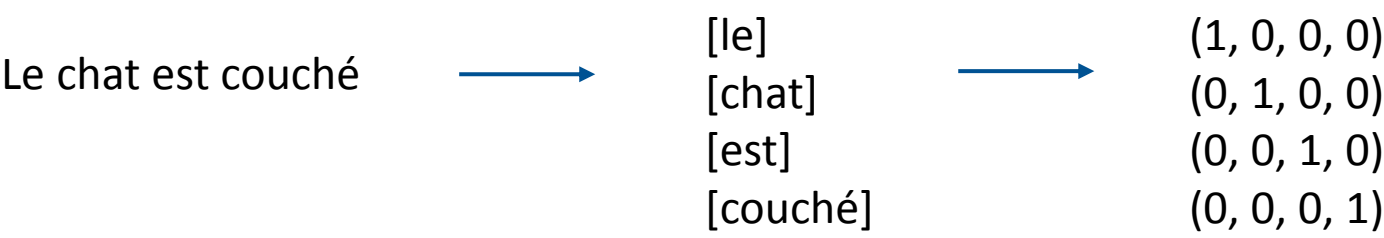
	Nom	Quantité de données	Origine
	Large Movie Review Dataset	25000 x 2	Stanford
	Rottent Tomatoes Dataset	215 000	Kaggle
	Twitter Sentiment Corpus	5500	Niek Sanders
	Twitter Sentiment Analysis Corpus	1 578 627	?
	Sentiment Analyses Dataset	9645	Stanford
	UMICH S1650	40 000	Kaggle
	Amazon reviews	> 6 millions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

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



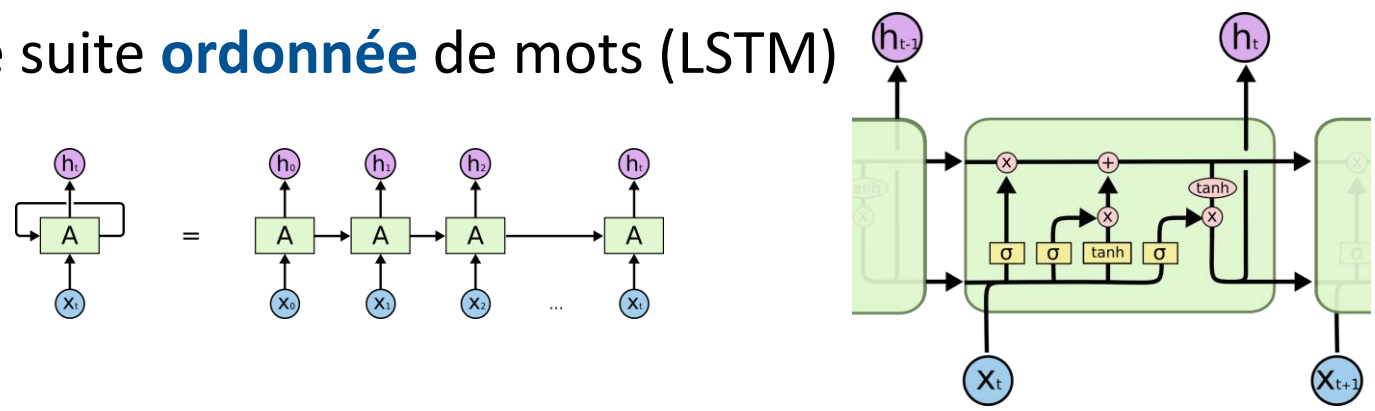
- Format uniforme entre les bases.
- Réduction de la taille.
- Pré-calcul sur la répartition des données.

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

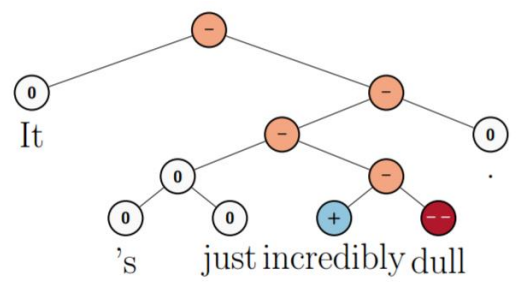
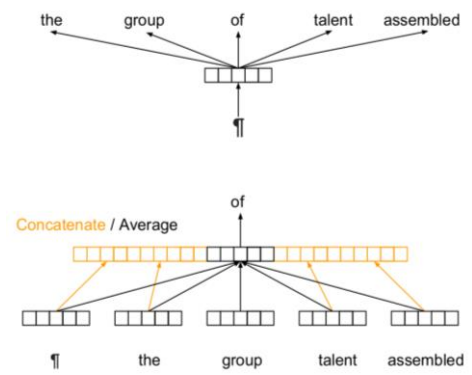


Trois façons de représenter une phrase :

- Une suite **ordonnée** de mots (LSTM)

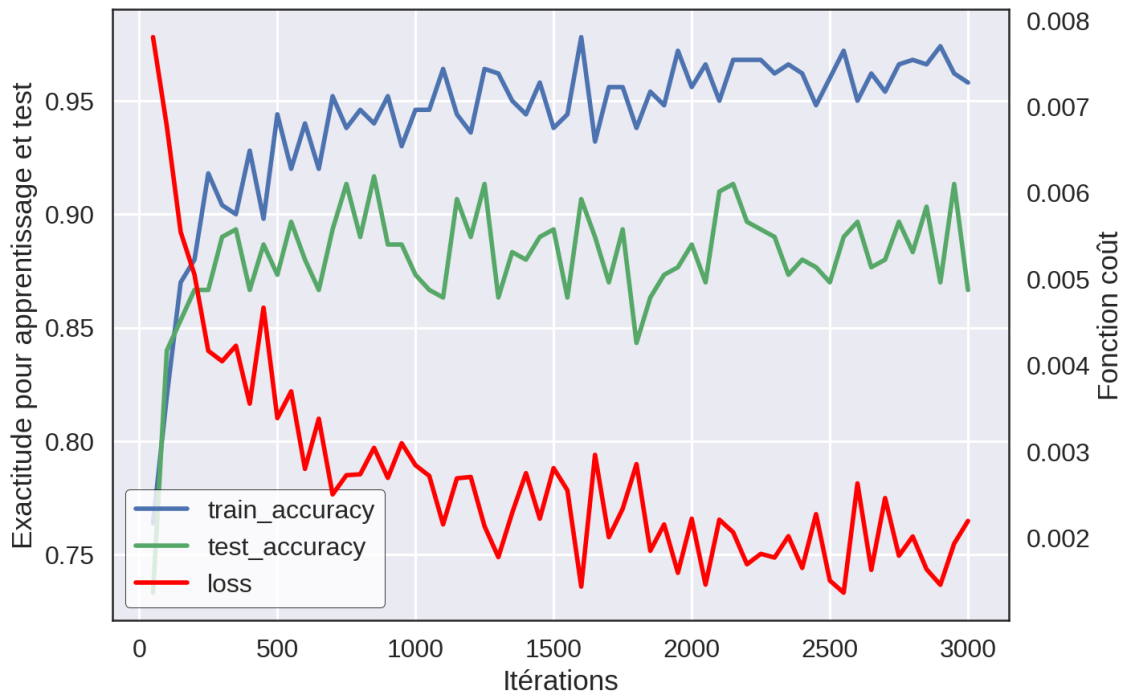


- Un vecteur à apprendre.
- Un arbre descripteur.



# UNE PREMIÈRE APPROCHE

## PROBLÈME « JOUET »



- ✓ 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.**

## 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.



**Merci pour votre attention**  
**Et place aux question!**



[https://github.com/DDouteaux/Projet\\_option\\_info](https://github.com/DDouteaux/Projet_option_info)