



Deep Learning, traitement de langues naturelles et text mining

Damien DOUTEAUX

Vincent HOCQUEMILLER

Louis REDONNET

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Sommaire

Projets envisagés

Reconnaissance d'auteur

Inférence

Questions et réponses

Traduction

Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Thème principal

- ⊙ Reconnaître l'auteur d'un texte
- ⊙ Estimer le genre d'un auteur

Plus-value possible

- ⊙ Peu (pas ?) d'essais en Deep Learning.
- ⊙ Peu d'essais en français.

Applications possibles

- ⊙ Vérification de classification
- ⊙ Analyse d'identité
- ⊙ Validation auteur (cf. plagiat et/ou similarité)

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Bases de données

Extraitre de petites entités (paragraphes, tweets,...) via :

- ⊙ *Livres* Projet Gutenberg, Wikibooks,...
- ⊙ *Twitter* Récupération par API.

Complexité

- ⊙ Problème ouvert
- ⊙ Peu de sources

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Thème principal

Extraire des relations entre phrases.

Plus-value possible

- ⊙ Peu d'essais en français.
- ⊙ Comparaison entre langues.

Applications possibles

- ⊙ Mise en avant de contradiction.
- ⊙ Comparaison d'informations

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Bases de données

- ⊙ Réutilisation de romans
- ⊙ Articles de presse
- ⊙ Stanford Natural Language Inference Corpus

Complexité

- ⊙ Dépend de l'axe retenu
- ⊙ Modélisation
- ⊙ Grande variété de possibilités.

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Thème principal

Réponses automatiques à des questions simples.

Plus-value possible

- ⊙ Poursuivre les travaux en apprentissage par renforcement

Applications possibles

- ⊙ Chatbox
- ⊙ Proposition de services

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Bases de données

- ⊙ Dialogues de films
- ⊙ Autres difficiles à trouver (Facebook,...)

Complexité

- ⊙ Utilisation de LSTM (technologie mature)
- ⊙ Difficulté à trouver des BDDs
- ⊙ Difficulté de modélisation
- ⊙ Manque de métrique pour évaluer

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Thème principal

Traduction à l'échelle d'une phrase par Deep Learning.

Plus-value possible

- ⊙ Un même réseau pour plusieurs langues.
- ⊙ Trouver un autre cas d'usage ?

Applications possibles

Amélioration de résultats en traduction.

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Bases de données

Tout ce qui est traduisible en plusieurs langues sur le web (cf. Linguee) :

- ⊙ Textes UE
- ⊙ Documents légaux
- ⊙ Brevets
- ⊙ Site traduit

Complexité

- ⊙ Dépend de la précision
- ⊙ Données présentes et technologies (LSTM) matures

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Thème principal

- ⊙ Classifier des phrases par sentiment.
- ⊙ Prédire un sentiment (binaire ou avec une échelle).

Plus-value possible

- ⊙ Combiner plusieurs approches.
- ⊙ Problématique « originale ».

Applications possibles

- ⊙ Prendre le pouls de la *twittosphère*.
- ⊙ Classification de mails (injurieux ou non).

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

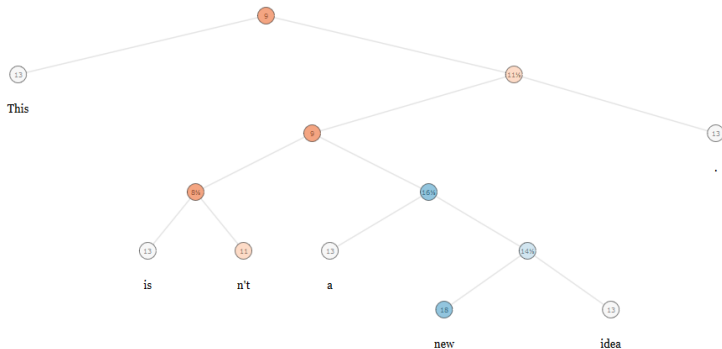
Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion



Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Bases de données

- ⊙ *IMDB, Amazon,...* Retours critiques de clients.
- ⊙ *Arbres syntaxiques* Stanford, 10000 arbres mais difficile à étendre (Amazon Turk).

Complexité

- ⊙ Dépend de la précision et des outils.
- ⊙ Base de données à agréger.
- ⊙ Grande variété de possibilités.

COMPARAISON DES SUJETS

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Sujet	Plus-value	BDD	Complexité
<i>Auteur</i>	◆	+	◆
<i>Inférence</i>	◆	+	—
<i>Q&A</i>	◆	◆	—
<i>Traduction</i>	—	+	—
<i>Sentiments</i>	+	+	◆

On retient le sujet d'analyse de sentiments.

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Récupération de diverses bases de données :

- ⦿ Des contes de fée
- ⦿ Des conversations de films
- ⦿ Du texte de Wikipédia
- ⦿ Des liens vers des BDDs

Des bases intéressantes, mais peu liées à nos objectifs.

DES BASES DE DONNÉES D'AVIS

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments







Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Nom		Type	Quantité	Origine
	Large Movie Review Dataset	Cinéma	25000 × 2	Stanford
	Rotten Tomatoes Dataset	Cinéma	1,6Mb	Kaggle
	Twitter Sentiment Corpus	Tweets	5500	Niek Sanders
	Twitter Sentiment Analysis Corpus	Tweets	1578627	?
	Sentiment Analyses Dataset	Divers	9645	Stanford
	UMICH S1650	Blogs	40000	Kaggle

Nécessité d'agglomérer les bases entre elles.

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

- ⦿ Utilisation de `tensorflow`
- ⦿ Peu de pistes sur l'architecture
- ⦿ Calcul sur les serveurs du LIRIS

UNE SEMAINE DE DÉCALLAGE

Sommaire

Projets envisagés

Reconnaissance d'auteur
Inférence
Questions et réponses
Traduction
Analyse de sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

Principales tâches	Durée (semaine)	Intervenants	Janvier					Février				Mars			
			S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
État de l'art	4	Vincent													
Recherche des sources	2	Vincent													
Examen documents	2	Vincent													
Synthèse	2	Vincent													
Déterminer les objectifs	1	Vincent													
Recherches complémentaires	1	Vincent													
Récupération BDD	4	Vincent													
Prise de contact PAR	2	Vincent													
Analyse des données	3	Vincent													
Nettoyage complémentaire	2	Vincent													
Récupération d'un réseau	4	Louis													
Recherche des sources	2	Louis													
Analyse du réseau	3	Louis													
Étude des méthodes pour l'utiliser	3	Louis													
Analyse des résultats	3	Damien													
Tests avec X-validation	2	Damien													
Mise en perspective avec littérature	2	Damien													
Pertinence méthode de travail	2	Damien													
Itération sur le modèle	1	Damien													
Administratif	7	Damien													
Préparation premier reporting	1	Damien													
Premier reporting	1	Damien													
Préparation second reporting	2	Damien													
Second reporting	1	Damien													
Préparation troisième reporting	2	Damien													
Soutenance finale	1	Damien													
Préparation livrable	4	Damien													
Réunions commanditaires	11	Damien													
Réunion de lancement	1	Damien													
Réunions intermédiaires	10	Damien													

Sommaire

Projets envisagés

Reconnaissance
d'auteur
Inférence
Questions et
réponses
Traduction
Analyse de
sentiments

Choix du sujet

Base de données

Réseau

Gestion de projet

Conclusion

- ⦿ Quasi-respect du calendrier.
- ⦿ Le sujet final est fixé.
- ⦿ Des bases de données repérées et en cours d'étude.
- ⦿ Un début de réflexion sur le réseau de neurones.

Sommaire

Projets
envisagés

Reconnaissance
d'auteur

Inférence

Questions et
réponses

Traduction

Analyse de
sentiments

Choix du
sujet

Base de
données

Réseau

Gestion de
projet

Conclusion

Merci pour votre attention
Et place aux questions!