Deep Learning, Text Mini
Traitement du Langage Na

Rapport de projet d'option informatique

Damien Douteaux, Vincent Hocquemiller et Louis Redonnet

Mars
30
2017

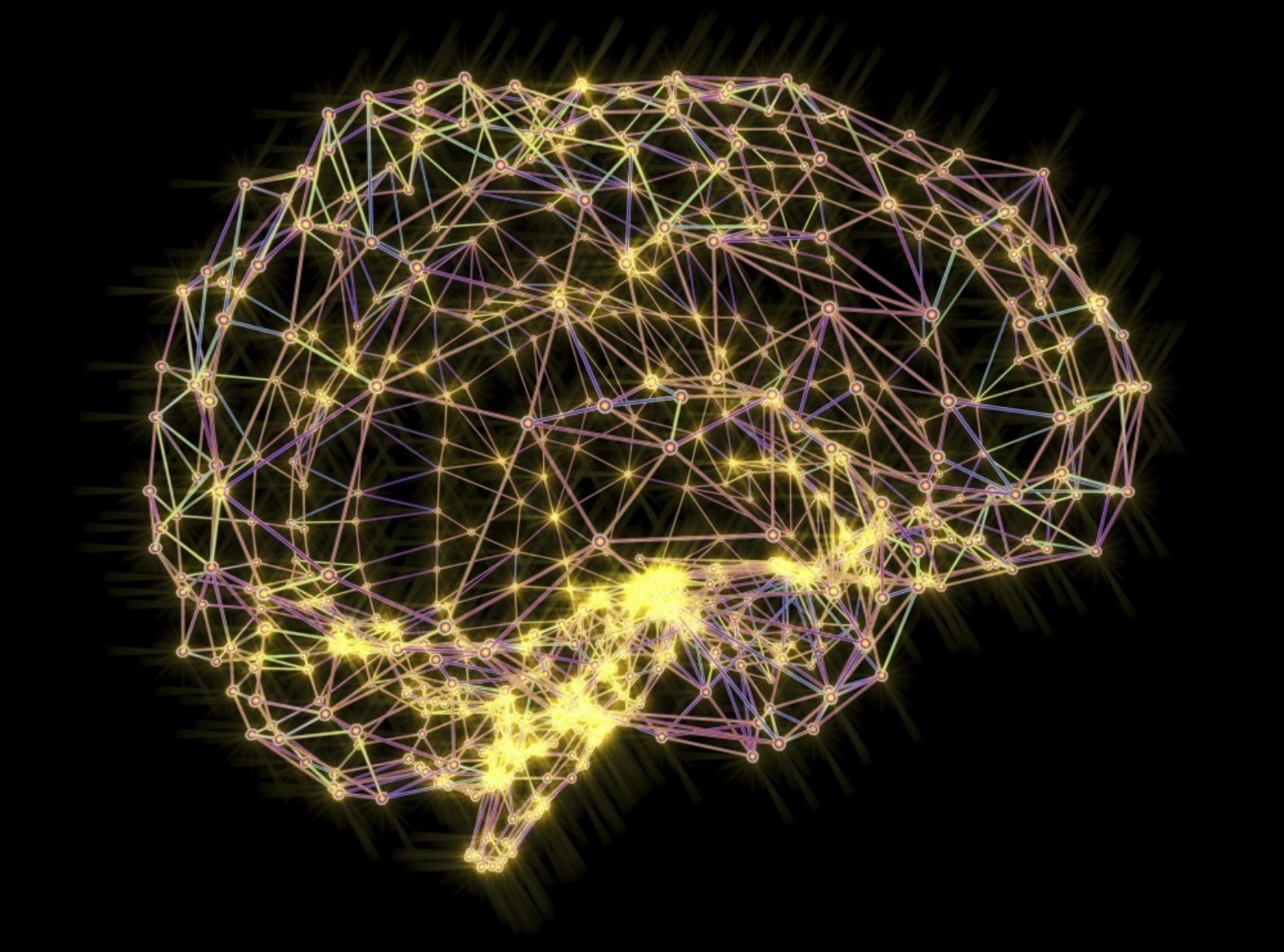

Table des matières

Introduction

1 • État de l'art

1.1. Reconnaissance d'auteur

1.2. Inférence

1.3. Questions et réponses

1.4. Traduction

1.5. Analyse de sentiments

1.6. Choix d'un sujet de travail

2 • Base de données

2.1. Rencontre avec un PAr

Dans notre recherche de base de données, la première étape proposée par nos commanditaires a été la rencontre d'un PAr travaillant sur des problématiques similaires. Nous avons ainsi rencontré Thomas Blondelle le jeudi 2 février.

Lors de cette réunion, ce dernier nous a présenté son objectif de travail, avant que nous abordions la liste des bases de données qu'il avait déjà pu récupérer et en partie pré-traiter.

Les bases de données proposées consistaient essentiellement en les deux jeux suivants :

- *Contes* Une base de données constituées de contes libres de droits et trouvables en ligne. Cette base contenait ainsi les contes, découpés en paragraphe.
- *Wikipédia* Une base de données constituées de phrases issues de Wikipédia, avec l'avantage d'être en français.

Cependant, dans le cadre de notre application choisie, nous étions à la recherche de bases de données en lien avec l'analyse de sentiments, comme des critiques (de films de produits), ou des phrases labellisés dans cet objectif. Ainsi, les bases de données proposées par le PAr ne répondaient pas à ces attentes et n'ont donc pas été retenues pour la suite de ce travail.

2.2. Les bases de données utilisée

2.2.1. *Recherche de bases de données en analyse de sentiment*

Suite à la rencontre avec le PAr décrite à la Section 2.1, nous avons alors cherché des bases de données plus orientées sur l'analyse de sentiments. Nos recherches nous ont permis de dégager les bases détaillées à la Table 1.

	Nom	Quantité de données	Origine	Source
🎞	Large Movie Review Dataset	25000 × 2	Stanford	[?]
	Rotten Tomatoes Dataset	20 399	Kaggle	[?]
🐦	Twitter Sentiment Corpus	5500	Niek Sanders	[?]
	Twitter Sentiment Analysis Corpus	1 578 627	Anonyme	[?]
💬	Sentiment Analysis Dataset	9645	Stanford	[?]
	UMICH SI650	40 000	Kaggle	[?]
a	Amazon reviews	> 6 millions	Julian Mc. Auley	[?]

Table 1 • Bases de données considérées pour le projet

Les bases de données de la Table 1 ont été regroupées par types de données. On retrouve ainsi :

- 🎞 Des bases de données liées à des critiques de films. Chaque commentaire est associé à une évaluation faite par l'utilisateur, ainsi on garantit le fait que la base soit labellisée manuellement par des humains, seul l'extraction du site a été automatisée.
- 🐦 Des bases de Tweets, dans le détail la base la plus large inclus la plus petite de 5500 Tweets. La base de 5500 Tweets a été labellisée manuellement par son auteur, ce qui garantit le réalisme des annotations. La deuxième base quant à elle a été labellisé par apprentissage semi-automatique, ainsi on ne peut être sûr de la validité de tous ces labels.

- Des bases données diverses de commentaires ou de phrases diverses et labellisées. Celle issue de Kaggle provient d'un challenge de Machine Learning et a été fournie par l'université du Michigan. La base de Stanford quant à elle est un peu particulière et correspond à des arbres syntaxiques sur le modèle de celui qui vous a été présenté à la Section **??**. Le découpage et la validation des données ont été réalisés avec Amazon Turk, il est difficile d'étendre facillement son volume.
- Cette dernière base de données est la plus conséquente et a été proposée par un professeur de l'université de Californie San Diego. Elle provient d'un scrapping d'Amazon et est donc elle aussi initiallement assurée d'être anotée par des humains. Il s'agit de la base la plus volumineuse que nous ayons pu récolter.

2.2.2. *Choisir quelle base utiliser*

Pour notre projet, il nous a alors fallu décider sur quels types de bases de données nous allions concentrer nos efforts. Nous avons alors dressé un tableau comparatif des caractéristiques de ces bases qui vous est proposé à la Table 2. Les critères retenus sont les suivants :

- *Quantité de données* Le volume est-il réaliste vis-à-vis d'une utilisation pour du Deep Learning ou le volume de données demandé est souvent grand (insuffisant ; suffisant ; très fourni). Nous avons généralement considéré que nos bases de données devaient contenir au moins 10 000 données pour être d'un suffisantes.
- *Qualité de l'annotation* Permet d'évaluer la fiabilité que l'on fournit envers l'annotation réalisée. La meilleure solution étant les cas où l'on est assuré que la base a été annotée (et/ou vérifiée) par des humains (peu fiable ; fiable).
- *Type de données* Le type de données qui est annoté, à savoir des phrases, des paragraphes,...
- *Format de données* Quel est la forme des données, et dans quelle mesure leur lecture sera facilement réalisable avec des outils à notre disposition. Le jugement viendra du fait que la structure soit courante ou plus exotique.

	Nom de la base	Quantité de données	Qualité de l'annotation	Types de données
	Large Movie Review Dataset	50 000	+	Paragraphes
	Rotten Tomatoes Dataset	20 399	++	Phrases (et arbre syntaxique
	Twitter Sentiment Corpus	5500	+	Tweets
	Twitter Sentiment Analysis	1 578 627	-	Tweets
	Sentiment Analyses Dataset	9645	+	Arbres syntaxiques
	UMICH SI650	40 000	+	Phrases
	Amazon reviews	> 6 000 000	+	Paragraphes

Table 2 • Comparatif des intérêts et des problèmes techniques des différentes bases de données repérées

Suite à ce comparatif, nous avons écarté les bases de données liées à Twitter. En effet, le format de texte de ces dernières (le Tweet), bien qu'intéressant, utilise une syntaxe particulière, qui n'est pas nécessairement représentative de la « vraie » syntaxe. Si son étude n'est pas dénuée d'intérêt, nous avons préféré privilégier une cohérence dans nos réalisations, en nous focalisant plus sur des phrases écrites dans une syntaxe non contrainte. De plus, le fait que la base Twitter pour l'analyse de sentiment utilisable en Deep Learning provienne d'un apprentissage semi-supervisé nous a conforté dans notre choix, dans la mesure où nous préférions privilégier des bases annotées par des personnes physiques.

Nous nous sommes ainsi orienté vers une étude sur les retours de clients d'Amazon ainsi que sur les critiques de cinéma.

2.3. Mise en forme des données

Nous avons désormais cerné l'ensemble des bases de données que nous souhaitions utiliser. La dernière étape de préparation de ces dernières était alors de les remettre en forme pour leur utilisation.

2.3.1. *Des formats divers*

Les bases de données retenues provenaient de sources variées, et ainsi leurs formats n'étaient pas uniformes. Pour en témoigner, nous allons fournir ci-après quelques exemples issues de ces dernières.

Retours clients d'Amazon Cette base de données est la plus conséquente que nous ayons récupérée. Il s'agit également des données les mieux structurées que nous avions) disposition, en effet, ces dernières étaient initiallement proposées sous la forme de fichiers JSON. De même, chaque fichier JSON correspondait à une thématique précise (outillage, livres, CDs,...), ainsi cette base de données étaient proposées sous formes de fichiers cohérents et structurés.

À titre d'exemple, vous pourrez trouver ci-dessous le contenu d'un critique trouvée dans un des fichiers originaux de la base de données d'Amazon.

```
{
  "reviewerID": "A2IBPI20UZIR0U",
  "asin": "1384719342",
  "reviewerName": "cassandra tu \"Yeah, well, that's just like, u...",
  "helpful": [0, 0],
  "reviewText": "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,",
  "overall": 5.0,
  "summary": "good",
  "unixReviewTime": 1393545600,
  "reviewTime": "02 28, 2014"
}
```

On retrouve ainsi dans cette base de données de nombreux éléments concernant l'auteur et le contexte du retour client. Dans le cadre de notre étude cependant, peu de champs nous intéresse, à savoir les suivants :

- *reviewText* Le texte du retour client sur le produit. Pour notre application, c'est à partir de ce dernier que l'on devra chercher à trouver la note attribuée.
- *overall* La note donnée par le client au produit, ie. celle que l'on va chercher à prévoir (elle reflète le sentiment du client sur le produit).
- *summary* Il s'agit d'une version condensée du champs `overall` permettant de savoir si une échelle « bon/pas bon » (binaire) le sentiment du client. En regardant la base de données, on remarque que la limite se situe aux alentours de 2,5-3, nous considérerons donc que les revues clients avec une note supérieure ou égale à 3 sont considérées comme positives.

Ces différents champs seront réutilisés dans le code pour harmoniser ces bases de données entre elles.

Critiques de cinéma (IMDB)

Critiques de cinéma (Rottent Tomatoes)

3 • Réseau de neurones

3.1. Les différentes technologies

4 • Entraînement d'un réseau de neurones

4.1. Structure d'entraînement

4.2. Résultats

4.2.1. *Sur les données d'apprentissage*

4.2.2. *Sur les données de tests*

4.3. Autres essais possibles

4.4. Comparaison à d'autres modèles simples

Conclusion