

# Participez à une compétition Kaggle !

Projet N° 8  
Irina Maslowski



# Plan

- ◆ La présentation de la compétition Kaggle
- ◆ L'apport de la communauté Kaggle
- ◆ « Plus fort, plus haut, plus vite. » (Pierre De Coubertin) La modélisation :
  - choix de la taille de dataset
  - choix du modèle
- ◆ Les résultats et le modèle final choisi
- ◆ Les leçons des gagnants
- ◆ Conclusion

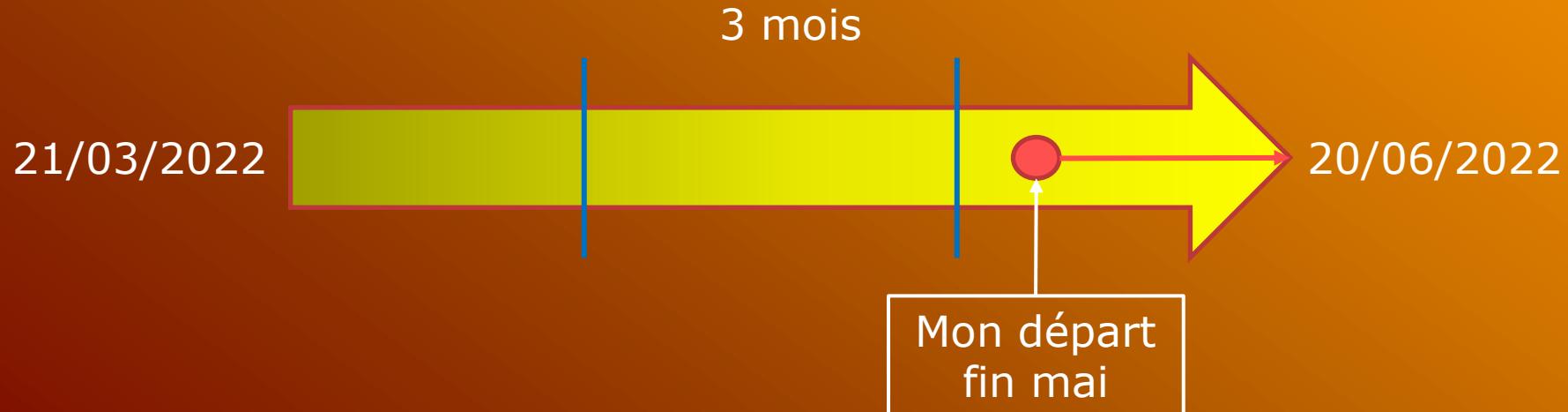
# PRÉSENTATION DE LA COMPÉTITION KAGGLE



# Présentation de la compétition Kaggle



## "U.S. Patent Phrase to Phrase Matching"



# Présentation de la compétition Kaggle

- ◆ But : évaluer le niveau de similarité sémantique entre deux phrases sur l'échelle de 0 (aucune) à 1 (forte)
- ◆ Contexte: brevets

Ancre	Cible	Contexte	Score
abatement	act of abating	A47	0.75
abatement	active catalyst	A47	0.25
abatement	eliminating process	A47	0.5
abatement	forest region	A47	0

# Présentation de la compétition Kaggle

- ◆ Métrique:
  - Coefficient de corrélation Pearson entre les scores de similarité prédit et réel



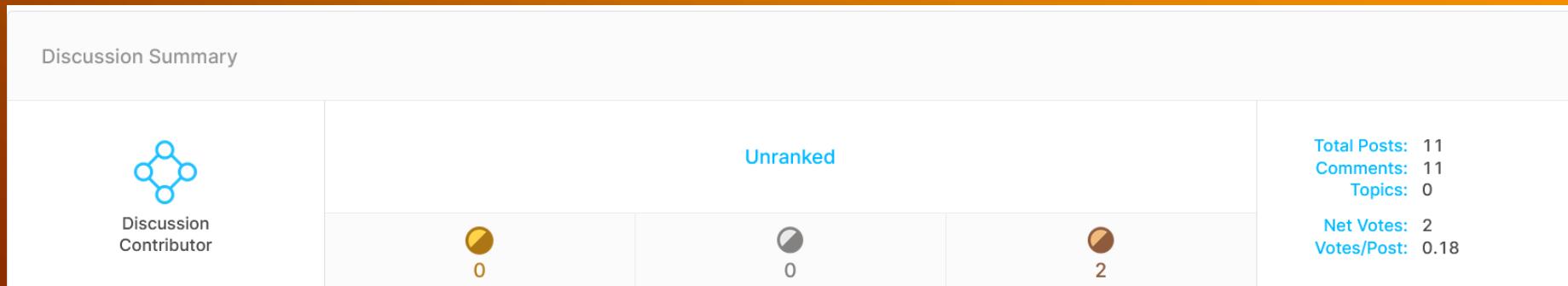
# L'APPORT DE LA COMMUNAUTÉ KAGGLE

# L'apport de la communauté Kaggle

- ◆ « L'herbe est toujours plus verte chez le voisin... » oui! si vous êtes débutant!
  - notebooks avec des explications pour les débutants
  - possibilité de répliquer un notebook pour l'élaborer
  - commentaires et votes pour un notebook publique
  - discussions

# L'apport de la communauté Kaggle

## ◆ Mon utilisation de la communauté: - commentaires



 In Depth EDA & 3 Model Ensemble

Comment a month ago on notebook

Thank you for your detailed notebook. It differs a lot from those where the authors use models already trained and fine-tuned elsewhere. As a novice in kaggle competitions, I appreciate a lot your analysis and takeaways.



# L'apport de la communauté Kaggle

- ◆ Mon utilisation de la communauté:
  - partage des notebooks



IRINA MASLOWSKI · COPIED FROM IRINA MASLOWSKI +63, -217 · LINKED TO [GITHUB](#) · 1 MO AGO 29 VIEWS



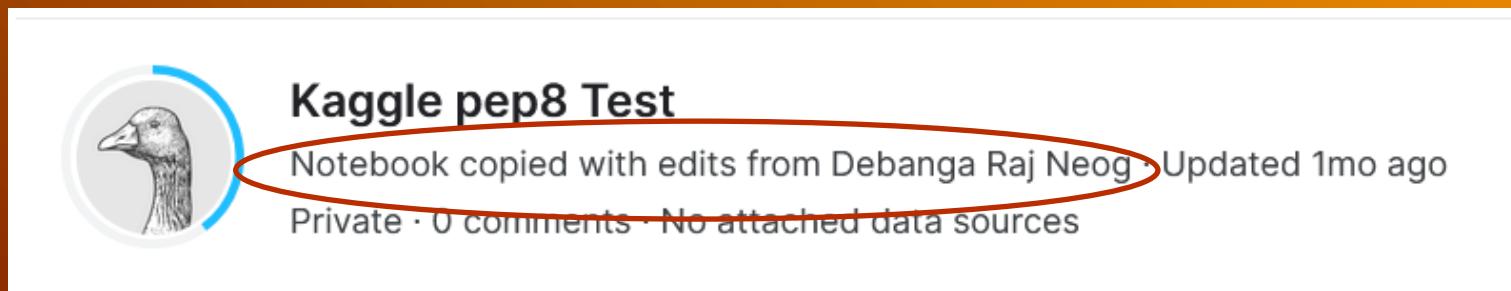
## U.S. Patent Phrase\_basic\_EDA

Python · [Cooperative Patent Classification Codes Meaning](#), [images\\_us\\_patent](#), [U.S. Patent Phrase to Phrase Matching](#)

# L'apport de la communauté Kaggle

## ◆ Mon utilisation de la communauté:

- reproduction et évolution des notebooks d'autres kagglers



Kaggle pep8 Test

Notebook copied with edits from Debanga Raj Neog · Updated 1mo ago

Private · 0 comments · No attached data sources

A screenshot of a Kaggle notebook card. It features a circular profile picture of a duck on the left. To the right of the picture is the title "Kaggle pep8 Test". Below the title, there is a note that reads "Notebook copied with edits from Debanga Raj Neog · Updated 1mo ago". At the bottom of the card, it says "Private · 0 comments · No attached data sources". A red oval has been drawn around the text "Notebook copied with edits from Debanga Raj Neog · Updated 1mo ago" to highlight it.

# L'apport de la communauté Kaggle

- ◆ Mon utilisation de la communauté:
  - réutilisation des bouts de codes d'autres kagglers dans mes notebooks

## U.S. Patent Phrase\_Deberta-V3\_small

Notebook Data Logs Comments (0) Settings



### Credits:

this notebook is based on the notebook "Getting started with NLP for absolute beginners" by @Jeremy Howard

### Imports

[2]:

```
import tensorflow as tf
import tensorflow_hub as hub
import seaborn as sns
import matplotlib.pyplot as plt
```



La modélisation

**« PLUS FORT, PLUS HAUT,  
PLUS VITE. » (PIERRE DE COUBERTIN)**

# La modélisation

## ◆ Plus de données:

### – Données de départ:

- ◆ 36 473 lignes, 5 colonnes
- ◆ lignes très courtes
- ◆ 5 valeurs de scores différentes: « 0,00 », « 0,25 », « 0,5 », « 0,75 » et « 1,00 »

	id	anchor	target	context	score
0	37d61fd2272659b1	abatement	abatement of pollution	A47	0.50
1	7b9652b17b68b7a4	abatement	act of abating	A47	0.75
2	36d72442aefd8232	abatement	active catalyst	A47	0.25
3	5296b0c19e1ce60e	abatement	eliminating process	A47	0.50
4	54c1e3b9184cb5b6	abatement	forest region	A47	0.00

# La modélisation

◆ Plus de données:

– Données supplémentaires:

- ◆ description du contexte. Source: « The Cooperative Patent Classification »
- ◆ 260 476 lignes, sept colonnes

– Données assemblés:

	id	anchor	target	context	score	section	context_title	section_title
0	37d61fd2272659b1	abatement	abatement of pollution	A47	0.50	A	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; CO...	HUMAN NEEDS
1	7b9652b17b68b7a4	abatement	act of abating	A47	0.75	A	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; CO...	HUMAN NEEDS
2	36d72442aef8232	abatement	active catalyst	A47	0.25	A	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; CO...	HUMAN NEEDS
3	5296b0c19e1ce60e	abatement	eliminating process	A47	0.50	A	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; CO...	HUMAN NEEDS
4	54c1e3b9184cb5b6	abatement	forest region	A47	0.00	A	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; CO...	HUMAN NEEDS

# La modélisation

- ◆ Résultat d'augmentation de données:
  - Test d'influence de différentes quantités de données sur un modèle
    - ◆ contexte
    - ◆ score
    - ◆ temps d'apprentissage



# La modélisation

◆ Quantité de données choisie:

- ◆ ancre
- ◆ cible
- ◆ contexte
- ◆ titre de contexte
- ◆ titre de section

# La modélisation

- ◆ Recherche d'un modèle le plus performant :
  - Baseline: USE (Universal Sentence Embendings) (Cer et al., 2018)



# La modélisation

- ◆ Recherche d'un modèle le plus performante :
  - Modèles pré-entraînés disponibles sur Hugging Face:
    - ◆ sur les données différentes:
      - généralistes (deberta-V3)
      - spécialisés sur les brevets (PatentS-BERTa de AI-Growth-Lab, bert-for-patents-64d de prithivida)
    - ◆ de différentes tailles (deberta-V3-small/base/large)

# La modélisation

- ◆ Recherche d'un modèle le plus performante :
  - Modèles pré-entraînés disponibles sur Hugging Face:
    - ◆ pour une tâche similaire:
      - deberta-v3-small-finetuned-mrpc (MRPC: corpus de paraphrases)
      - mberta-v3-base (un modèle multilingue)

# LES RÉSULTATS ET LE MODÈLE FINAL CHOISI



# Les résultats et le modèle final choisi

Modèle	Nombre d'époques d'entraînement	Score d'entraînement	Temps d'entraînement, min	Score dataset test	Score dataset final
USE	baseline		9,00	0.4918	0.5205
deberta-v3-small (pas de "context_title")	4	0.8094	5,48	0.7977	0.8137
mdeberta-v3-base (pas de "context_title")	4	0.8012	11,85		

# Les résultats et le modèle final choisi

Modèle	Nombre d'époques d'entraînement	Score d'entraînement	Temps d'entraînement, min	Score dataset test	Score dataset final
AI-Growth-Lab PatentSB ERTa (pas de "context_title")	4	0.8068	6,95		
PatentSB ERTa + deberta-small			5,81	0.8003	0.8227

modèle ensembliste

# Les résultats et le modèle final choisi

## ◆ Le modèle final:

Modèle	Nombre d'époques d'entraînement	Score d'entraînement	Temps d'entraînement, min	Score dataset test	Score dataset final
<b>PatentSB ERTa + deberta-small + mberta-base</b>			<b>6,86</b>	<b>0.8135</b>	<b>0.8312</b>

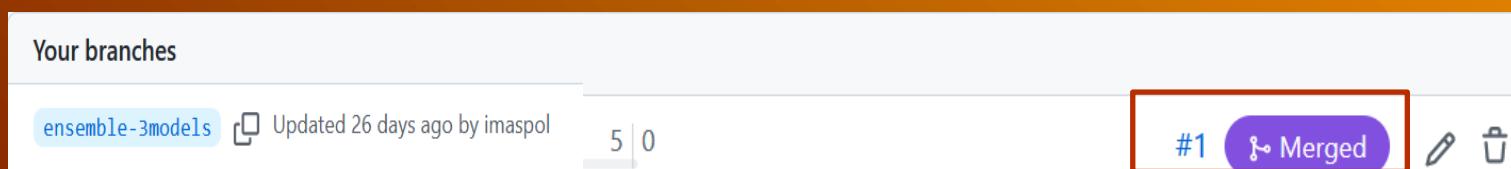
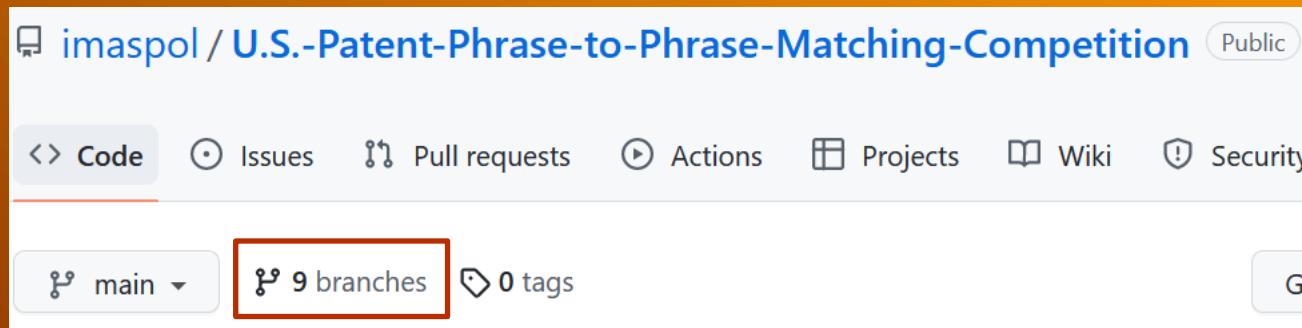
modèle ensembliste

# Les résultats et le modèle final choisi

- ◆ Mes notebooks sont disponibles sur :
  - Kaggle:
    - ◆ Inference Ensemble 3models
    - ◆ U.S. Patent Phrase Deberta-V3 small
    - ◆ U.S. Patent Phrase basic EDA
  - Git-hub:
    - ◆ @imaspol:  
<https://github.com/imaspol/U.S.-Patent-Phrase-to-Phrase-Matching-Competition>

# Les résultats et le modèle final choisi

## ◆ Gestion de versions:



# Les résultats et le modèle final choisi

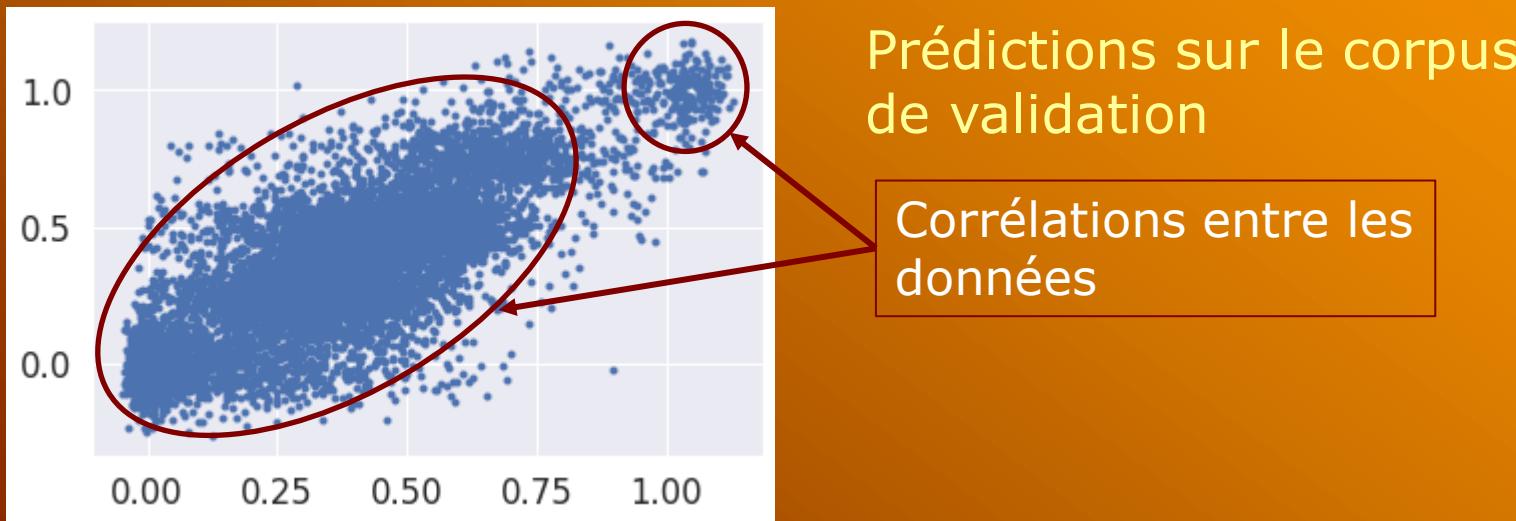
- ◆ Vérification de correspondance du code au standards pep8:
  - pylint

The background image shows a perspective view of a running track's lanes. The lanes are marked by white lines on a reddish-brown surface. The lines converge towards the bottom left of the frame, creating a sense of depth. In the foreground, the white chalk marks of the starting blocks are visible, including the diagonal 'X' and the lane numbers 1, 2, 3, and 4.

# LES LEÇONS DES GAGNANTS

# Les leçons des gagnants

- ◆ Savoir arranger les données en fonction de leurs spécificités



Source de l'image: notebook « **In Depth EDA & 3 Model Ensemble** » de Valentin Werner

# Les leçons des gagnants

- ◆ Utiliser la validation croisée
- ◆ Faire la recherche des meilleurs paramètres:
  - Learning rate (vitesse d'apprentissage)
  - Nombre d'époques
- ◆ « Adversarial-training » (Miyato et al., 2016) pour augmenter la robustesse d'un modèle
- ◆ Fine-tuner un modèle pré-entraîné
- ◆ Assembler des modèles diverses dans un ensemble

# Conclusion

- ◆ expérience très enrichissante
- ◆ conditions de travail intéressantes:
  - temps limité
  - concurrence
  - partage
- ◆ axes de progrès:
  - analyse et compréhenstion des données
  - spécificités techniques:
    - ◆ d'apprentissage d'un modèle de Machine Learning (ex. la validation croisée)
    - ◆ des réseaux de neurones

# Références

- ◆ Daniel Cer, Yinfei Yang, and others. Universal sentence encoder. CoRR, abs/1803.11175, 2018.
- ◆ Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification, 2016.
- ◆ HE, Pengcheng, LIU, Xiaodong, GAO, Jianfeng, et al. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.



Merci!

Questions?