Une introduction aux données linguistiques

TALN Semaine 1

Plan pour aujourd'hui

1. **Qu'est-ce que le TALN et que couvrirons-nous dans ce cours?**

    1.1. Ce que nous couvrirons

    1.2. Format du cours

    1.3. Évaluations

    1.4. Respect and attentes

2. **Que sont les données linguistiques et pourquoi sont-elles si spéciales ?**

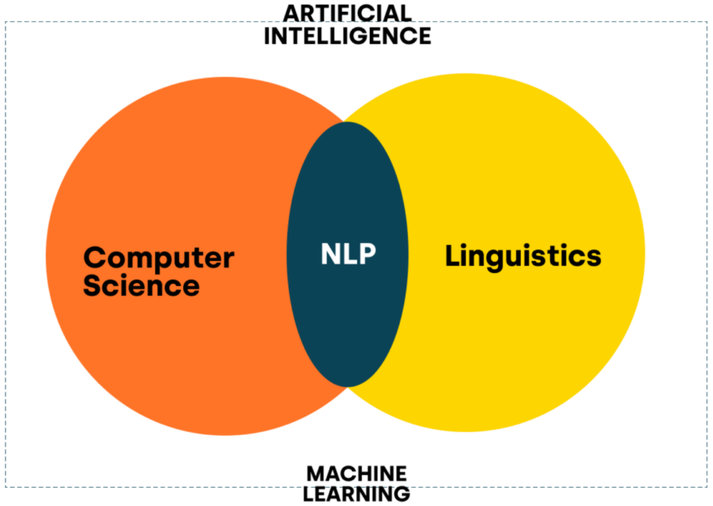
3. ***Exercices de groupe :*** **Mise en place de votre environnement de codage**

C’est quoi le TALN?

- **TALN= <u>Traitement automatique</u>+ <u>language naturel</u>**

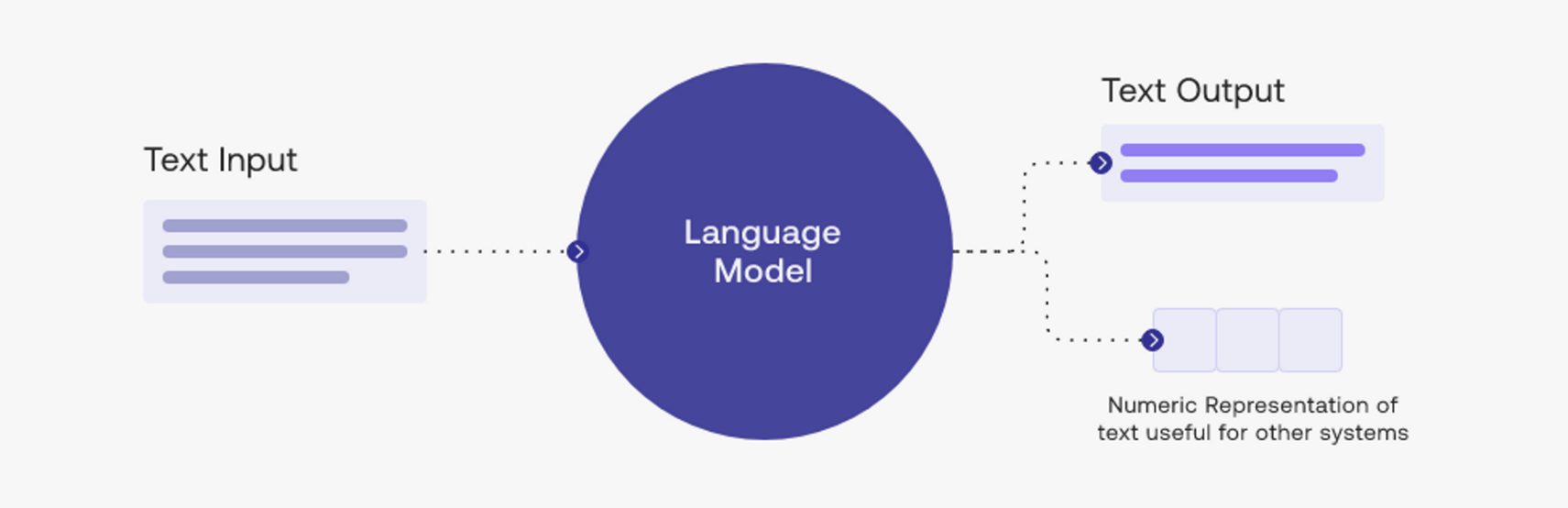
- **Définition :** Un domaine interdisciplinaire de la linguistique et de l'informatique qui développe et étudie des algorithmes pour *apprendre*, *traiter* et *générer* des données en langage naturel.

- Aujourd'hui, les algorithmes les plus importants sont l'apprentissage automatique, et en particulier basés sur les réseaux neuronaux.



Nous couvrirons:

LES MODÈLES DE LANGUAGE



Nous couvrirons:

LES MODÈLES DE LANGAGE

- **Modèles de langage N-gram**—Modèles probabilistes de langage contextualisé
- **Embeddings** — Sémantiques vectorielles
- **Modèles de langage neuronaux récurrents, LSTMs**—Modèles de langage neuronaux
- **Transformers** — Grand modèle de langage ou LLM
- M**odèles encodeurs** — Tâches de notation
- M**odèles encodeur-décodeurs** — tâches génératives
- **Peanfinage et apprentissage en-contexte**— le paradigme actuel de modèle fondateur

format du cours

- Durant chaque cours:
  - Première moitié— présentation du nouveau matériel
  - [pause de 15 minutes]
  - Deuxième moitié— exercices de codage en group
    (apportez vos ordis portables!)

Évaluations

- **Deux types d’évaluations:** Devoirs (50%) and Examens (50%)

- **Devoir 1 (25%)** - Dû le 26 février ..problèmes de codage appliqués
- **Examen d’intra (25%)** - XX mars ..questions théoriques
- **Devoir 2 (25%)** - Dû le 13 avril ..problèmes de codage appliqués
- **Final exam (25%)** - April XX ..problèmes d'étude de cas

Respect et attentes

- **Comment m'appeler en classe :** Professeure or Professeure Portelance.

- **Politique de courriel:** Veuillez prévoir jusqu'à 2 jours ouvrables pour toutes les réponses par courriel. Je ne répondrai pas durant les fins de semaines.

- **Polique de travail d’équipe:** Vous pouvez travailler sur vos devoirs en groupes de jusqu'à 3 personnes. Si vous travaillez avec d'autres, vous devez écrire leurs noms là où c’est indiqué. Vous devez toujours soumettre votre propre devoir et vous serez noté individuellement.

- **Politique de soumissions en retard:** C'est moins 15 % de sur votre note de devoir par jour de retard, sans exception. Vous recevrez les devoir 3 semaines à l'avance avant leur date d'échéance - planifiez en conséquence, il n'y a aucune raison pour que vous soyez en retard.

L’utilisation de l’IA générative

Il s'agit d'un cours sur le fonctionnement de l'IA générative.

Si vous le souhaitez, vous pouvez utiliser des outils d’IA générative pour la génération de code sur les exercices et les devoirs. Mais ne changez aucun code existant ou vous perdrez des points!

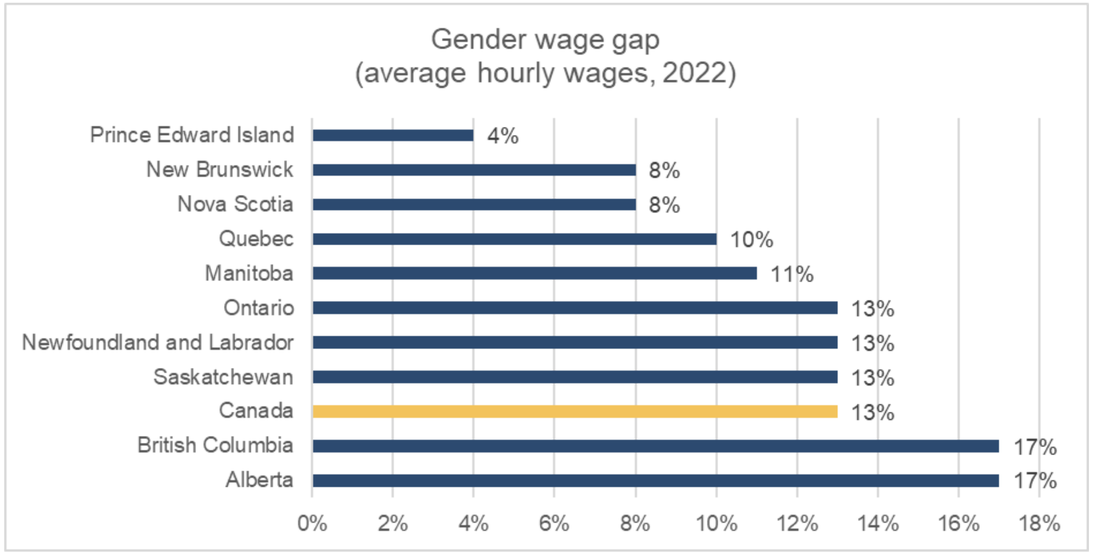
**Notez: pour les examens, vous ne serez autorisé qu'à une seule page de notes, pas d'ordis portables. Il peut y avoir des questions qui vous demande d’écrire du pseudo-code, alors assurez-vous de bien comprendre le code que vous soumettez à l'aide de l'IA générative avant de l'utiliser (Et évidemment assurez-vous qu'il fonctionne !!!!)

Maintenant, la partie l’fun.

Les données linguistiques… spéciales?

Mortgage Rates Hold Near 7%

The average rate on a 30-year fixed mortgage edged lower to 6.78%. It was one basis point higher last week.

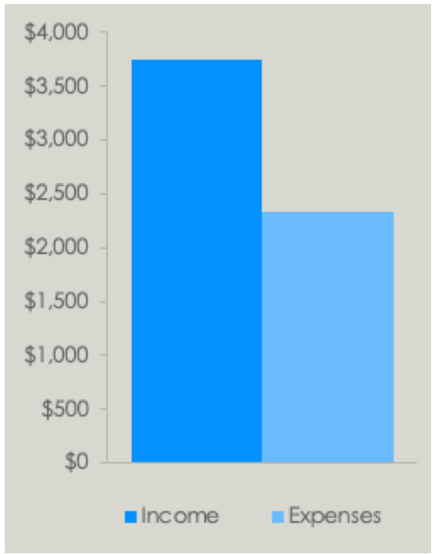


Year	Population
1860	56,802
1870	149,473
1880	233,959
1890	298,997
1900	342,782
1910	416,912
1920	506,676
1930	634,394
1940	634,536
1950	775,357
1960	740,316
1970	715,674
1980	678,974
1990	723,959
2000	776,733

CHAPTER 1

Loomings

Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It



Les données linguistiques… spéciales?

Le langage est un type de données structuré!

Les données linguistiques… spéciales?

C'est une phrase composée de mots.

-> [C’, est, une, phrase, composée, de, mots]

Les mots sont composés de morphèmes.

-> Les mot-s sont compos-é-s de morphème-s.

Morphèmes: Les plus petits constituants significatifs au sein d'une expression linguistique et en particulier dans un mot. Eg. Morpho-logi-que

Les données linguistiques… spéciales?

Document > sections > paragraphes > phrases > mots > morphèmes > charactères

Les données linguistiques… spéciales?

Le langage est un type de données hiérarchique!

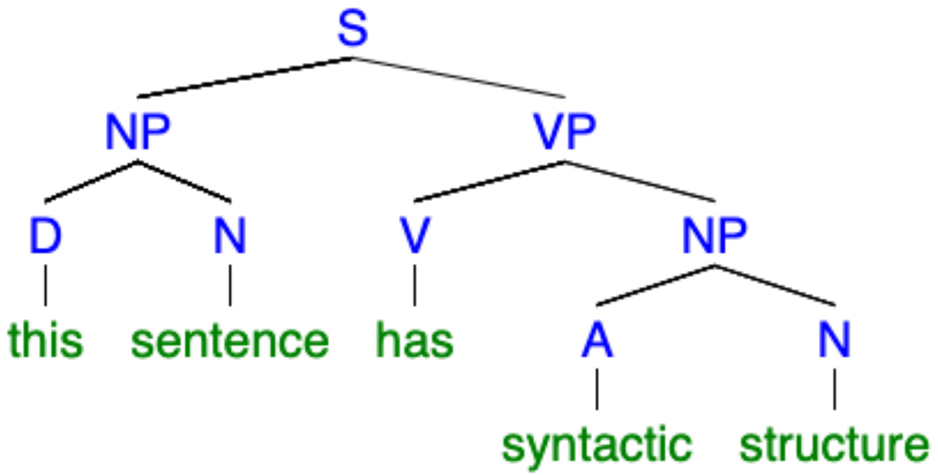
Linguistic data… why so special?

Il y a des types ou des catégories de mots: Noms, verbes, adjectifs, adverbes, mots fonctionnels

Toutes les catégories ne sont pas égales.

Les données linguistiques… spéciales?

La structure d'une phrase est appelée la **syntaxe**.



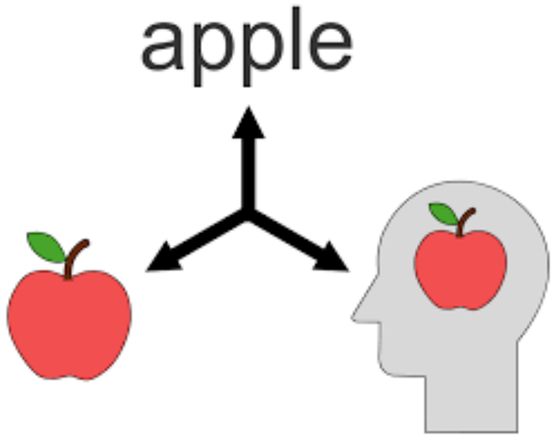
Les données linguistiques… spéciales?

Le langage est un type de données symbolique!

Les données linguistiques… spéciales?

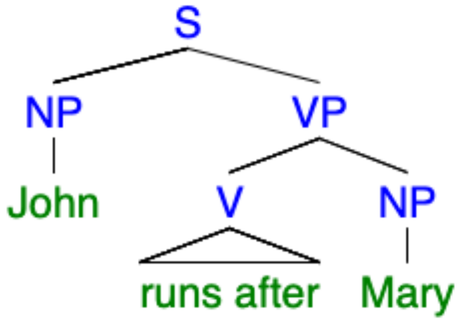
Le sens d'une phrase est appelé la **sémantique**.

⟷ John runs after Mary.

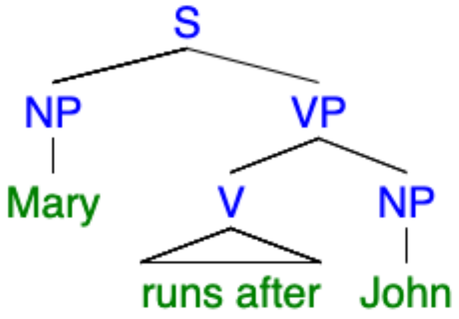


Les données linguistiques… spéciales?

Il existe une cartographie entre la structure et le sens.



vs



Les données linguistiques… spéciales?

Le language est très différent de toutes autres formes de données. De plus, la science n'a toujours pas de réponse à :

- Comment *apprenons*-nous notre langue ?
- Comment *comprenons*-nous notre langue ?
- Comment *produisons*-nous notre langue ?

Alors, comment sommes-nous censés trouver des algorithmes qui font exactement cela ?

Nous couvrirons:

LES MODÈLES DE LANGAGE

- **Modèles de langage N-gram**—Modèles probabilistes de langage contextualisé
- **Embeddings** — Sémantiques vectorielles
- **Modèles de langage neuronaux récurrents, LSTMs**—Modèles de langage neuronaux
- **Transformers** — Grand modèle de langage ou LLM
- M**odèles encodeurs** — Tâches de notation
- M**odèles encodeur-décodeurs** — tâches génératives
- **Peaufinage et apprentissage en-contexte**— le paradigme actuel de modèle fondateur

[Pause - 15 minutes]

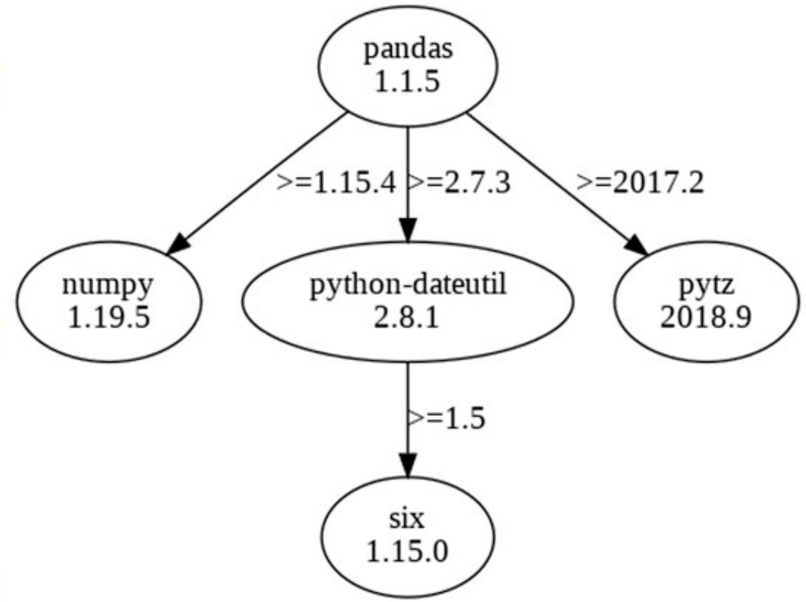
Qu'est-ce qu'un gestionnaire de paquets ?

- Un outil qui automatise le processus d'installation, de mise à niveau, de configuration et de suppression des paquets Python et de leurs dépendances.

- Les paquets sont des bibliothèques, des modules ou des frameworks externes qui étendent les fonctionnalités de base de Python, de sorte que vous pouvez inclure du code pré-écrit dans les projets sans tout écrire à partir de zéro.

- Les gestionnaires de paquets sont essentiels démêler les dépendances et résoudre les conflits complexes des paquets.

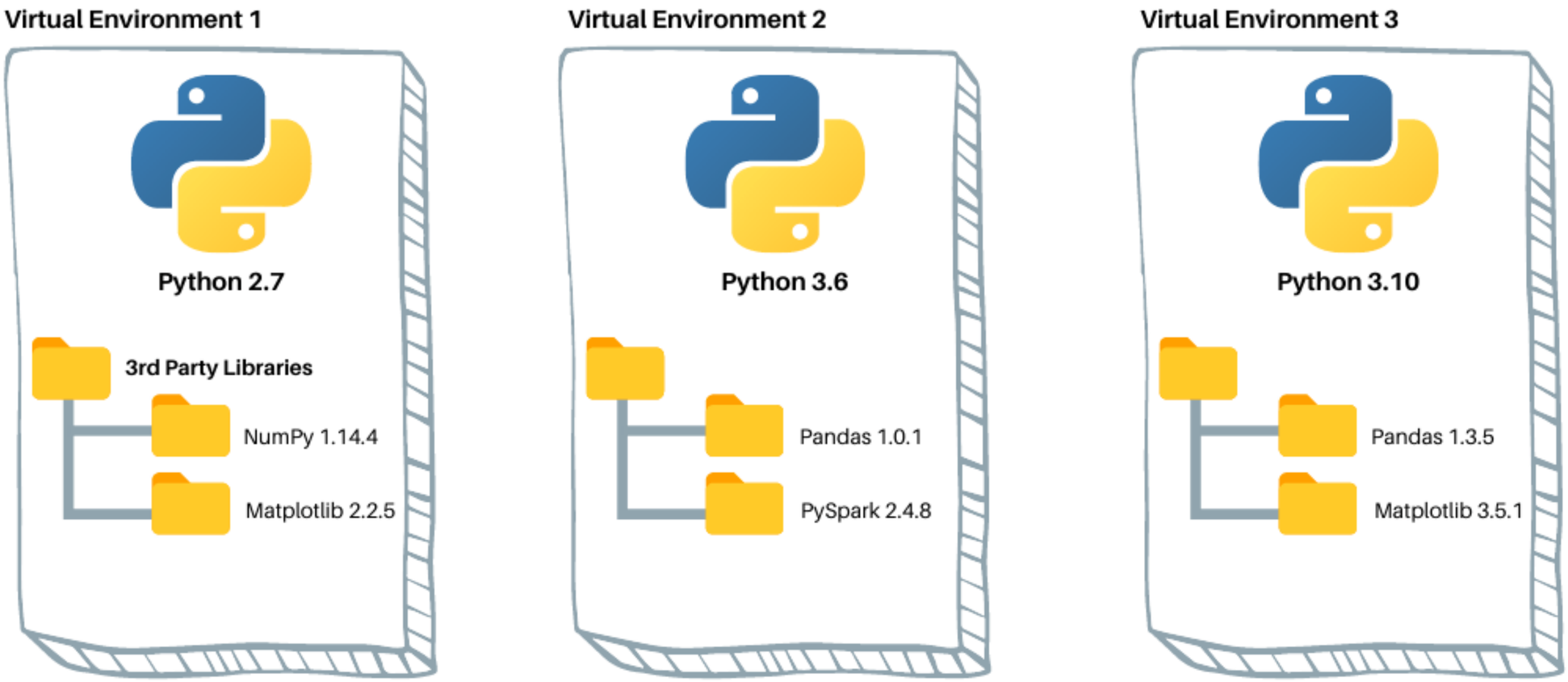
- Ex. `pip` ou `pip3`



Qu'est-ce qu'un environnement virtuel ?

- Un environnement virtuel Python est un répertoire autonome qui contient des dépendances spécifiques à un projet (paquets et leurs versions).
- Il est isolé des autres environnements de projet ou de l'installation globale de Python sur votre système, c'est-à-dire qu'un paquet installé dans un environnement n'affectera pas un autre.
- Il a son propre interprète Python dédié, c'est-à-dire que vous pouvez avoir une version différente de Python et `pip` de celle par défaut sur votre système.
- Ils aident à éviter les conflits de dépendance et encouragent la reproductibilité.
- Ex. `venv` ou `virtualenv`

Qu'est-ce qu'un environnement virtuel ?



Qu'est-ce qu'un gestionnaire d’environnements ?

- Un gestionnaire d’environnements/projets combine la gestion des paquets avec celle d’environnements virtuels, ce qui vous permet de créer facilement différents environnements et dépendances de paquets pour chaque projet.

- En tant qu'outil, il assure le suivi de tout vos environnements et vous aide à séparer et organiser vos projets.

- Ex: Anaconda, Poetry, ou uv .

Configuration de votre environnement

1. Allez au GitHub repo du cours et suivez les instructions de démarrage

www.github.com/evaportelance/HEC-NLP

2. Ouvrez les exercices/week1 et commencez à exécuter le code !