Applied NLP

Session 1

Lecturer: Narges Chinichian

Winter Semester 2025-2026

Day 1 Agenda:

1. Welcome & course introduction
2. Introduction: words as data
3. Setting up GitHub repository
4. Choosing your author(s) and language
5. Group discussion: why this author?
6. Homework: Reflection + repo initialization

About Me:

Dr. Narges Chinichian

PhD in Physics (Complex Systems), with 8+ years experience in **data science & machine learning.**

Worked on projects in **natural language processing since 2016.**

In this course, I want us to **treat literature as data**: collect, clean, analyze, and visualize language and see what algorithms can reveal about texts.

If I was a *word*, that word would be “chaii”.

My two hobbies are reading and climbing (I’m a freelance climbing trainer in Berlin)

About You?

Please share briefly:

1. Your name & where you're from.
2. What languages you speak/read.
3. A favorite author, book, or text (any language).
4. How would you rate your Python and Git skills?
5. Would you rather work solo or in a small group?

Bonus: If you were a *word*, which word would you be?

Course Overview

- 7.5 weeks | 1 day per week | 8 × 45min units
- Hands-on, project-based learning

- Final deliverables:
  - GitHub project (code, data, notebooks) (40%)
  - Medium article (1000–1500 words) (40%)
  - Presentation (5–7 min) (20%)
- Languages and authors of your choice

What To Expect From Deliverables

We will invite external audience.

Best presentation receives an award.

Best Medium article receives an award and is highlighted by the university.

Session Organization

First ~90 minutes: Each person or team presents their progress since the last session and receives feedback from the class.

Around 2pm: 15 minutes break.

Next ~90 minutes: New topic is taught.

Around 3:45: 60 minutes break.

Remaining time (from 4:45-8) : Hands-on work and individual or team project development.

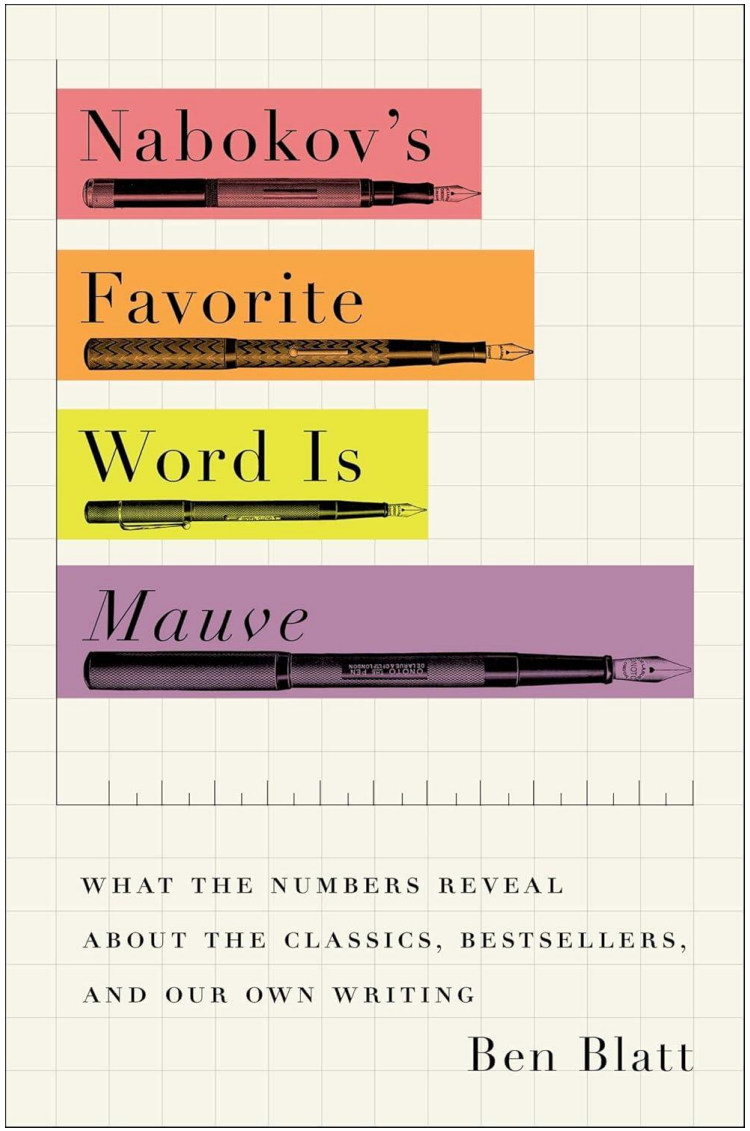
I would expect not more than **one week of absence** from each student unless in emergency cases with proof.

Important Note:

- For each session, apply **≥3 of the ~5** course measures to **your own text**.
- If one doesn’t fit, **justify** and use a **suitable substitute**.

Recommended Text To Read:

Nabokov's Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writing



Texts in Numbers

- What do you think can be measured in a text?
- If you didn’t know the author’s name, could you guess it from the numbers?
- How does **translation** change the “numbers” of a text . Does the fingerprint (of an author) survive?

From Al-Kindi to Shannon

9th century – Al-Kindi:

First to describe **frequency analysis** for breaking substitution ciphers

Realized that **letters occur with predictable frequencies** in a language

Turned linguistic patterns into a **tool for cryptanalysis**

20th century – Claude Shannon:

Formalized these ideas as **information theory**

Introduced **entropy** to measure predictability and redundancy in messages

- High entropy means every symbol is equally likely – like random noise.
- Low entropy means some symbols are predictable – like natural language, where certain letters appear more often.

Showed that the **same patterns** Al-Kindi used to *break* codes define the **limits of secure communication.**

Word Frequency

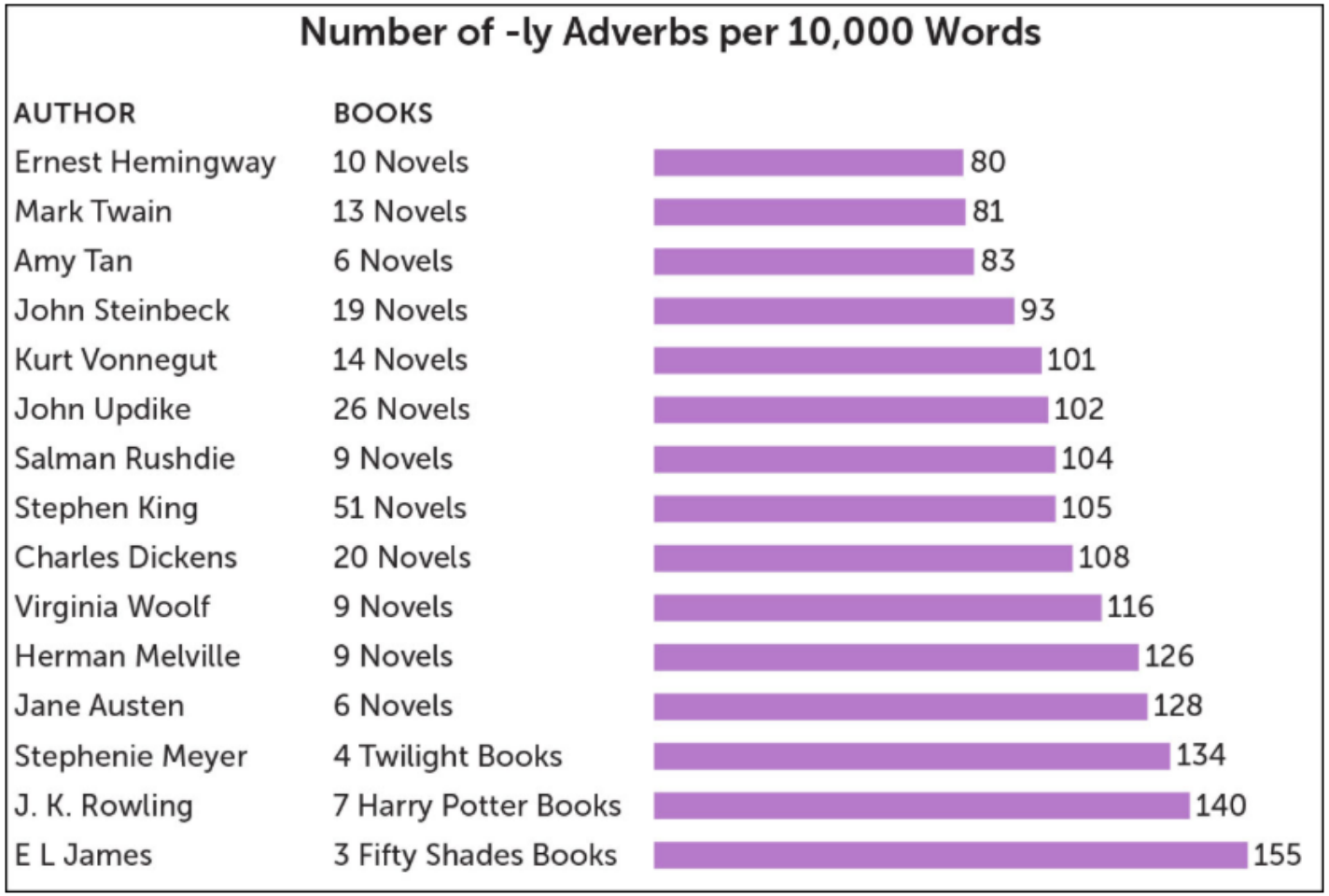
- In all languages, a few words are very common, most are rare
- What do you think the most frequent English word is? And the least frequent?

Rank	Nouns	Verbs	Adjectives	Prepositions	Others
1	time	be	good	to	the
2	person	have	new	of	and
3	year	do	first	in	a
4	way	say	last	for	that
5	day	get	long	on	I
6	thing	make	great	with	it
7	man	go	little	at	not
8	world	know	own	by	he
9	life	take	other	from	as
10	hand	see	old	up	you

Source: Oxford English Corpus (OEC), "Facts about the language," Oxford Dictionaries

The road to hell is paved with adverbs.

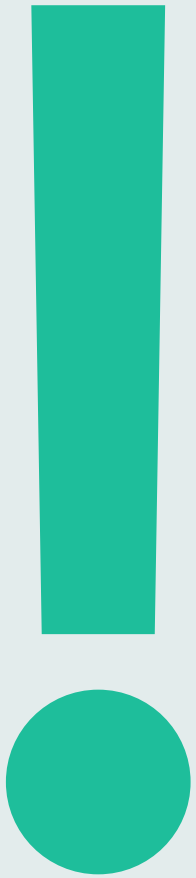
—STEPHEN KING



Blatt 2017 – *Nabokov's Favorite Word Is Mauve*

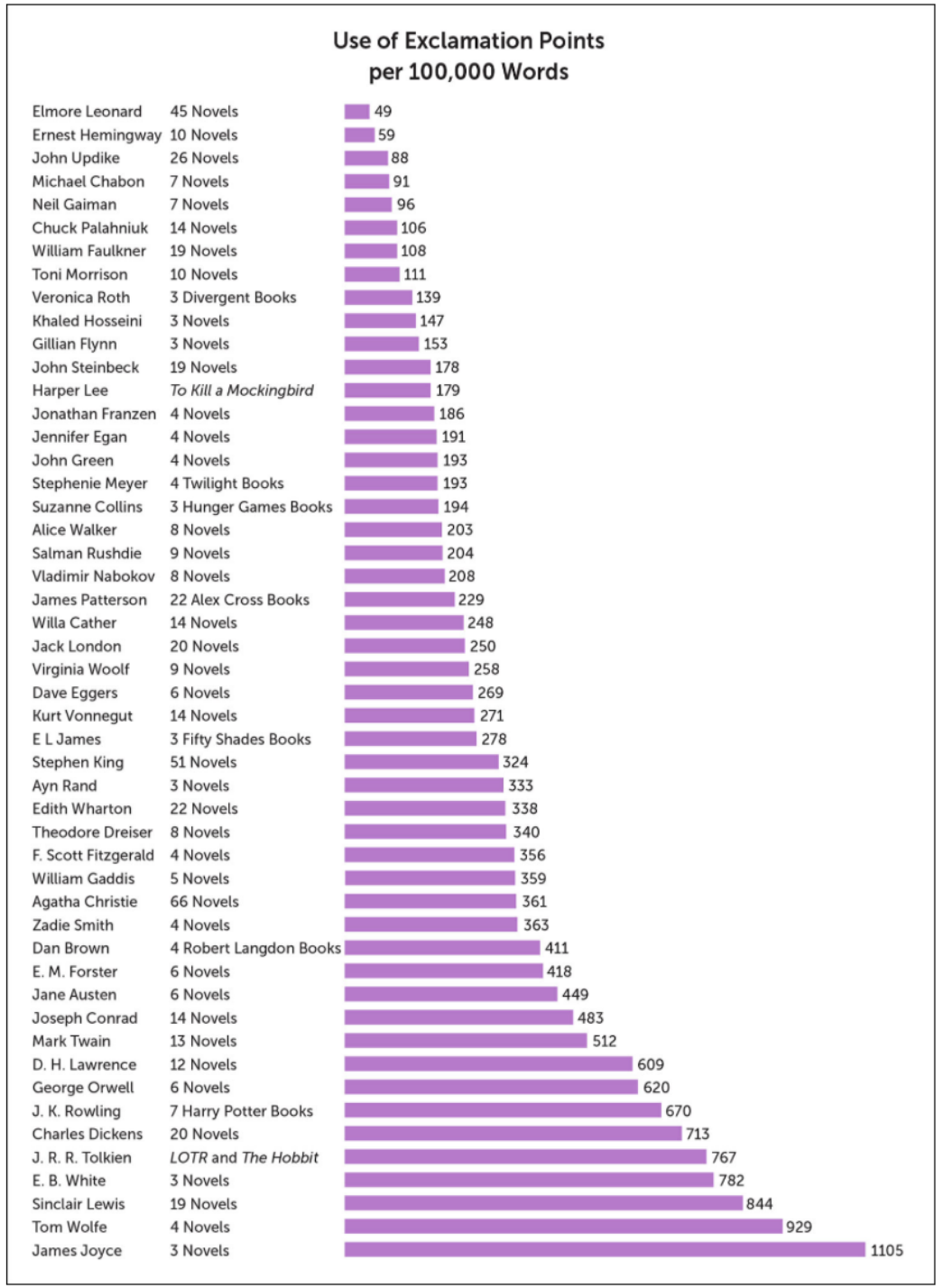
Exclamations!!!

Elmore Leonard in “Elmore Leonard's Ten Rules of Writing”:
“You are allowed no more than two or three per 100,000 words of prose.”



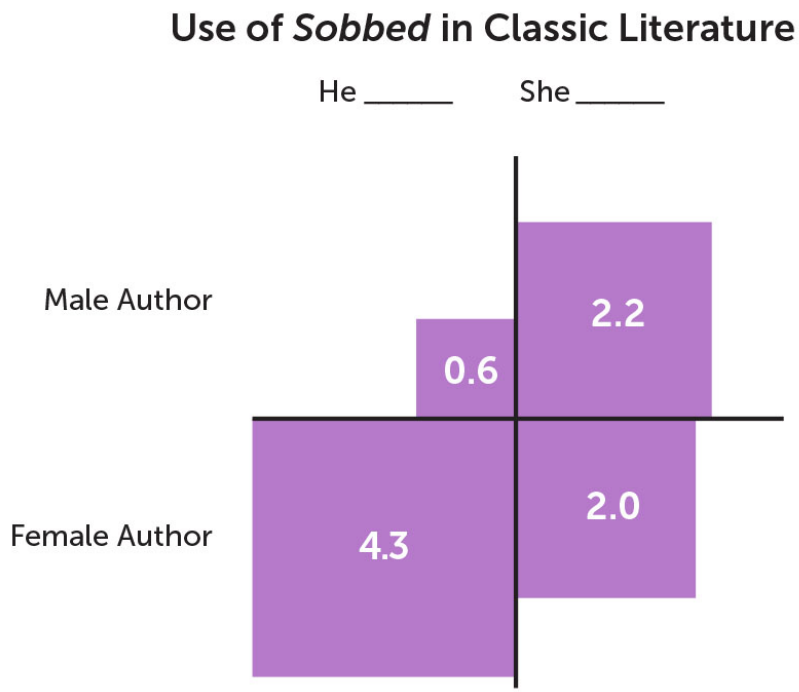
- Output: **45 novels ≈ 3.4 million words**
- Leonard’s rule: **≤ 2–3 exclamation points per 100,000 words**
- Allowed by his rule: **≈ 102** total (3.4M ÷ 100k = 34; 34 × 3 ≈ 102)
- **Actual used: 1,651**
- **Result: ~16×** more exclamation points than he recommends [*]

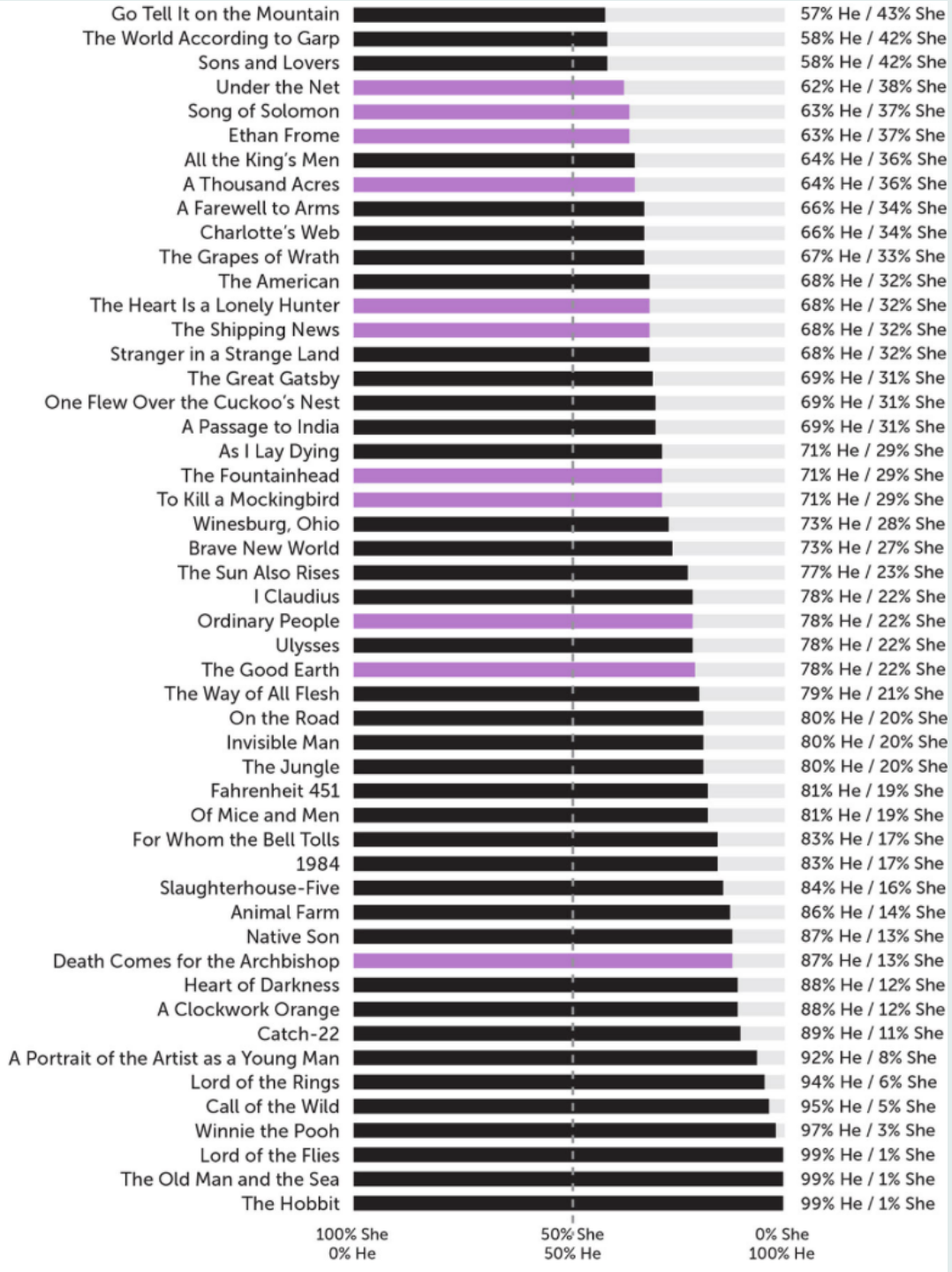
* From Blatt 2017 – *Nabokov’s Favorite Word Is Mauve*



* From Blatt 2017 – *Nabokov's Favorite Word Is Mauve*

Gender in the Text: Pronouns & Verbs





* From Blatt 2017 – *Nabokov's Favorite Word Is Mauve*

Bechdel's Test

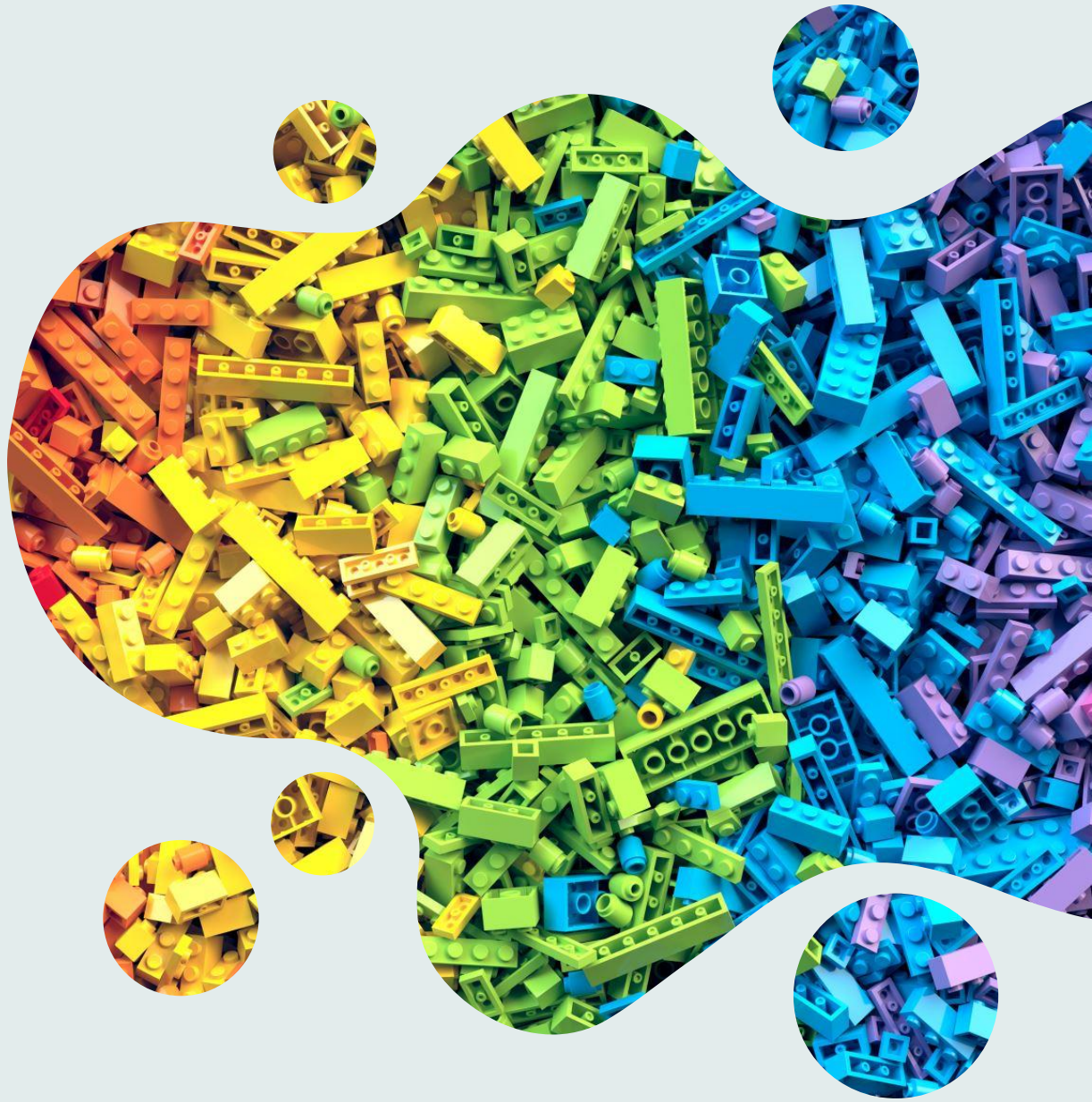
- The work has **at least two women** in it – often specified as two **named** women.
- They **talk to each other**.
- They talk about **something other than a man**.

Origin: The test comes from Alison Bechdel's 1985 comic strip **"The Rule"** (Bechdel credits the idea to her friend **Liz Wallace**).

Color Words & Descriptions

- **Alice through the Looking Glass** has way more color words than **Alice in Wonderland**.

- **Nabokov (Ben Blatt's counts): ~460 color words per 100k** tokens; his most distinctive favorite is **"mauve."**



Choose Your Group and Author

Frank Herbert – Dune

Question: How does the frequency of ecological and political terms change across the series?

→ Try topic modeling or keyword trend analysis across volumes.

Ferdowsi – Shahnameh

Question: Which words or phrases co-occur most often with mythical creatures?

→ Build a co-occurrence network to compare humans vs. non-humans.

Homer – The Odyssey

Question: How does sentiment vary between homecoming scenes and battle scenes?

→ Use sentiment or emotion lexicons to compare sections.

Chinua Achebe – Things Fall Apart

Question: How often and in what contexts are Igbo words used in the English text?

→ Identify code-switching patterns and visualize them over chapters.

Choose Your Group and Author

Haruki Murakami – Kafka on the Shore

Question: Can we detect shifts between “real” and “dreamlike” passages using embeddings or clustering?

→ Train sentence embeddings to cluster narrative modes.

Mary Shelley – Frankenstein

Question: How do the emotional tones of Victor and the Creature differ?

→ Run sentiment or emotion analysis per narrator.

Gabriel García Márquez – One Hundred Years of Solitude

Question: How do recurring family names and relationships connect across generations?

→ Build a character network graph from named-entity recognition.

The Epic of Gilgamesh

Question: What are the dominant themes before and after Enkidu’s death?

→ Use topic modeling to compare pre- and post-event sections.

Choose Your Group and Author

You have 90 minutes

Task:

Decide if you want to work Solo or as a group.
Select one author whose work you'll focus on for this whole course.

Instructions:

Pick your **language**.

Pick **one author** (I accept more if you convince me) and at least 2 works of that author.

Once you've decided, **write your/your team name and author here:**

https://tinyurl.com/4j4pjb6e

Tip:
Choose someone whose themes or style *interest* you. This will make your life much more fun.

Setting Up Your GitHub

Please all add your GitHub handle here:

It's what you get in your url when you are on your profile page:

So if you see:

https://github.com/NoCh-Git

Your handle is NoCh-Git.

Please enter your name and handle here:

https://tinyurl.com/34peb7cn

Join Course Organization

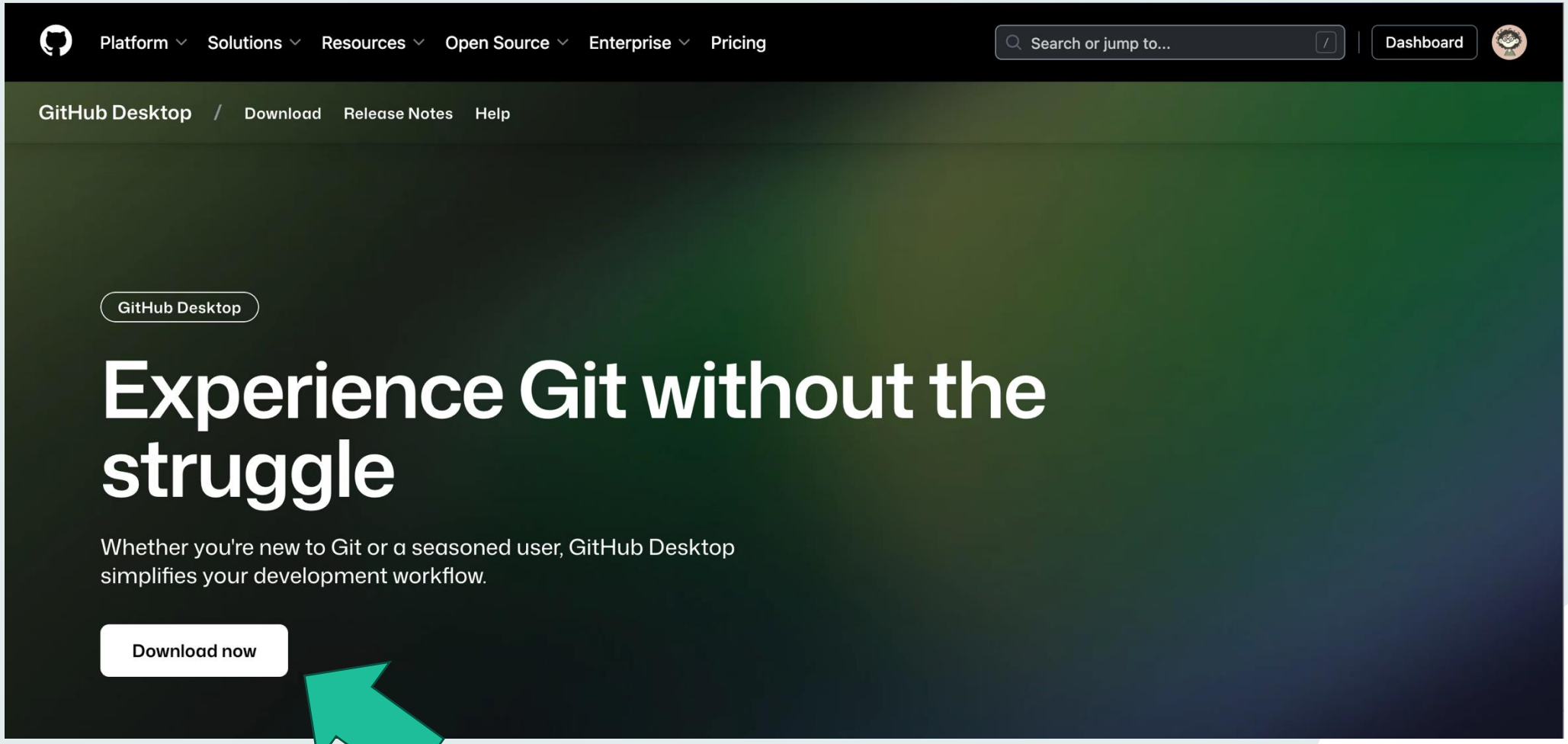
After you add your handle, I will invite you to join the course organizion.

https://github.com/AppliedNLP-SRH

This is where you will keep your project repo to be evaluated.

Install GitHub Desktop (Optional)

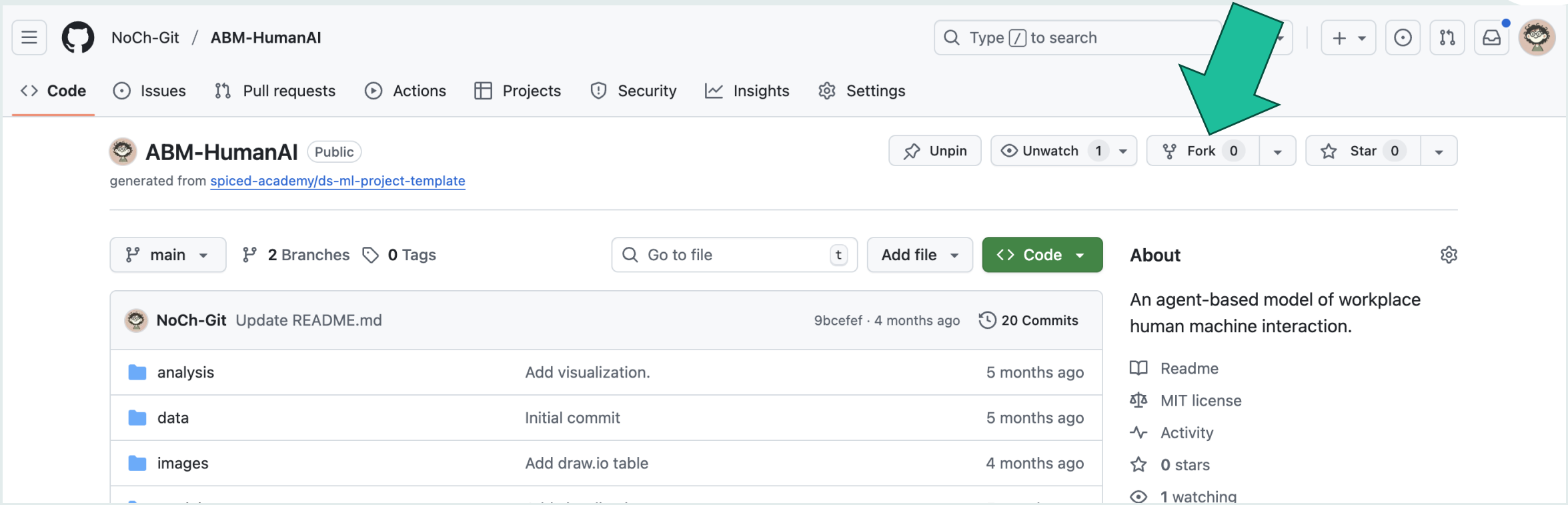
https://github.com/apps/desktop



Fork Repo of Today

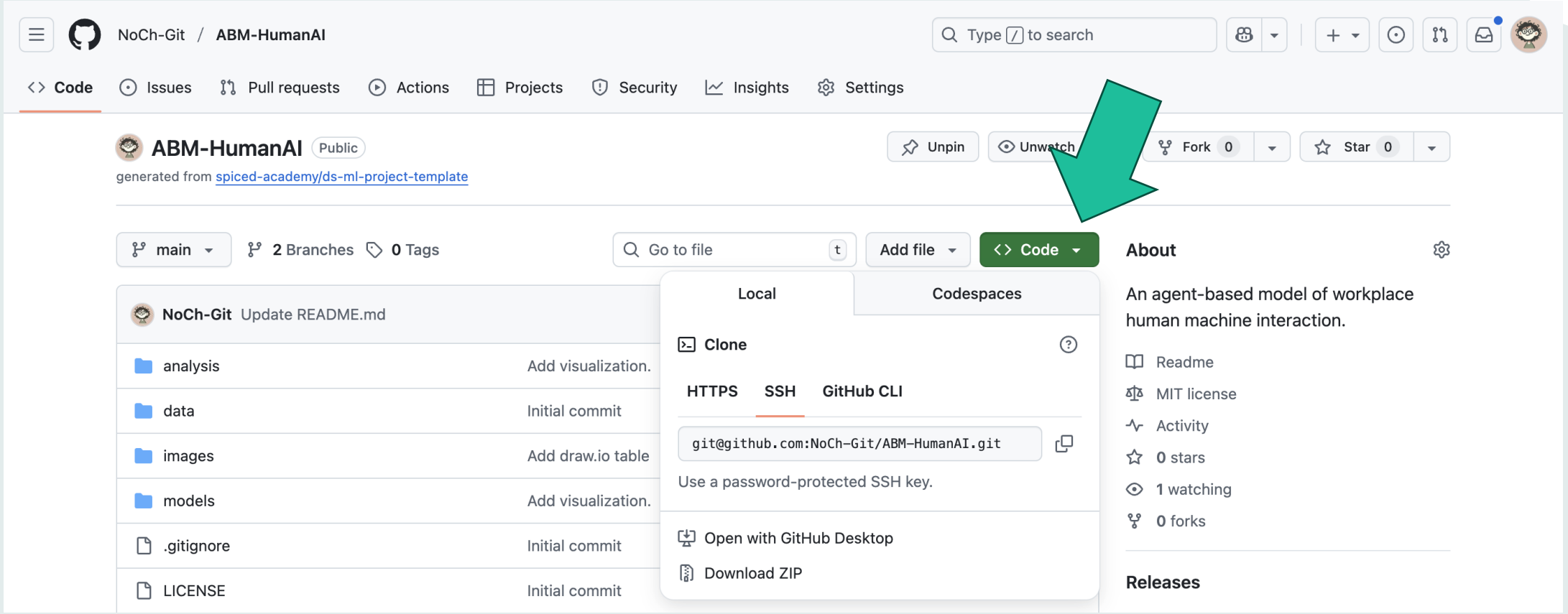
Forking a repo would create a copy of that repo for you that you can play with.

Choose yourself as the owner and untick the “Copy the main branch only” box.



Clone the Copy of Repo to Your Machine Using GitHub Desktop or CLT

You need to have a local copy of the Python notebooks.



Start Exploring