Instructions

- Solve the problems on the following page. Where applicable use python.
- Feel free to use an external library of your choice.
- There is no time limit or necessary right answer. You do not need to complete it, but make a solid effort so that we can talk about what you have done in a meaningful way!
- We are looking at the quality of your solution, the approach used and key decisions you made. Make sure you can explain it when asked.
- Feel free to ask questions and use the team here at Vector as a resource / sounding board. Assume you are working here and this is your project. How would you go about it?
- What we are interested in seeing is the way you think. Create a repository on gitlab / github and share it with us.
- Most importantly, enjoy it!

Problem

Part 1

- We want to build an entity normalization engine. The input to this engine is short strings / phrases that could encompass the following entities: company names, company addresses, serial numbers, physical goods and locations.
- Fictional Examples:
  - Company names: “Marks and Spencers Ltd”, “M&S Limited”, “NVIDIA Ireland”, etc.
  - Company addresses: “SLOUGH SE12 2XY”, “33 TIMBER YARD, LONDON, L1 8XY”, “44 CHINA ROAD, KOWLOON, HONG KONG”
  - Serial numbers: “XYZ 13423 / ILD”, “ABC/ICL/20891NC”
  - Physical Goods: “HARDWOOD TABLE”, “PLASTIC BOTTLE”
  - Locations: “LONDON”, “HONG KONG”, “ASIA”
- Build a system that can identify unique entities for each category above. Build one system for company names, one for company addresses, etc. Some of these will be trivial (remove spaces, edit distance, etc.) while others are more complicated and will need a trained model / some other form of knowledge and guidance.

- Examples:
    - “Marks and Spencers Ltd” and “M&S Limited” are the same entity, but they differ from “NVIDIA Ireland”
    - “LONDON” and “LONDON, ENG” are the same but they differ from “ASIA”

Part 2

- Let’s do something real world now! Your system will receive strings one by one and you have to group together any entities that you have come across earlier.
- For example, imagine this stream: “MARKS AND SPENCERS LTD”, “LONDON”, “ICNAO02312”, “LONDON, GREAT BRITAIN”, “TOYS”, “INTEL LLC”, “M&S CORPORATION Limited”, “LONDON, ENGLAND”. The groups that you would generate would then be:
    - “MARKS AND SPENCERS LTD” + “M&S CORPORATION Limited”
    - “LONDON” + “LONDON, GREAT BRITAIN” + “LONDON, ENGLAND”
    - “ICNAO02312”
    - “TOYS”
    - “INTEL LLC”

- Note that you will not have access to the full stream to begin with. Samples will appear incrementally when you have processed the previous one. You have to pick the latest sample received, scan the entries you already have, identify if the entity is a duplicate and then add it to a cluster / create a new cluster depending on your result.