Background:

Given a dataset of rental listings containing (among other things) their asking price, title, and listing URL/source, we are asked to estimate the number of rooms in a given unit. An annotated dataset containing a column, 'numrooms' serves as the ground truth set.

Current Approach:

We model this as an information extraction problem, where we limit ourselves to extracting the number of total bedrooms in each given unit. This limitation stems from the unavailability of rental unit descriptions and can be relaxed given listing descriptions. We begin by parsing the text of the title into structured data, using the wit.ai API. Wit.ai's entity extraction tool uses Facebook's Duckling, an open-source entity extraction library. However, we opted to use the API in lieu of the library for 2 reasons:

1. wit.ai is free and provides a RESTful API with no rate limits as of 5/23/2020.
2. Online training is easy. Using the "Understand" tool, developers provide example phrases and annotated data to continually improve wit.ai's capabilities, benefiting everyone who uses the API.

However, it is worth outlining wit.ai's shortcomings for posterity:

1. Wit.ai is owned and operated by Facebook. By 12/30/2020, they plan on phasing out standalone authentication, requiring developers to use their Facebook accounts, or sign up for Facebook if they don't have an account.
2. It is likely that they will institute API limits at some point in the future.
3. Other information extraction solutions, such as GATE have achieved near-human performance on MUC-6 (a competition-based information extraction conference sponsored by DARPA). The same cannot be said for wit.ai

Notwithstanding its limitations, we believe wit.ai is an excellent starting point, as evidenced by the discussion of the results below. A documented implementation of our approach is available at https://github.com/weirdindiankid/mapc

Results:

Using the *MAPCBedroomQuantifier* app at https://wit.ai/weirdindiankid/MAPCBedroomQuantifier/, we estimate the number of bedrooms for each listing, given its title. Given a total of 317 annotated entries, our results agree with

MAPC's results 56.1% of the time (i.e. 178 estimates out of 317 match those provided by MAPC). To understand the discrepancy, it is worth noting that the manually annotated data **had** to have taken the listing's description into account, since in many instances, the titles are equivocal. For instance, "Apartment in [location]" and "Private rent in shared family home [sic]" are examples of some of the titles we had to contend with. Our extant solution is *nearly* perfect when estimating the number of bedrooms available for rent, using the titles alone. However, given listing descriptions, modifying our wit.ai app to estimate the total number of rooms should be trivial.

Outlook:

While the existing system will likely suffice once descriptions are provided, following the tips below will improve performance significantly:

1. Build a rules-based inference system, where the first step might involve obtaining "known truths". For instance, for any listing in Boston or Cambridge, it might be worth plugging the listing's address into the assessor's database to get the number of rooms, before using an NLP-based solution.
2. Switching from wit.ai to GATE or other comparable solutions might improve the information extraction module's performance.