COL-772
Assignment-2

J. Shikhar Murty
2013EE10462

April 20, 2016

1 Features

I used 16 features for my best NER system, for example a complicated `regex` for detecting telephone numbers and some more general, for example a feature for detecting English words. Using these, my system's macro f score on the validation data was 76.7%, and on the test data was 75.8%

I'll focus on the more complicated features, and leave out simple ones.

LAND_AREA_FEATURE: This feature fires if the keywords `sq, sqft, feet, meter, acres,size` (total of 13 such keywords for measurement of land area) appear in either the current token or if the current token matches with a magnitude regex, and the next token is a measurement unit. Note that similar conditions may hold for a "C" so the feature tries to make sure that it is not a "C" token by scanning the next few tokens and checking if a `"per"` or a `"/"` appears.

PRICE_OR_COST: This features fires if a complex regex matches the current string token (the regex essentially looks for words like `"lac", "cr", "crores", "k", "rs"`) or if `"/-"` occurs. If the above regex doesn't match, it checks whether the token matches against a regex that looks for numbers and if the next token matches amount keywords like `"lac", "cr", "crore"`.

COST: This is similar to the above feature, except that it looks for keywords like `"per", "/", "pr"` and may also fire if the keywords `"sqft", "sq", "acre"` etc occur AND "per", "/" etc occur before it.

Other than the above, I also had a feature looking for hashtags, a feature for URLS, a feature for whether the shout is for rent, a feature that fires if the token is the name of a builder.

2 Gazetteers

I mined list of locations in Delhi NCR from the web, along with another list of Indian names and surnames. Then I created particular features that check whether the token occurs in any of the respective gazetteer. This again helped the model generalise better.