Since it's twitter dataset we have to look at specific features that would help us in this dataset. For example, adding tags for words that are capital, Punctuations, URL, hashtag or @(these are pretty frequent in twitter dataset, for example @ would show us a person or user), adding POS tags and also first letters of words are my assumption for improvement. Beside all of above features, I'm going to add more features from reading lexicon files.

From (architecture.museum) we can add sentence by sentence each separately in our list and then check if it can improve accuracy or not. But in each name(each line) we can see lots of words that can be misleading in finding name of museums and these kind of places, but we can use different method as well, by looking at the file we can see that some words are very frequent in this file, and their identical so by selecting them and checking them we can tag them as location, words like museum, park and gallery.

From (automotive.make) and (automotive.model) we can guess cars models and company, these names are pretty identical just by their first words, so just by selecting the first words in these files and checking that with our words we can tag them as car.

From (base.events.festival_series) and (award.award) we can guess awards and festival prizes and awards. Again, in this file if we read line-by-line many words can mislead us in finding proper tags for them. So again, we chose specific words to find if "EVENT" tag is proper or not, these words can be "prize", "award" and "festival".

From (Bigdict) I could not find specific area to categorize its words but it seems it's mostly about events such as festival, conferences and tours. Like above we are not going to use that file since it seems it cannot help us and it's pretty large file to process and add features from that to our prediction.

From (broadcast.tv_channel) we can find words like "tv" and "channel" which are identical and pretty have been repetitive in every sentence.

From (business.consumer_product) in this file products have been shown and brands name have shown with them as well. So it could be a good file for finding company's brand name. but we have to be careful about words that are not related to brands so we can remove them from lower 500 words. For example in this line "Olympus C-3040 Zoom" if we just checking word by word we will categorize the word "zoom" as company's name. so we have to remove all those words that are in frequent words and remain will be good enough for our task.

From (cvg.computer_videogame) I couldn't find any specific feature from this file, these video games words are not very identical, they are when we compare them with their neighbors.

From (education.university) we can see that this file has words like “university”, “college” and “school” are in every sentence and they are pretty identical and can show us some sense of location.

From (firstname.5k) we can add features for person and people tag since they are pretty identical in showing this kind of relation. We don’t have to add other files like firstname.10, firstname.100 and …

From (government.government_agency) we can see words like “federal”, “office”, “department”, “ministry” and “agency” are identical so by seeing those words we can add the tag “LOC”.

From (lastname.1000) we can add feature like all of them as “person”.

From (location.country) we can add feature like “location”

From (sports.sports_league) and (sports.sports_team) we can add features like “SPORT” for some words like “f.c.”, “fc”, “club” and “league”.

From (time.holiday) and (time.recurring_event) we can add features for “time” and “events”.