Feature Extraction

Considering that we are dealing with drugs we decided to analyze a bit how drug names are formed and to look at the text provided and based on this we observed the following:

* Special characters: many drug names have hyphens in it.
* Suffixes: they are added to the end of the generic name of a drug and they are used to provide additional information about the drug. For example, endings in ‘caine’, ‘ol’, ‘ine’ are very common.
* Prefixes: they are also used to provide additional information about the drug. For example, ‘anti’, ‘di’, ‘peni’ are very common.
* Word length: the name of the drugs tend to be quite long as well as the groups that can also have more than one word. Oppositely, brands are prone to have shorter names.
* Uppercase/Lowercase: in the data provided we observed that most of the times drug names are lowercase while drug brands are in uppercase. So, it would be good to check for lowercase and uppercase during feature extraction.
* Plural: many groups of drugs tend to finish with the letter ‘s’, like ‘opiods’, ‘tranquilizers’, ‘analgesics’, etc.
* Lookup tables: we are provided with two lookup tables. So, we decided to add this to our feature extractor. We look in both lookup tables and based on the label it has on them we add 4 new features which tell if it is a drug, drug\_n, brand or group.

### Adding suffix of length 3 to 2 words previous and after

First thing we did was to add the suffix of length 3 to the 2 previous and after words. This improved the F1 score of the CRF model ( to 57.3%) but it diminished the F1 score of the NB model by 2% (to 47.2%). After testing this feature with the other ones and seeing the results, we decided to not use this feature anymore. So, the following results are without using this feature.

1. Adding the length of the word and the previous word

We added as a feature the length of the word and the length of the previous word. This increased the baseline F1 score of the CRF model to 63.5% and the F1 score of the NB model to 51.5%.

1. Uppercase, lowercase and camelcase

Then three more features were added: one that tells if the word is in uppercase, another that tells if it is in lowercase and the last one that tells if it is in camelcase.

These features improved the F1 score of the CRF model to 73.2%, which is an increase of 9.7 %, and the F1 score of the NB model to 62% which is an increase of 10.5%.

1. Special characters

Two more features were added to check for special characters: one that tells if the word has dashes and another one that indicates if it has numbers in it.

With these two features we didn't see a massive improvement like with the previous one, but they did improve both models' F1 score. CRF F1 score improved to 75.6% and NB F1 score improved to 62.1%.

1. Plurals

We added a feature to tell if a word ends in plural. This did not improve neither of the F1 scores. CRF score remained the same and the NB score had a very small decrease of 0.1%.

1. Prefixes and suffixes

We made a list of the most commonly used suffixes with three, four, five, and six characters as well as a list of the most frequently used prefixes. We were able to achieve a small increase in the macro average F1 score Naive Bayes (0.2%) while the CRF score decreased by 1.2%.

We tried using only suffixes with three or four characters to see if the length of the suffix had an effect but the F1 scores remained almost the same as when using suffixes with more characters.

1. Lookup tables

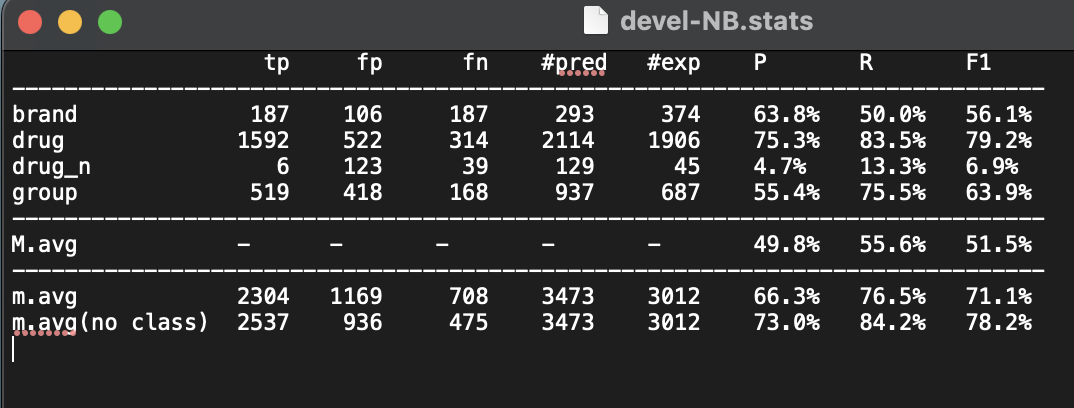
Finally, we added four new features which tell if it is a drug, drug\_n, brand or group based on both lookup tables. We first tried scanning the whole text file without using any auxiliary structure but it was extremely time consuming to the point where it needed more than half an hour to terminate. In order to overcome this, a dictionary was created. The results were clearly better.

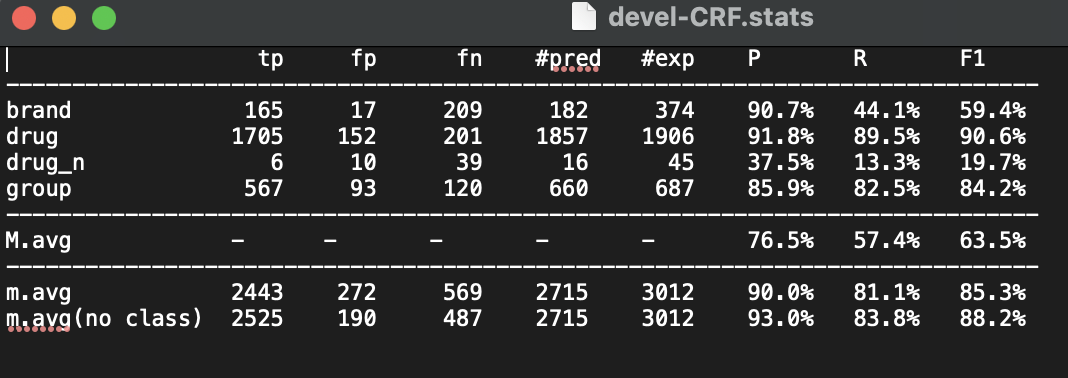
We tried adding these additional features taking into account the prefixes and suffixes and without taking them into account but the result was better when we did consider them.

As a result, we ended up with a F1 score of 64% for the NB model and a F1 score of 77.1% for the CRF model.

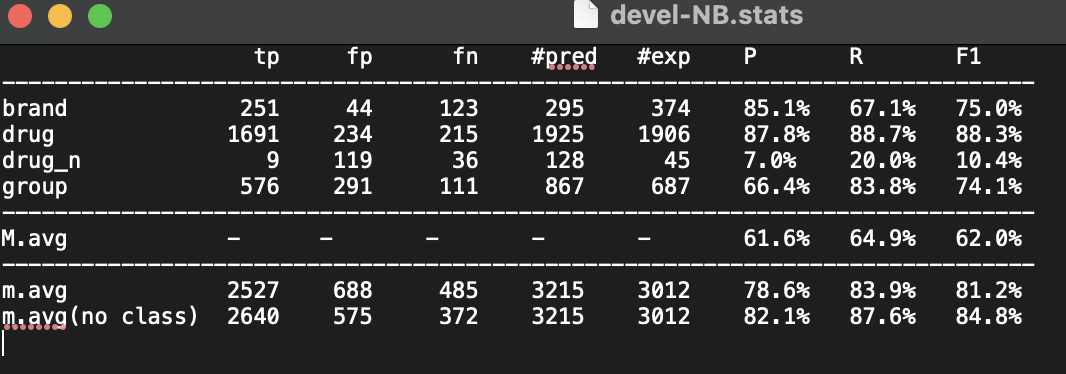
Annex:

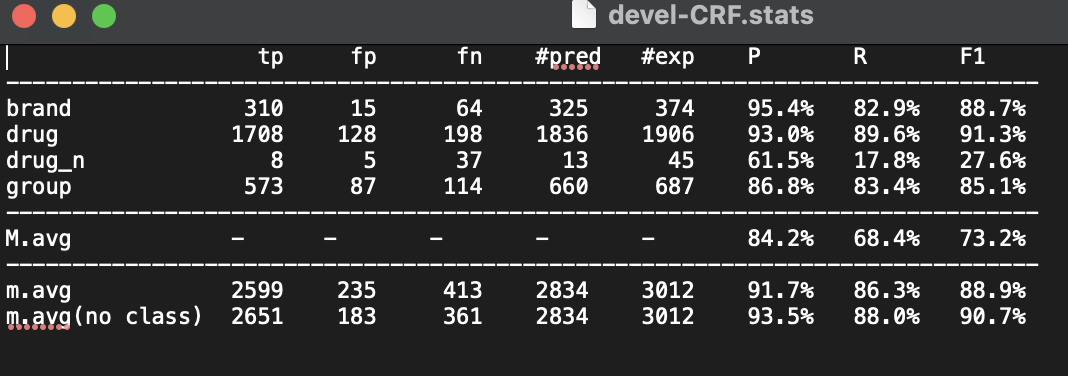
Adding the length of the word and the previous word



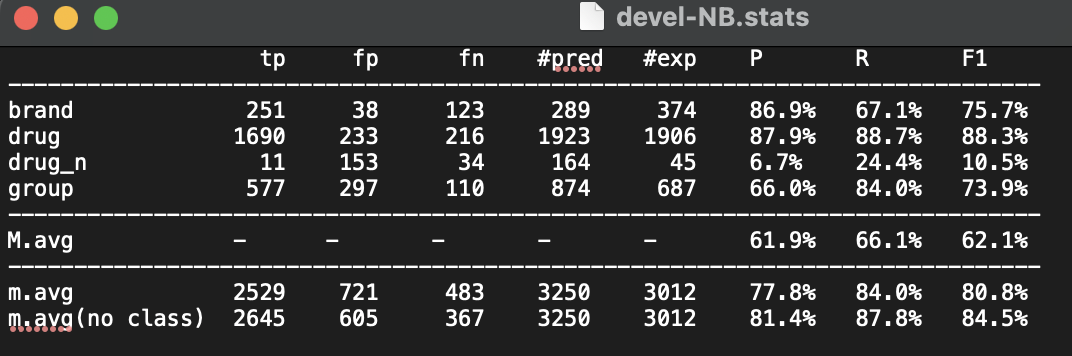


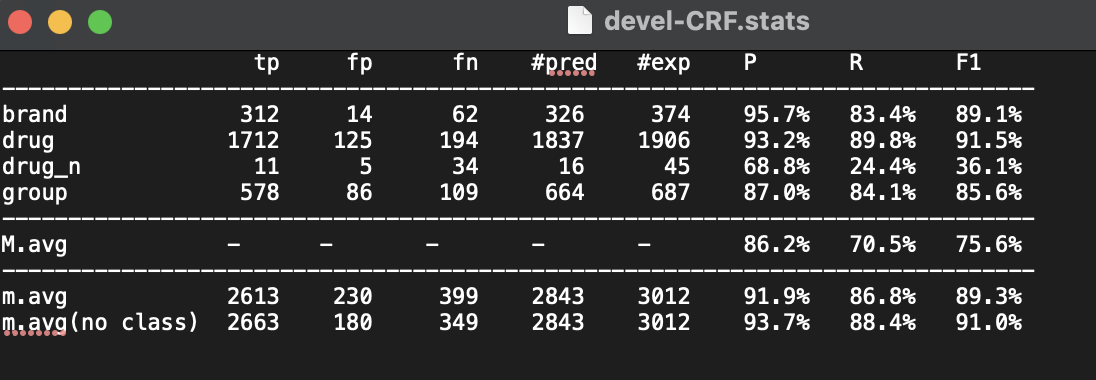
Uppercase, lowercase and camelcase



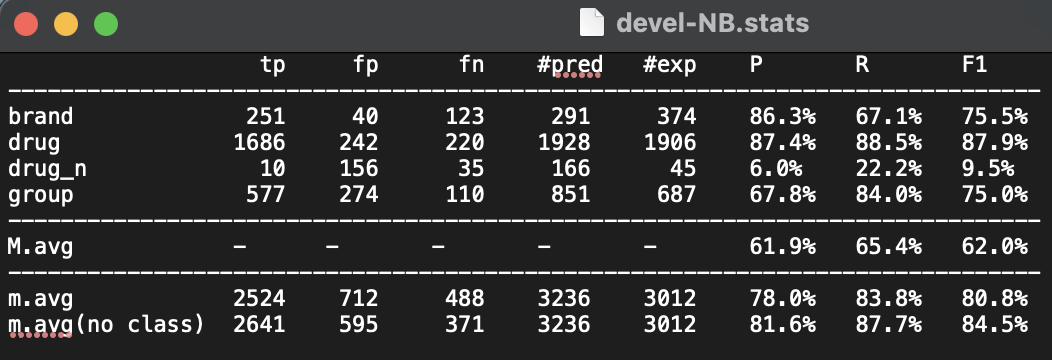


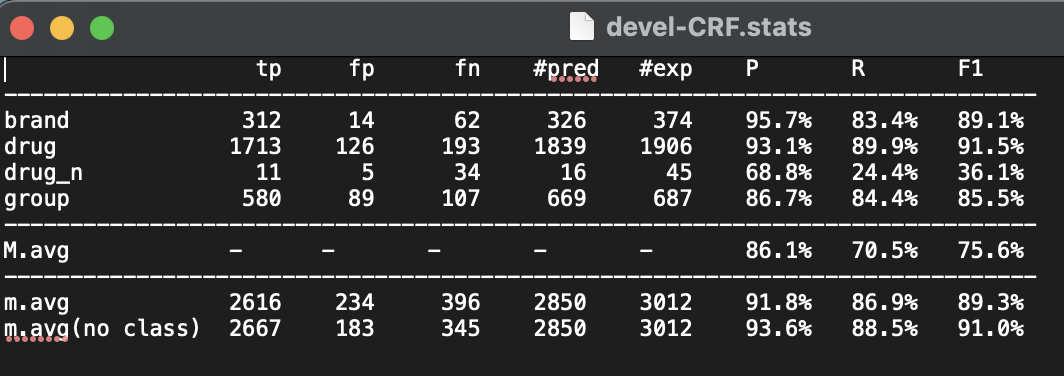
Special characters



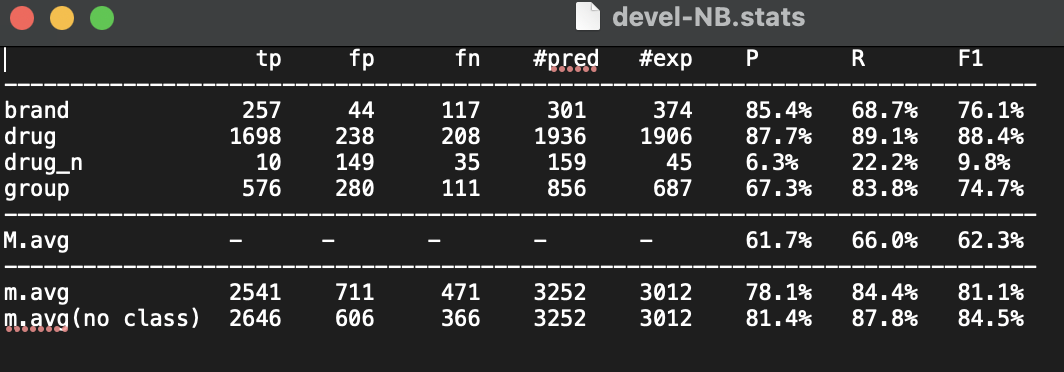


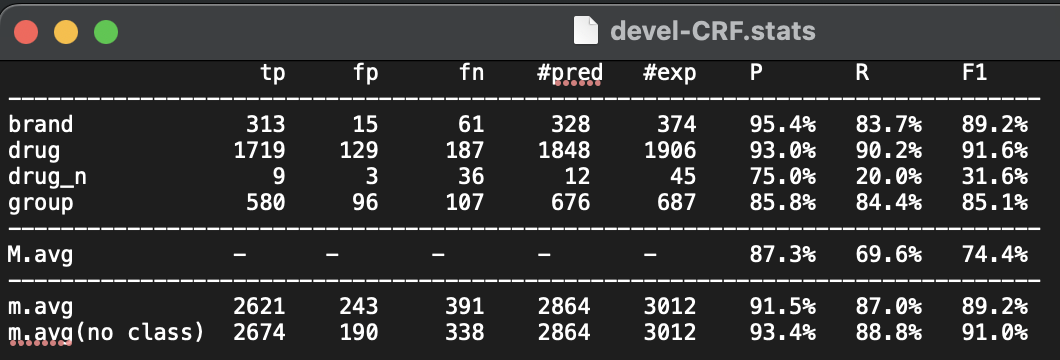
Plural





Prefixes and suffixes





Lookup

