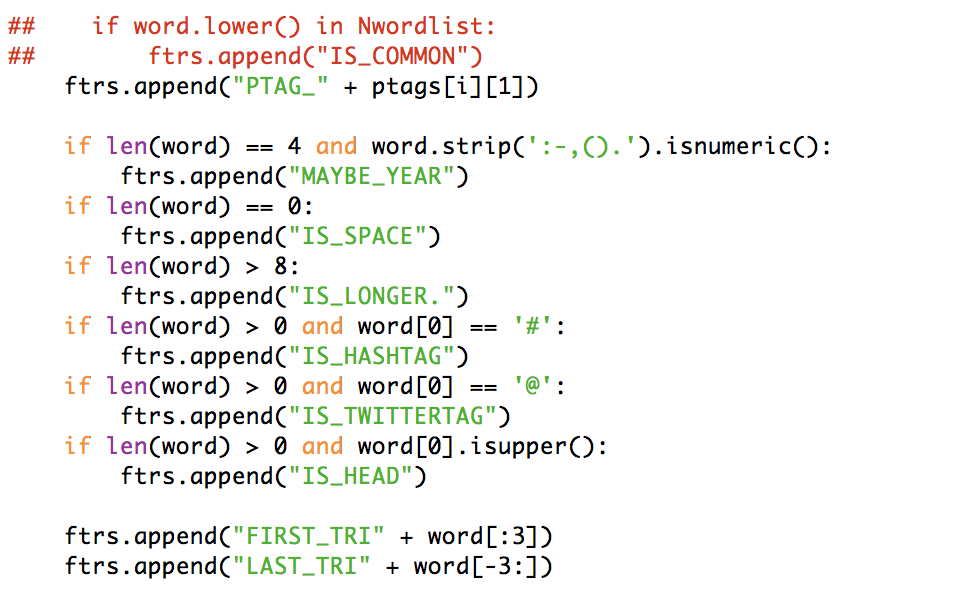
**CSCI 544 HW6**

**Qifan Wang**

**9369314001**

1.1.

I implemented a bunch of features, first let me explain some basic improvement.





here are some minor improvements I have implemented,:

year identify

space identify

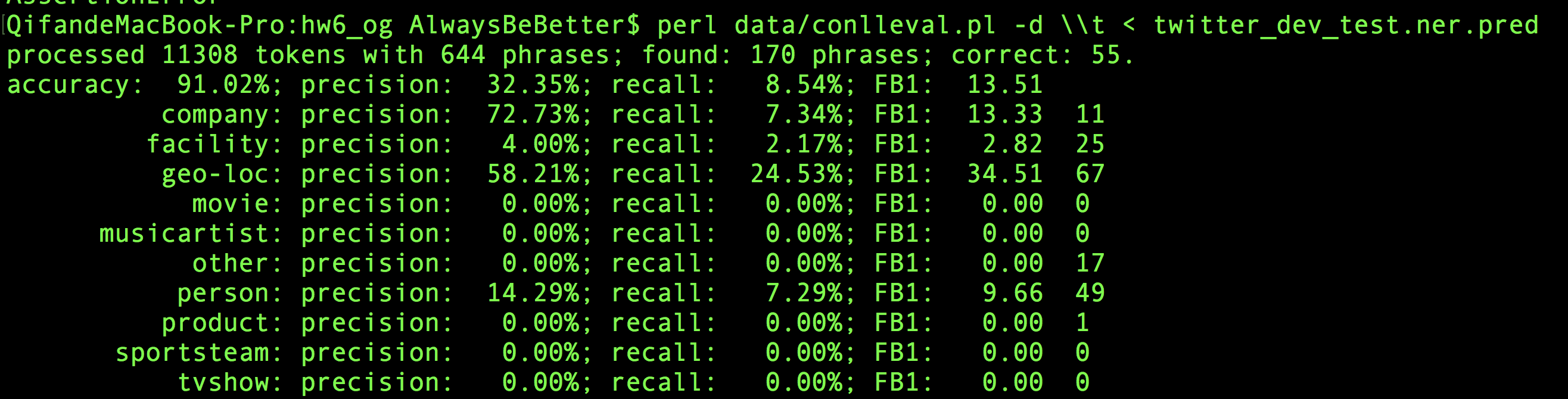
longer word identify

hashtag, twitter-tag, http identify (Since a lot of twitter use these tags to post something and some of them are links)

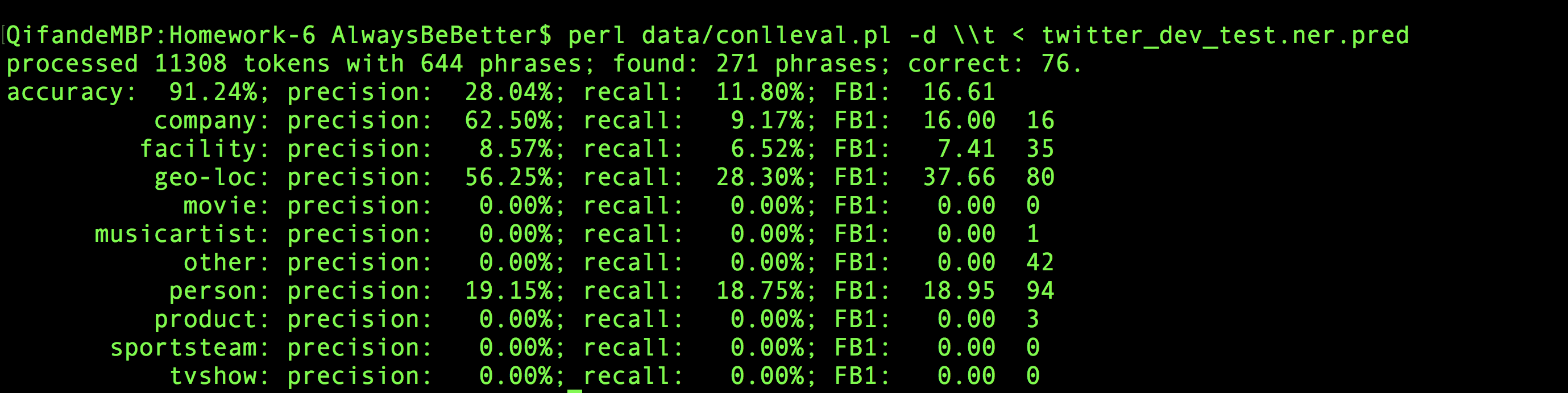
head of a sentence identify

some partial word letter identify

POS tag identify

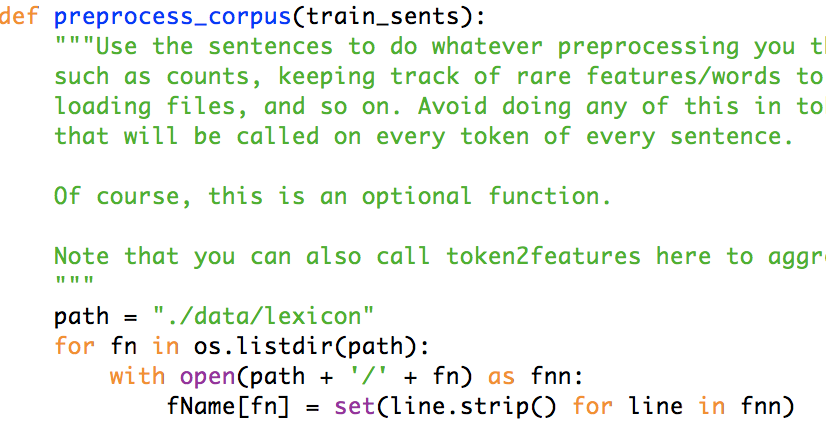


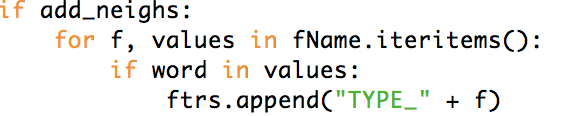
The original logistic regression on the test file gives about 13.51 score.



And after I add these feature, it improve up to 16.61, especially when I added word[:3] and word[-3:], it goes up quite a bit.

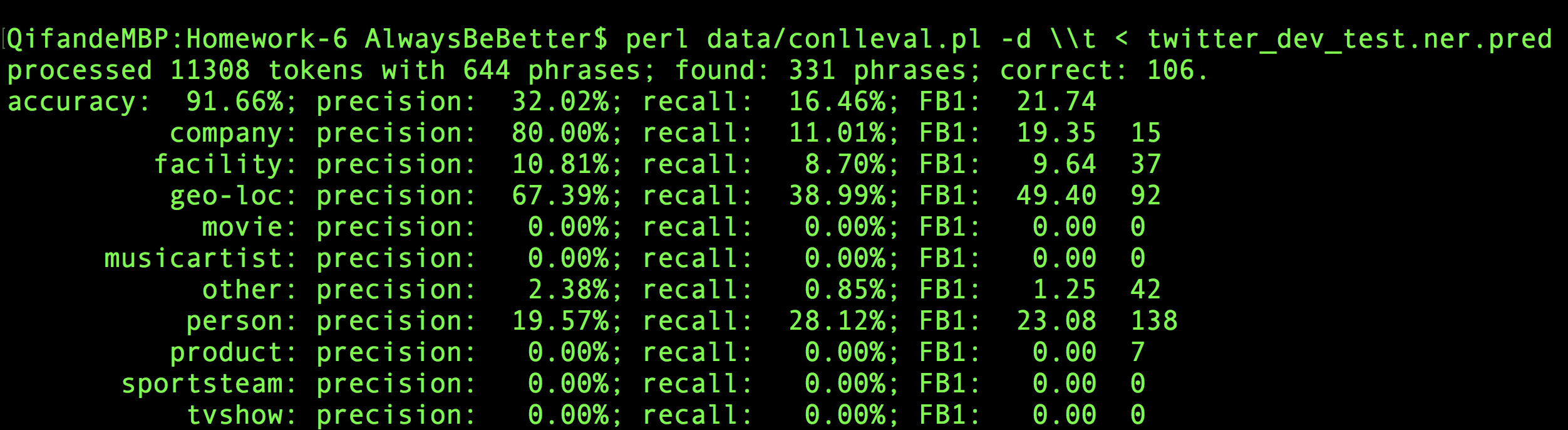
Also, the a very important part lexicons joined after I implement these basic feature improvement.





first of all, I just achieve some very simple goals, which are searching words inside the lexicons dictionary, if it exist, then tag added.

And here is what the result improvement looks like.



from 16 up to 21 almost 22.

However, I realize that there it might not be so simple to use lexicons, it would be too rough. So I choose to use some of them, such as last-name.5k, sport team and so on. But the result turned out is worse. After few tries, I chose to not modify them.