Naive.py

This machine learning code implements the naive-baeyers algorithm to distinguish between two articles of hockey and baseball.

The program outputs the correctness of the algorithm by implementing the k-fold cross algorithm in which k=5 means in the group of 399 or 398. In this first, I am training the computer on 4/5 of the data and then testing on 1/5. The program first calculate p(xi/y=hockey)(In code it distinct[word][0]) and p(xi/y=baseball)(In code it is distinct[word][1])

Important Points:

- words are stored as a dictionary distinct={key:[p(xi/y=hockey),p(xi/y=baseball)}.
- After this, we tested on 1/5 of the data.
- Smoothing is done by factor of dimension of X (read from the net)
- 97% chance is that my problem giving right answer.
- Used log to find probability.
- log(p(y=baseball/X)=sigmalog(p(xi/y=baseball))+log(p(y=baseball))-log(p(X)) and log(p(y=hockey/X)=sigmalog(p(xi/y=hockey))+log(p(y=hockey))-log(p(X)) as you can see that log(p(X)) is common in both so this is not calculated as while comparing these term got cancelled.