NLP assignment1

姓名：王涵 学号：2015211984

```
Keywords:
Forward Maximum Match, Backward Maximum Match, Bi-Direction Maximum Mactch
```

Problem 1

Intro:

The task on this assignment is basically build a Chinese word segmentation system on the PKU dataset and for this problem I only use the provided dataset to train and test the very model.

Input and Output

> - Input: the PKU dataset
> - Output: 3 indices to represent how well this model is doing:
>   Precision = (Number of words correctly segmented)/(Number of words segmented) * 100%
>   Recall = (Number of words correctly segmented)/(Number of words in the reference) * 100% F measure = 2*P*R / (P+R)

Source Code and Environment

Compatible both python2.7 & python3.5+

> benchmark.py
> fmm.py
> bmm.py
> bdmm.py

benchmark.py include a class Benchmark to collect data, seperate training set/validation set and compute the performance indices.

Algorithms and Models

In order to segment a sentence into separated words, there is some mature methods and algorithms and one of the most intuitive and fundamental algorithm is the Maximum Match alg.

There are generally 3 ways to perform the Maximum Match algortihm, listed as Forward Maximum Match(FMM),Backward Maximum Match(BMM) and Bi-Direction Maximum Match(BDMM). Take FMM as a example to illustrate how the algorithm works.

The basic idea of the MM is to try to find a segmentation with the maximum length that match with the word bank, which we collected and formed from the training set.In doing so, after we match over the sentence, a segmentation is formed.FMM is to find this segmentation from the beginning of the sentence and scan forward,which differs from BMM for it start from the end of the sentence.

However, there are experiments stands that BMM may be doing a little bit better than FMM in Chinese segmentation.

BDMM is an method to optimize between FMM and BMM.

> Fisrtly, perform both FMM and BMM segmentations, get f_seg and b_seg.
> Compare the number of words produced.
>
> > If equal choose from the two seg randomly.
> > If not, choose the one with less unit word.

In doing so,we select the best segmentation to match with the reference segmentaion.

Perfomances and Benchmarks

Thre are 3 indices to show how well the segmentation is doing.

> Precision = (Number of words correctly segmented)/(Number of words segmented)*100%.
> Recall = (Number of words correctly segmented)/(Number of words in the reference)*100%.
> F-measure = 2*P*R / (P+R)

FMM on PKU dataset

	Precision Rate	Recall Rate	F-measure Index
1	89.7842476842 %	94.4341003675 %	92.0504905934 %
2	89.8564022766 %	94.5874969399 %	92.1612719257 %
3	90.2318688982 %	94.6232037734 %	92.3753770369 %
4	90.0562664725 %	94.4566952046 %	92.2040081752 %
5	89.7349915632 %	94.3175321957 %	91.96921382 %
Average	89.9327553789 %	94.4838056962 %	92.1520723102 %

BMM on PKU dataset

	Precision Rate	Recall Rate	F-measure Index
1	90.1084831091 %	94.5519352922 %	92.2767481843 %
2	90.1896501639 %	94.6951046818 %	92.3874807376 %
3	90.2000930261 %	94.6222388414 %	92.3582627661 %
4	90.1767005335 %	94.6499528712 %	92.3591949359 %
5	90.2605521653 %	94.6779626698 %	92.4165006525 %
Average	90.1896946161 %	94.6394388713 %	92.3596374553 %

BDMM on PKU dataset

	Precision Rate	Recall Rate	F-measure Index
1	90.2362871778 %	94.6176000586 %	92.3750217927 %
2	90.2541751176 %	94.6777544895 %	92.4130587032 %
3	90.0801658545 %	94.5681842509 %	92.2696327047 %
4	89.9934628717 %	94.4120088794 %	92.1497994072 %
5	90.0533105829 %	94.6171731741 %	92.2788472633 %
Average	90.1234803209 %	94.5785441705 %	92.2972719742 %

Analysis and Conclusion

From the charts showed above, the 3 models perform almost equally on this dataset. We can see that BMM outrun FMM by a bit. But due to that they share the main idea and method.This diviation can be neglected.

However,BDMM,what is expected to perform the best, did not show a noticeable advance to other two models.Since BDMM is to optimize from two almost equally performed models, this medicore performance is understandable.

Note that BDMM does twice jobs as FMM or BMM does, I say the improvement does not worth the work.

I can't really draw any further conclusion based only on the PKU dataset.Can't wait to see how they perform in other datasets.