# Training the CRF Model - How To

1. Start with one raw file with raw title sentence, e.g. WD Blue 1TB SATA 6 Gb/s 7200 RPM 64MB Cache 3.5 Inch Desktop Hard Drive (WD10EZEX)

2. Tokenize the raw title sentence file into a sequence of keys using the tokenizer (The output should be similar to the following except that it should be in a text file)
   WD
   Blue
   1TB
   SATA
   6Gb/s
   7200
   RPM
   64MB
   Cache
   3.5
   Inch
   Desktop
   Hard
   Drive
   (WD10EZEX)

3. Label the sequence of words (The labeled format should be similar to the following except that it should be in a text file)

| | |
|---|---|
| WD | company |
| Blue | color |
| 1TB | size |
| SATA | misc |
| 6Gb/s | rate |
| 7200 | speed |
| RPM | speed |
| 64MB | misc |
| Cache | misc |
| 3.5 | dimension |
| Inch | dimension |
| Desktop | O |
| Hard | O |
| Drive | O |
| (WD10EZEX) | model |

4. Train the CRF model using the labeled input (necessary information regarding training could be found at https://nlp.stanford.edu/software/CRF-NER.shtml), command is similar to:

```
        java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop
harddisk/configfile.prop;
```

5. Use the trained model to produce the output file

   E.g. the command used should be similar to the following(this is the test file, meaning the file is already pre-labeled)

```
        java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
        harddisk/AmazonTraining50-model.ser.gz -testFile harddisk/BestBuyTest30.tsv
        -outputFormat tsv > harddisk/outputfiles/AmazonTraining50BestBuyTest30Output.tsv
        2>harddisk/outputfiles/MatricAmazonTraining50BestBuyTest30.txt;
```

To change from the test file into only produce the output, change the -testFile flag into -textFile