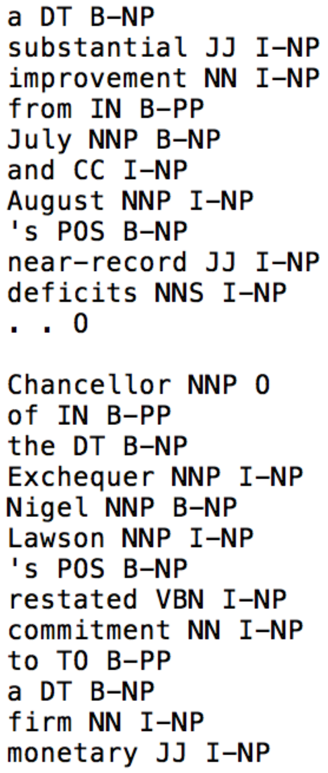
Part-of-Speech Tagging

(CMPSC 448; modified 9/13/2023)

Teams:

- <=5 students per team. You can build team across classes (001 & oo2). Each team votes for a team leader who is in charge of i) leading the project, ii) submitting the results, iii) presenting the project**.**
- Each team has a name
- The team leader sends TA the team name and team members by 9/20.

Training data: https://www.cnts.ua.ac.be/conll2000/chunking/train.txt.gz
Format of training file (as the following screenshot shows): Each row is for one token in the sentence; sentences are separated by an empty row. Three columns in total: token, POS tag, Chunking tag (we only use the first two columns for this midterm project)



```
a DT B-NP
substantial JJ I-NP
improvement NN I-NP
from IN B-PP
July NNP B-NP
and CC I-NP
August NNP I-NP
's POS B-NP
near-record JJ I-NP
deficits NNS I-NP
. . O

Chancellor NNP O
of IN B-PP
the DT B-NP
Exchequer NNP I-NP
Nigel NNP B-NP
Lawson NNP I-NP
's POS B-NP
restated VBN I-NP
commitment NN I-NP
to TO B-PP
a DT B-NP
firm NN I-NP
monetary JJ I-NP
```

Dev data: you can use a small part of training data as dev set.

Unlabeled Test data: will be released on 9/25

Requirements:

- The three algorithms you have to use:
    - use Bayesian Classifier for POS tagging
    - use Logistic Regression for POS tagging
    - use Support Vector Machines for POS tagging

- What you can use:
    - Features defined by you or other papers
    - Online packages such as NLTK, Pytorch, spaCy, Gensim, etc.
    - Combine above algorithms/models to get your "best model"

- What you should not use:
    - Pretrained word embeddings
    - Transformer-based pretrained language models, e.g., BERT, GPT3, ChatGPT, etc.
    - Any data other than the provided training data for tuning the model

What you need to submit (deadline 11:59pm on 10/7):

URL of your github repository, including

- **Labeled test data** by your best model: two columns (token, predicted_tag); TA will compute accuracy for each team. Filename "teamname.test.txt"
- **Code files** for the three algorithms: Bayesian Classifier, Logistic Regression, SVM

Email your TA the above URL.

Evaluation:

- **System performance (80%)**: each team gets **your_acc/max_acc_of_two_classes**
- **Presentation (20%)**: Each team presents in a few minutes. The following factors are considered for scoring: slides quality, the work you did (what features did you define, how models were optimized, what lessons/experience you have learned, what errors/issues you found, etc.)
- **Each team member gets the same score**.