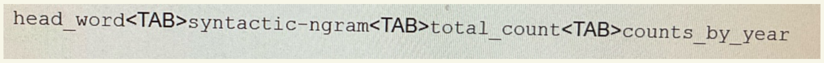
Assignment 3 FOR DUMMIES

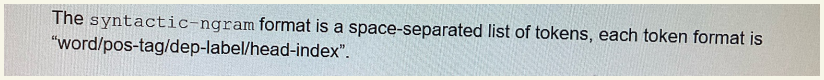
So we got NGRam corpus with a lot of sentences.

Each sentence format is:



head_word<TAB>syntactic-ngram<TAB>total_count<TAB>counts_by_year

and the syntatic-ngram format is:



The syntactic-ngram format is a space-separated list of tokens, each token format is "word/pos-tag/dep-label/head-index".

so we got tokens,

each token has 'word' (like 'likes'), 'dep-label' (like 'subject'), 'index' for the next word related to it (like 'dog').

So we will get from it "word1 word2 dep-label" and add to it "total count"

So step1 we will get all syntatic-ngram inputs, and their total count.

and output "word1 word2 dep-label total_count" for mapper

this is after we used stemmer on the words.

(stemmer takes word like "walked" | "walking" and change them to "walk")

and step1 reducer will combine them (see example in the end).

So there all kinds of dep-label. So in Step1A we collect all of the used types of dep-labels to use in step 2.

step1A ouput can be like "ccomp det pobj prep ..."

(dep-types used names)

In step2 we create for each words pair a vector.

the vector is the size of the amount of dep-labels types!

So if in step 1A we got 5 types, each cordinates will be the amount of apperances of that specific dep-labels.

"word1 word2 0 0 25 3 8"

mapper put all the types to the relevant key

reducer created the vectors.

In Step3 we take our vectors dep-label sizes. (from here I'm not sure)

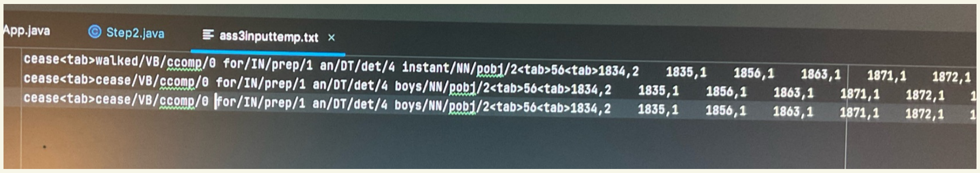
and calculate on that data all 24 equations to create a vector sized 24 to be used in the training model next?..

(plus, I'm not sure if the equation get the true data)

(I hate math)

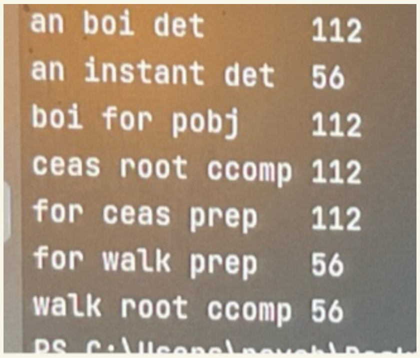
Example:

Step 1 input (ass3inputtemp.txt):



```
cease<tab>walked/VB/ccomp/0 for/IN/prep/1 an/DT/det/4 instant/NN/pobj/2<tab>56<tab>1834,2    1835,1    1856,1    1863,1    1871,1    1872,1
cease<tab>cease/VB/ccomp/0 for/IN/prep/1 an/DT/det/4 boys/NN/pobj/2<tab>56<tab>1834,2    1835,1    1856,1    1863,1    1871,1    1872,1    1
cease<tab>cease/VB/ccomp/0 for/IN/prep/1 an/DT/det/4 boys/NN/pobj/2<tab>56<tab>1834,2    1835,1    1856,1    1863,1    1871,1    1872,1    1
```

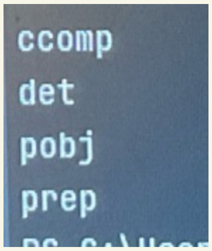
output:



```
an boi det          112
an instant det      56
boi for pobj        112
ceas root ccomp     112
for ceas prep       112
for walk prep       56
walk root ccomp     56
PS C:\Users\noyah\Desk
```

step 1A input

step 1A output:



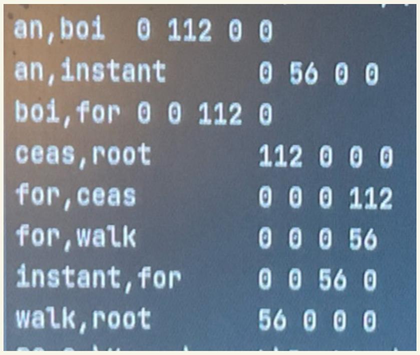
```
ccomp
det
pobj
prep
PS C:\User
```

(all dep-types)

(will be our vector)

step2 input

step 2 output:

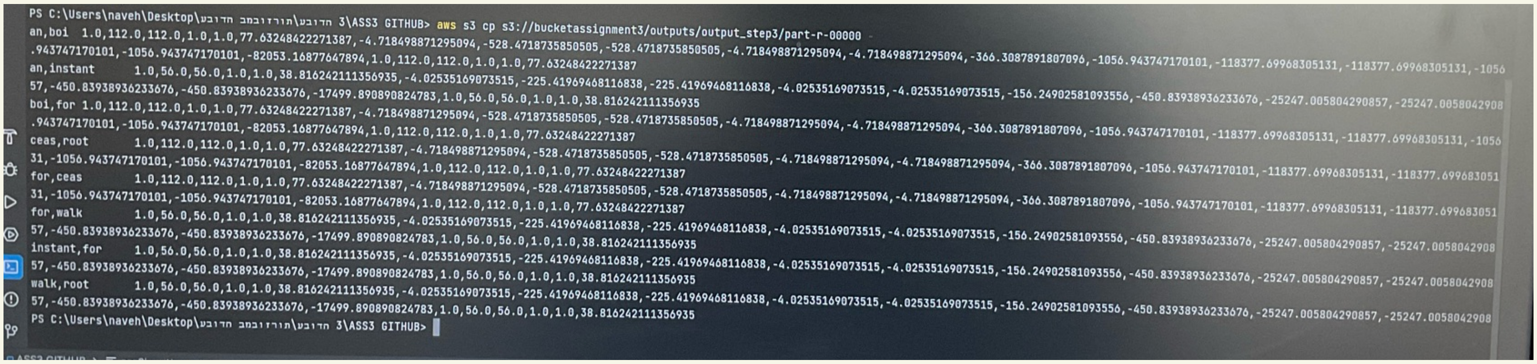


```
an,boi   0 112 0 0
an,instant      0 56 0 0
boi,for 0 0 112 0
ceas,root       112 0 0 0
for,ceas        0 0 0 112
for,walk        0 0 0 56
instant,for     0 0 56 0
walk,root       56 0 0 0
```

our vectors!

step3 input

step3 output:



```
PS C:\Users\naveh\Desktop\חדובע תורזובמב\חדובע 3\ASS3 GITHUB> aws s3 cp s3://bucketassignment3/outputs/output_step3/part-r-00000 -
an,boi  1.0,112.0,112.0,1.0,1.0,77.63248422271387,-4.718498871295094,-528.4718735850505,-528.4718735850505,-4.718498871295094,-4.718498871295094,-366.3087891807096,-1056.943747170101,-118377.69968305131,-118377.69968305131,-1056
.943747170101,-1056.943747170101,-82053.16877647894,1.0,112.0,112.0,1.0,1.0,77.63248422271387
an,instant      1.0,56.0,56.0,1.0,1.0,38.816242111356935,-4.02535169073515,-225.41969468116838,-225.41969468116838,-4.02535169073515,-4.02535169073515,-156.24902581093556,-450.83938936233676,-25247.005804290857,-25247.0058042908
57,-450.83938936233676,-450.83938936233676,-17499.890890824783,1.0,56.0,56.0,1.0,1.0,38.816242111356935
boi,for 1.0,112.0,112.0,1.0,1.0,77.63248422271387,-4.718498871295094,-528.4718735850505,-528.4718735850505,-4.718498871295094,-4.718498871295094,-366.3087891807096,-1056.943747170101,-118377.69968305131,-118377.69968305131,-1056
.943747170101,-1056.943747170101,-82053.16877647894,1.0,112.0,112.0,1.0,1.0,77.63248422271387
ceas,root       1.0,112.0,112.0,1.0,1.0,77.63248422271387,-4.718498871295094,-528.4718735850505,-528.4718735850505,-4.718498871295094,-4.718498871295094,-366.3087891807096,-1056.943747170101,-118377.69968305131,-118377.699683051
31,-1056.943747170101,-1056.943747170101,-82053.16877647894,1.0,112.0,112.0,1.0,1.0,77.63248422271387
for,ceas        1.0,112.0,112.0,1.0,1.0,77.63248422271387,-4.718498871295094,-528.4718735850505,-528.4718735850505,-4.718498871295094,-4.718498871295094,-366.3087891807096,-1056.943747170101,-118377.69968305131,-118377.699683051
31,-1056.943747170101,-1056.943747170101,-82053.16877647894,1.0,112.0,112.0,1.0,1.0,77.63248422271387
for,walk        1.0,56.0,56.0,1.0,1.0,38.816242111356935,-4.02535169073515,-225.41969468116838,-225.41969468116838,-4.02535169073515,-4.02535169073515,-156.24902581093556,-450.83938936233676,-25247.005804290857,-25247.0058042908
57,-450.83938936233676,-450.83938936233676,-17499.890890824783,1.0,56.0,56.0,1.0,1.0,38.816242111356935
instant,for     1.0,56.0,56.0,1.0,1.0,38.816242111356935,-4.02535169073515,-225.41969468116838,-225.41969468116838,-4.02535169073515,-4.02535169073515,-156.24902581093556,-450.83938936233676,-25247.005804290857,-25247.0058042908
57,-450.83938936233676,-450.83938936233676,-17499.890890824783,1.0,56.0,56.0,1.0,1.0,38.816242111356935
walk,root       1.0,56.0,56.0,1.0,1.0,38.816242111356935,-4.02535169073515,-225.41969468116838,-225.41969468116838,-4.02535169073515,-4.02535169073515,-156.24902581093556,-450.83938936233676,-25247.005804290857,-25247.0058042908
57,-450.83938936233676,-450.83938936233676,-17499.890890824783,1.0,56.0,56.0,1.0,1.0,38.816242111356935
PS C:\Users\naveh\Desktop\חדובע תורזובמב\חדובע 3\ASS3 GITHUB>
```

vector sized 24 (we can change to 6x4 later)

how train a model?

WHAT NOW?

(not sure)

I think we need to change the output format of Step3 to be like

$$
\begin{array}{c|cccccc}
 & e1 & e2 & e3 & e4 & e5 & e6 \\
\hline
\text{equation 1} & X & X & X & X & X & X \\
e2 & X & X & & & X & X \\
e3 & X & X & X & X & & \\
e4 & X & X & X & X & X & X
\end{array}
$$

4x6 (size 24)

To train a model with sized 24 matrix using WEKA.

And make sure Step3 is what we need to do + check the equations...

Written by:

Vaz Dias Hadas