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 "<comp 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!

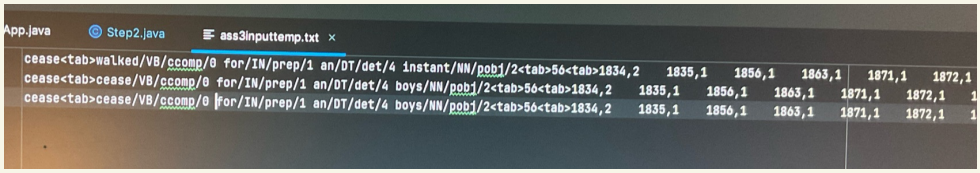
So if in step 1A we got 5 types, each cordinates will be the amount of appearances of that specific dep-labels.

"word1 word2 0 0 25 3 8"

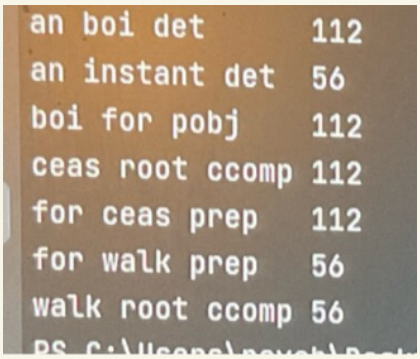
mapper put all the types to the relvant key reducer created the vectors.

Example:

Step 1 input (ass3inputtemp.txt):

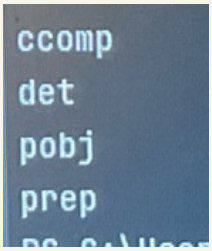


output:



step1A input

step1A output:

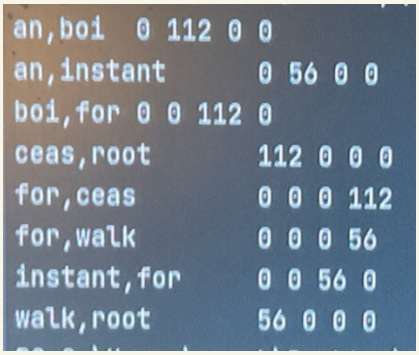


(all dep-types)

(will be our vector)

step2 input

step 2 output:



our vectors!

WHAT NOW?

(not sure)

I've added Equations.java with the requested equations.

I think we need to create new vectors will the calculations of the step 2 equations.

like

"word1 word2 L1 L2 Cosine Jaccard Dice Jensen" 1x6

somehow. ? or do it in the classifier?

After that using the word-relatedness.txt and the other 4 requested equations.

To train a model with sized 24 matrix

	e1	e2	e3	e4	e5	e6
equation 1	x	x	x	x	x	x
e2	x	x	x	x	x	x
e3	x	x	x	x	x	x
e4	x	x	x	x	x	x

(size 24)

idk im going to sleep.