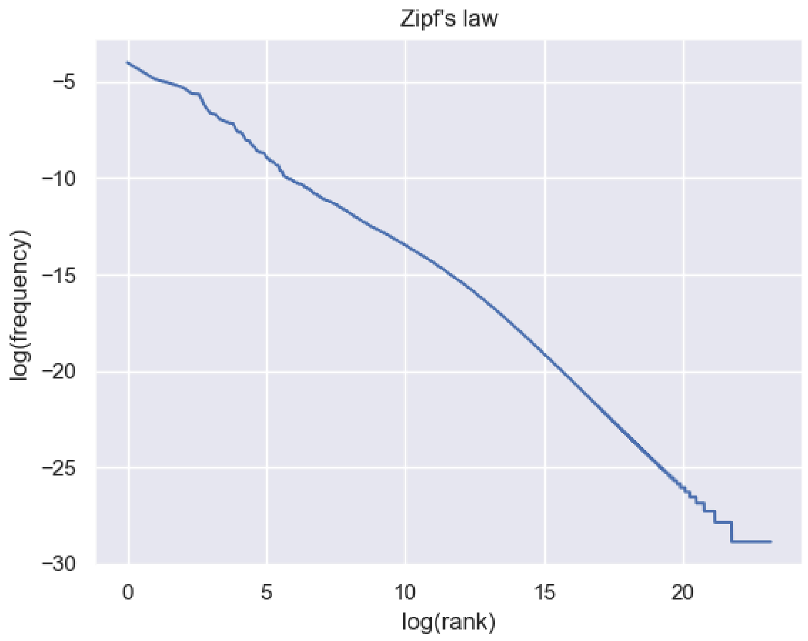
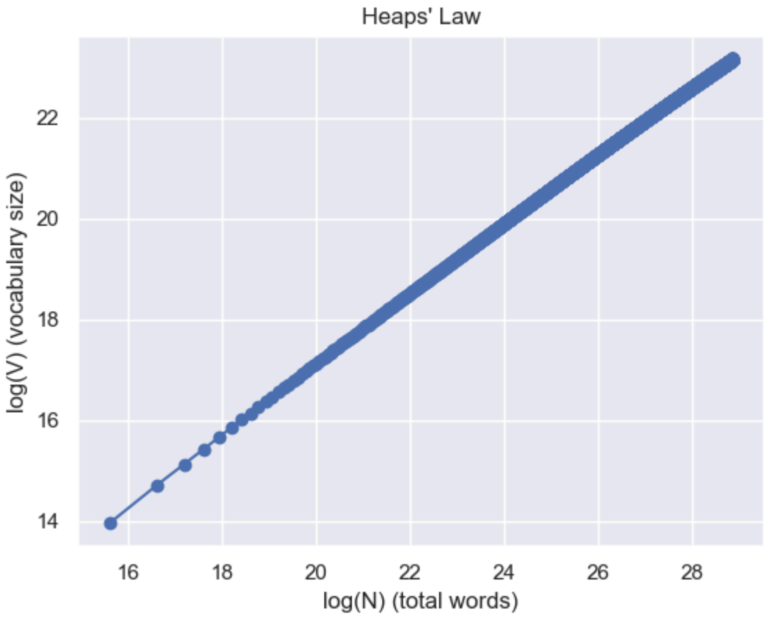
Assignment #1

Task #1: testing the Heaps' law in natural language





We checked whether our plot matches Heaps' law. According to the lecture, Heaps' law says V=k·(N^B) with typical values 0.67<B<0.75 and 10<k<100 for natural language.
:If we take logarithms on both sides we get

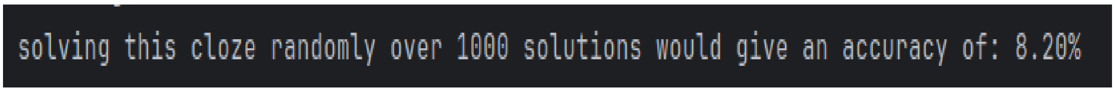
logV=logK +BlogN
.which is a straight line in log–log space, with slope B and intercept logK

In our plot of logV vs. logN we see an approximately straight, increasing line, which is exactly the pattern expected from this linear form. We did not estimate precise numeric values for K and B, but the overall shape and behavior of the curve are consistent with the ranges given .in class, so we conclude that our results are in line with Heaps' law

Task #2: statistical language models – solving a cloze

Chance Accuracy:

I estimated that when sampling a large amount of solutions, the mean accuracy will be around 1/n (where n is the number of candidates per blank). And that was the reality, as the mean accuracy from 1000 random solutions was 8.2% and then the number of candidates in our example was 12 (and 1/12 is around 8.3%). Thus, With multiple candidates per blank, random selection performs poorly.

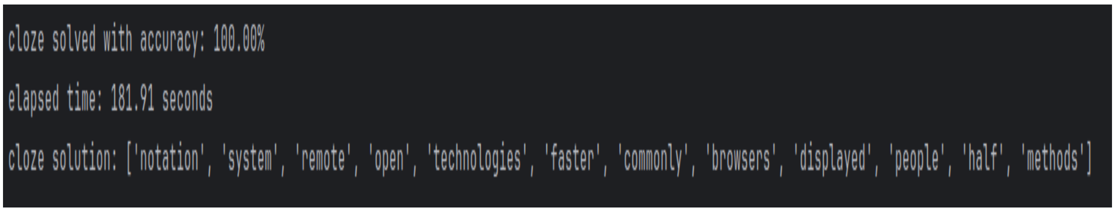


```
solving this cloze randomly over 1000 solutions would give an accuracy of: 8.20%
```

Comparison with My Solution:

My n-gram-based solution achieved 100% accuracy, a large improvement over chance. This shows that:

- The n-gram model captures meaningful linguistic patterns
- The scoring mechanism effectively distinguishes correct candidates
- The approach is far better than random selection



```
cloze solved with accuracy: 100.00%

elapsed time: 181.91 seconds

cloze solution: ['notation', 'system', 'remote', 'open', 'technologies', 'faster', 'commonly', 'browsers', 'displayed', 'people', 'half', 'methods']
```

Technical Details of My Solution:

- N-gram models up to 5-grams (bigrams through 5-grams)
- Laplace smoothing (add-k) to handle unseen n-grams
- Log probabilities for numerical stability
- Weighted combination favoring higher-order n-grams
- The difference (100% vs 8.2%) indicates the model uses contextual information rather than guessing. The low chance accuracy (8.2%) confirms the task is non-trivial and that the model's performance is meaningful.