

Brief report of Assignment 2

Victor Mai || DC127853 || Brief report of Assignment #2

2024/2025 CISC3025 Natural Language Processing

*I acknowledge the use of Grammarly to refine word usage and polish grammar to fulfill the academic English requirements.

I. Introduction:

- The goal of my project:
implement a bigram *n-gram language model* and evaluate the train and test set perplexity.
- N-gram models are the foundation of many NLP tasks, such as Machine Translation, Words Recognition, and Text Generation. Thanks to the simplicity and efficiency of n-gram models, n-gram models had become the foundation of language modeling.
- In this project, i am going to focus on words and bigram(2-gram) counting as well as calculating perplexity.

II. Background:

About *n-gram model* and *Markov assumption*:

- The *n-gram model* uses the chain rule to estimate the probability:

$$P(x_1, x_2, \dots, x_N) = P(x_1)P(x_2|x_1) \dots P(x_N|x_1, x_2, \dots, x_{N-1})$$

- For simplicity, we use the Markov assumption to approximate each component in the product:

$$P(w_1 w_2 \dots w_N) \approx \prod_i P(w_i | w_{i-w})$$

$$P(w_i | w_{i-w}) = \frac{\text{count}(w_{i-1} w_i) + n}{\text{count}(w_{i-1}) + n \cdot |V|},$$

V is the vocabulary of training data

About **Perplexity**:

- Perplexity is a metric used to evaluate the performance of a language model. It measures how well the model predicts a given sequence of words. Lower perplexity indicates better performance.
- By using add-n smoothing perplexity can be higher when n becomes larger.

$$PPL(w_1 w_2 \dots w_N) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_{i-1})}}$$

III. Challenge on calculating Perplexity:

According to the formular of perplexity to a sentence given, we can find that if the length of sentence is much longer than normal ones. the product of all fraction P can be close to 0. This situation can easily result in loss of precision during computer processing.

So, in the codes of my project, I tried to use logarithm to address the challenge as follow:

$$\begin{aligned}
\ln PPL(w_1 w_2 \dots w_N) &= \ln \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_{i-1})}} \\
&= \ln \left(\prod_{i=1}^N P(w_i | w_{i-1}) \right)^{-\frac{1}{N}} \\
&= -\frac{1}{N} \ln \prod_{i=1}^N P(w_i | w_{i-1}) \\
&= -\frac{1}{N} \sum_{i=1}^N \ln P(w_i | w_{i-1}) \\
PPL(w_1 w_2 \dots w_N) &= \exp \left[-\frac{1}{N} \sum_{i=1}^N \ln P(w_i | w_{i-1}) \right]
\end{aligned}$$

```

def add_n_perplexity(
    sentence:str,
    n:int,
    word_dict = None,
    bigram_dict = None):

    if word_dict is None:
        word_dict = read_word_count()
    if bigram_dict is None:
        bigram_dict = read_bigram_count()
    sentence = sentence_preprocess(sentence,
                                    word_dict)

    prev_word = ''
    v = len(word_dict)
    log_prob = 0.0 # log(1.0)
    for word in sentence:
        if word != '<s>':
            bigram = f"{prev_word} {word}"
            numerator = bigram_dict[bigram] + n
            denominator = word_dict[prev_word] + n * v
            log_prob += log(numerator / denominator)

```

```
prev_word = word
perplexity = exp(-log_prob / len(sentence))
return perplexity
```

- Replacing the product operation by summing and exponentiation does significantly to alleviate the loss of precision in real world testing and use.

IV. Results:

- For functionality of word counting, here includes both simple and complex samples:

SIMPLE one:

```
The sun rises and birds sing.
Birds sing as the sun rises.
The sun rises, and birds sing joyfully.
Birds sing joyfully when the sun rises.
When the sun rises, birds sing happily.
```

The result is:

```
<s> 5
the 5
sun 5
rises 5
and 2
birds 5
sing 5
. 5
<\s> 5
as 1
, 2
joyfully 2
when 2
happily 1
```

Complex input (around 43MB):

⚠ 文件太大: 42.92 MB。只读模式。

\$10,000 Gold?

SAN FRANCISCO – It has never been easy to have a rational conversation about the value of gold.

Lately, with gold prices up more than 300% over the last decade, it is harder than ever.

Just last December, fellow economists Martin Feldstein and Nouriel Roubini each penned op-eds bravely questioning the status quo. Wouldn't you know it?

Since their articles appeared, the price of gold has moved up still further.

Gold prices even hit a record-high \$1,300 recently.

Last December, many gold bugs were arguing that the price was inevitably headed for \$2,000.

Now, emboldened by continuing appreciation, some are suggesting that gold could be headed even higher than that.

One successful gold investor recently explained to me that stock prices languished for a more than a decade before the dot-com crash. Since then, the index has climbed above 10,000.

Now that gold has crossed the magic \$1,000 barrier, why can't it increase ten-fold, too?

Admittedly, getting to a much higher price for gold is not quite the leap of imagination that it seems.

After adjusting for inflation, today's price is nowhere near the all-time high of January 1980.

Back then, gold hit \$850, or well over \$2,000 in today's dollars.

But January 1980 was arguably a "freak peak" during a period of heightened geo-political instability.

At \$1,300, today's price is probably more than double very long-term, inflation-adjusted, average gold prices.

So what could justify another huge increase in gold prices from here?

One answer, of course, is a complete collapse of the US dollar.

With soaring deficits, and a rudderless fiscal policy, one does wonder whether a populist administration might take such a path.

And if you are really worried about that, gold might indeed be the most reliable hedge.

Sure, some might argue that inflation-indexed bonds offer a better and more direct inflation hedge than gold.

But gold bugs are right to worry about whether the government will honor its commitments under more extreme circumstances.

In fact, as Carmen Reinhart and I discuss in our recent book on the history of financial crises, *This Time is Different*.

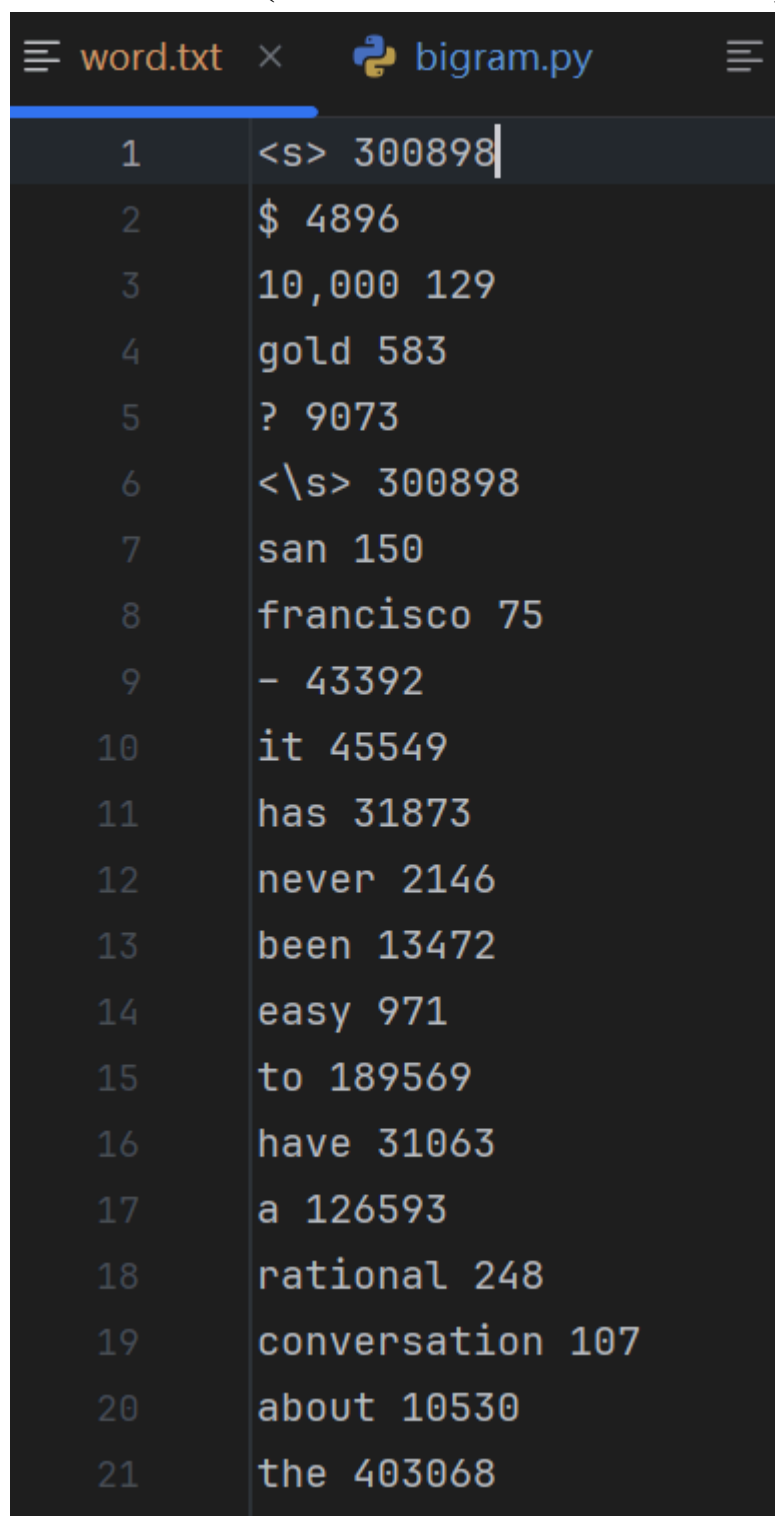
Even the United States abrogated indexation clauses in bond contracts during the Great Depression of the 1930's.

So it can happen anywhere.

Even so, the fact that very high inflation is possible does not make it probable, so one should be cautious in assuming that.

Some have argued instead that gold's long upward march has been partly driven by the development of new financial instruments.

The result is (Toke about 16.36 seconds):

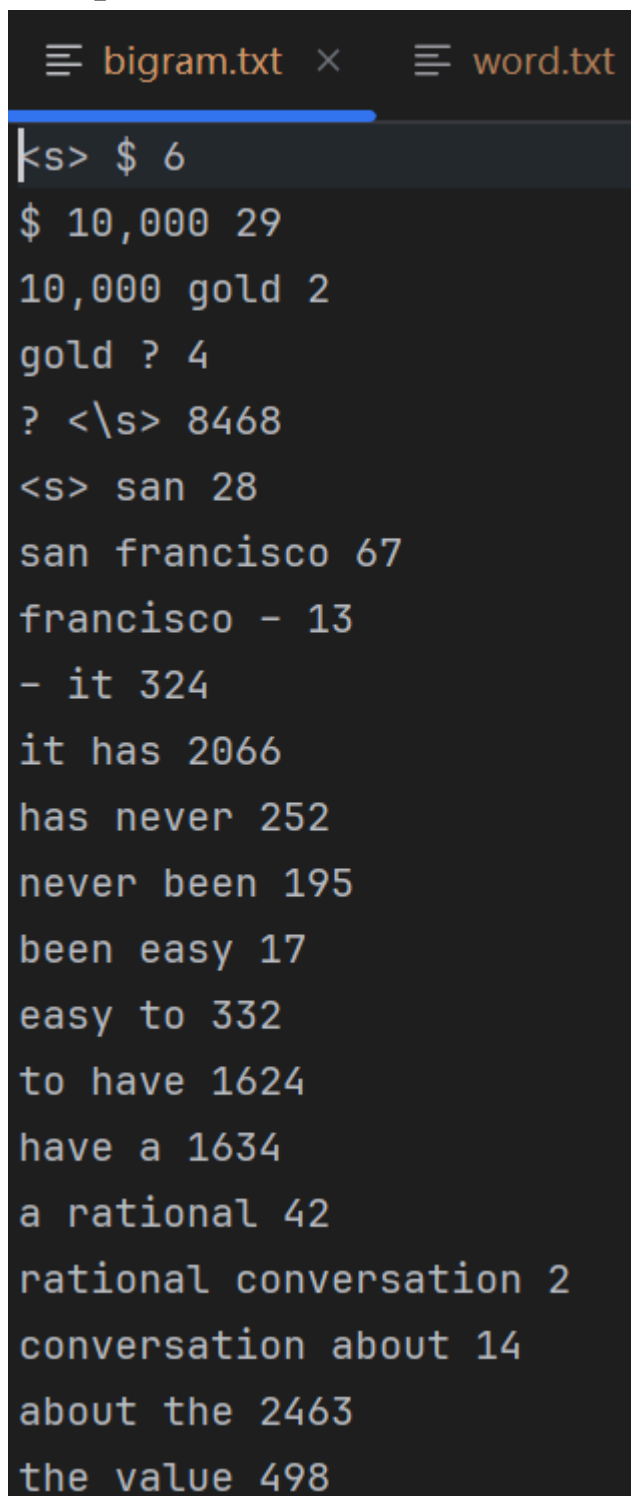


1	<s> 300898
2	\$ 4896
3	10,000 129
4	gold 583
5	? 9073
6	<\s> 300898
7	san 150
8	francisco 75
9	- 43392
10	it 45549
11	has 31873
12	never 2146
13	been 13472
14	easy 971
15	to 189569
16	have 31063
17	a 126593
18	rational 248
19	conversation 107
20	about 10530
21	the 403068

- For functionality of bigram counting, here show the case of complex sample:

Input is same as complex word counting above.

The partial result is shown as below (toke about 18.297 sec):



```
bigram.txt x word.txt
|s> $ 6
$ 10,000 29
10,000 gold 2
gold ? 4
? <\s> 8468
<s> san 28
san francisco 67
francisco - 13
- it 324
it has 2066
has never 252
never been 195
been easy 17
easy to 332
to have 1624
have a 1634
a rational 42
rational conversation 2
conversation about 14
about the 2463
the value 498
```

- For functionality of Perplexity calculating, here is the final output (perplexity of n=2 in add_n_smoothing):

Test-Set-PPL: 3516.82

Now, as one mordant and widespread slogan puts it, Austria is “the better Germany.” 2452.39

Many French writers also express a vivid feeling of national decline, and many ordinary citizens believe that the rules of the global economy work against France. 1718.48

In short, the large states still have illusions, encouraged by their political elites, about what the state can do to guide economic development. 1718.48

By contrast, the small European states – whether in Eastern or Western Europe – have been much more flexible in responding to the challenges of the modern world. 4783.81

Politics can offer all kinds of goods that voters find very attractive: tax breaks, subsidies, and social benefits. 4783.81

But small states are less likely to think that they can create the rules of the game, and accordingly they are more willing and able to make adjustments. 10

They are more keenly aware that if they try to redistribute too much, they will simply drive away the factors of production: capital will flow elsewhere, and the same sort of logic applies to the large states: France and Germany are losing skilled labor at the same time as they are drawing in cheaper labor from Eastern Europe. 20

The result is that they are feeling less French or German, creating greater political room, on both the left and the right, for nationalist resentment. 20

In the earlier rounds of European integration, from the 1950’s to the 1980’s, the big states received very obvious gains, which their politicians could easily exploit. 2317.78

But since 1989 – or since the Maastricht Treaty came into force in 1992 – the political dynamic has changed. 2317.78

It is now the EU’s smaller states that gain the most from wider and deeper European integration. 1192.62

If the large states are to obtain similar gains, and their politicians are to recover voters’ respect, their governments will have to accept the small-state model. 1192.62

Europe’s Smart Asian Pivot 1177.30

MADRID – For the first time in centuries, the focus of the global economy is shifting to the East. 522.85

The United States has commenced its “pivot” to Asia, and its relations with China, in particular, seem constantly to be flirting with Thucydides’s trap, the one that has doomed so many great powers. 2317.61

But, with the US and China regarding each another warily in the foreground of world affairs, where does Europe fit in? 2317.61

The European Union is at a critical historical juncture, one that demands its own pivot eastward – a coherent and decisive Asian strategy that builds on Europe’s strengths and addresses its weaknesses. 1869.51

Although the EU’s population is only one-fifth the size of that of China and India combined, and its military presence in Asia is minimal, its €12.6 trillion of economic power is formidable. 1869.51

This has not gone unnoticed by Asia’s governments, which are heavily dependent on economic growth to meet their young and growing populations’ demand for jobs. 1869.51

Currently, Asia is the EU’s main trading partner, surpassing North America and constituting one-third of its total trade. 1869.51

Trade with China alone is worth more than €1 billion per day, second only to trade with the US. 1400.07

Moreover, the EU has a somewhat paradoxical asset at its disposal: it is not a Pacific power and does not carry the burden of great-power status in Asia. 17

Far from being a weakness, this is precisely the source of the EU’s potential strength in Asia, for it provides a degree of diplomatic agility that the Americans lack. 2287.83

In attempting to execute its strategic pivot, the US is haunted at virtually every turn by its status as a historical hegemon, a military power, and the guarantor of the global order. 2287.83

Even when rebranded as a “rebalancing,” America’s eastward shift is inevitably met with suspicion by some Asian countries, particularly China. 2287.83

Europe, by contrast, can use its agility to perform a “smart pivot.” 1942.69

The EU must engage with Asia on at least three mutually beneficial fronts, with trade being the most important. 2273.88

The trade-liberalization agreements that the EU has in the pipeline with Asia’s vibrant economies (including South Korea, Singapore, Malaysia, India, Vietnam, and the Philippines) are a good start. 2273.88

As large-scale regional free-trade arrangements take shape, the EU continues to signal unequivocal commitment to free trade through sophisticated bilateral trade agreements. 1794.50

But trade flows are vulnerable. 1794.50

Of course, economic relations between states foster interdependence, decreasing the risk of conflict. 2222.46

But if conflict does erupt, the cost is far higher. 1773.43

When the territorial dispute between Japan and China over the Senkaku/Diaoyu islands flared anew last year, Japanese auto exports to China plummeted 80% in a matter of weeks. 1773.43

Herein lies Asia’s specific paradox: a high level of economic integration has not given rise to regional institutions that can support the stability required for such integration. 1081.68

Instead, Asia remains deeply scarred by unsettled disputes, periodic fits of nationalism, and contested borders, all of which tend to be amplified by apprehensions about the future. 1081.68

Regional integration is a recipe for long-term stability in Asia, and a second line of engagement for the EU. 1081.68

In Europe, once torn apart by war, armed conflict among EU member states is now almost unthinkable. 2193.42

V. Conclusion:

In this project, I successfully implemented a **bigram n-gram language model** and evaluated its performance using **perplexity** as the primary metric. The project focused on two main tasks: **word and bigram counting** and **perplexity calculation** with **add-n smoothing**.

I understand that there are still some place can have better optimization for the entire project. For example, running word counting and bigram counting together could reduce time consumption. Additionally, storing the results in JSON or NPY format instead of TXT could speed up the access time to the vocabulary in the `add_n_smoothing_batch` function.

All in all, this project is a great demonstration of the appeal of n-grams and the great potential of NLP, and has added a lot of interest to my learning in the field of NLP!