

Assignment 2

Steven Au
sttau
NLP 201

November 2023

1 Introduction

The goal is to create an n-gram model with OOV handling and then add linear interpolation. We are not allowed to use it outside the library.

The dataset consists of three token files divided into train, cross-validation, and test. We are only given guiding values on how to verify a working implementation for perplexity and interpolated perplexity.

The assignment is divided into three parts. Part 1 is preprocessing the training dataset and replacing tokens with less than three frequencies with `<UNK>`. Then replace tokens that are not in train vocabulary with `<UNK>`, in the dev and test set. Then calculate perplexity for each unigram, bigram, and trigram data set. Part 2 is calculating linear interpolation for three datasets while testing different lambdas for each n-gram. There are also small questions about interpolation perplexity when changing the training dataset. Part 3 is experimenting with Chat GPT 3.5 using different prompts and comparing the results to GPT 3.

2 Part 1

Starting from scratch, was very difficult as I needed to refer to lecture notes on constructing n-gram models properly. There was a lot of re-implementation of various codes and test prints. Thankfully, we were given the values for perplexity on the test string "HDTV ." so we can debug any issues with our functions.

My first approach was to wing it and create all the functions first for part 1 then test the string to see if my code was correct. Turns out, I wasn't really sure what the problem was asking for as I read the other components of part 1. I was testing my training data's perplexity and wasn't sure what number I was trying to match up. The more I read the problem the more confused I got. I tried printing my whole dataset's perplexities and was confused about what I was trying to match. I was able to get guidance from my peer Khin who explained the problem simply for me. All the time I spent playing around tweaking the code didn't help me reiterate the lists I kept calling.

My second implementation paid attention to how I structured my codes so I could use my n-gram pipeline on the test string and the dataset. This allowed me to modularize my code and test for bugs as each function was a different step in producing the perplexity. I began by making the n-grams into unigrams, bigrams, and trigrams and counting their respective frequencies in their own dictionary. Next, I pass the n-gram into their own functions to calculate their probabilities from the frequencies. I made each n-gram its own functions to track that each n-gram was being evaluated properly and set up with their edge cases. Last I have the perplexities call the probability functions so it can keep a running sum of the log probability. Having three perplexity functions allows me to see how I call my previous n-gram when calculating conditional probability and makes the complexity of the problem easier.

However, I was still getting errors for my perplexity and getting infinity for the test string. I decided to print all my values at every iteration to make sure was accessing my frequencies properly. I also did manual calculations to see how certain values affect perplexities to determine my mismatch values. After I solved that, I needed to make sure my perplexity scores were matching the given scores. I spent a couple of hours learning that unigram required a STOP token, and bigrams, and was calculated without the START token frequencies. Also, I learned that the bigram implementation for “START HDTV” did not apply to the STOP token as I was counting ” . <STOP>” in my calculations.

2.1 Perplexity Table

Data Type	Unigram Perplexity	Bigram Perplexity	Trigram Perplexity
String Data	658.045	63.708	39.479
Train Data	976.544	77.073	7.873
Dev Data	892.247	inf	inf
Test Data	896.499	inf	inf

Table 1: N-gram Perplexities for Various Data Sets

2.2 Debug Prints

Generated Unigrams: [(<START>,), ('HDTV',), ('.',), (<STOP>,)]

Generated Bigrams: [(<START>, 'HDTV'), ('HDTV', '.'), ('.', '<STOP>')]

Generated Trigrams: [(<PAD>, '<START>', 'HDTV'), (<START>, 'HDTV', '.'), ('HDTV', '.', '<STOP>')]

Unigram: ('HDTV',)

Numerator (Frequency of ('HDTV',)): 4

Denominator (Total words minus <START>): 1,622,905

Probability: 2.46×10^{-6}

Unigram: ('.',)

Numerator (Frequency of ('.',)): 60,949

Denominator (Total words minus <START>): 1,622,905

Probability: 0.0376

Unigram: ('<STOP>',)

Numerator (Frequency of ('<STOP>',)): 61,530

Denominator (Total words minus <START>): 1,622,905

Probability: 0.0379

Log Probability Sum: -19.5, Word Count: 3

Bigram: ('<START>', 'HDTV')

Numerator (Frequency of ('<START>', 'HDTV')): 1

Denominator (First Word Frequency): 61,530

Probability: 1.63×10^{-5}

Bigram: ('HDTV', '.')

Numerator (Frequency of ('HDTV', '.')): 1

Denominator (First Word Frequency): 4

Probability: 0.250

Bigram: ('.', '<STOP>')

Numerator (Frequency of ('.', '<STOP>')): 58,015

Denominator (First Word Frequency): 60,949

Probability: 0.952

Log Probability Sum: -12.5, Bigram Count: 3

Trigram: ('<PAD>', '<START>', 'HDTV')

Numerator (Frequency of ('<START>', 'HDTV')): 1

Denominator (First Word Frequency): 61,530

Probability: 1.63×10^{-5}

Trigram: ('<START>', 'HDTV', '.')

Numerator (Frequency of ('<START>', 'HDTV', '.')): 1

Denominator (Bigram Frequency): 1

Probability: 1.00

Trigram: ('HDTV', '.', '<STOP>')

Numerator (Frequency of ('HDTV', '.', '<STOP>')): 1

Denominator (Bigram Frequency): 1

Probability: 1.00

Log Probability Sum: -11.0, Trigram Count: 3

3 Part 2

Part 2 required me to handle the zero probability words and pass them as infinity when calculating the perplexity scores. Since I set everything in a pipeline I could call my previous functions to get the probability values. I need to change my alignment for the interpolated perplexity score based on how I did my preprocessing and handled the edge case if there were no corresponding n-grams. I cleaned up a lot of my code implementation as calculating the scores took a long time. I had a weird issue where I was processing part of the test string differently than my dataset. I fixed the implementation to read only tuples when passing through my pipeline. My next issue was how I was handling the Out-of-Vocabulary words, as my function needed to prepare the dev and test data.

If we split the training dataset in half there will be less of a distribution of the vocabulary meaning there are fewer reoccurring words. Therefore we will be replacing a lot more tokens with <UNK> and reducing our vocab size. There will be a lot more generalizing of OOV tokens and lowering the perplexity. If we didn't handle OOV words, the vocab would be too sparse, and there less probability representing of tokens and sequences therefore increasing perplexity. However, in our case, it lowers it for overfitting the <UNK> tokens and reducing the language model complexity since we process OOV tokens to UNK. The same happens when we increase the minimum frequency amount for OOV tokens. We will be converting more of our vocab to UNKs and lowering perplexity.

3.1 Interpolation Tables

Dataset	Interpolated Perplexity
String	48.114
Train	12.973
Dev	122.819
Test	164.900

Table 2: Interpolated Perplexities with Lambdas (0.3, 0.3, 0.4)

4 Part 3

Using OpenAI's playground environment for testing older ChatGPT models. Also used the SQuAD database for additional prompting. The response is highlighted to show what the. Questions or Answers are numbered if the response was regenerated to give a new question with the same prompt or a new answer given the same question and passage.

Data Type	Lambdas	Interpolated Probability
Test Data		
	(0.1, 0.3, 0.6)	10.233
	(0.2, 0.2, 0.6)	10.202
	(0.4, 0.3, 0.3)	15.463
	(0.3, 0.3, 0.4)	12.973
Dev Data		
	(0.1, 0.3, 0.6)	165.242
	(0.2, 0.2, 0.6)	147.457
	(0.4, 0.3, 0.3)	116.744
	(0.3, 0.3, 0.4)	122.819

Table 3: Interpolated Perplexity for Test and Dev Data Sets with Different Lambdas

4.1 GPT3 text-davinci-003

Comparing davinci-003 and davinci-002 since davinci-001 is not available on the platform

4.1.1 One-shot Learning

Read the passage and answer the question.

Passage: When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. Henry Ware was elected to the chair in 1805, and the liberal Samuel Webber was appointed to the presidency of Harvard two years later.

Question: Who succeeded Joseph Willard as president?

Answer: Samuel Webber

Read the passage and answer the question.

Passage: The 8- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United States. The megaregion’s area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana.

Question 1: What is the name of the region that is not defined by the

eight or 10 county definitions?

Answer: The Greater Southern California Megaregion.

Question 2: How many megaregions are there in the United States?

Answer: 11

Question 3: What is the name of the state that the megaregion expands to in the east?

Answer: Nevada.

4.1.2 Few-shot Learning

Read the passage and answer the question.

Passage: In the late 17th century, Robert Boyle proved that air is necessary for combustion. English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus.

Question: Who proved that air is necessary for combustion?

Answer: Robert Boyle

Question: John Mayow died in what year?

Answer: 1679

Read the passage and answer the question.

Passage: When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. Henry Ware was elected to the chair in 1805, and the liberal Samuel Webber was appointed to the presidency of Harvard two years later.

Question: Who succeeded Joseph Willard as president?

Answer 1: Samuel Webber

Answer 2: Henry Ware

Question: Who was the Hollis Professor of Divinity before Henry Ware?

Answer: David Tappan

Write a complete sentence

Question: Obama did not want to answer all of Trump's questions.
Trump

replied "You gotta give me more time or I will sue." To what did Trump respond when he said, "You gotta give me more time or I will sue?"

Response: Answers may vary. Four options are shown.

Responses #1, #3 and #4 are incorrect. "Once I am in the White House, I will have the time to answer all your questions."

What are some ways online banking is more convenient than visiting a bank branch? Select all that apply.

What are some ways online banking is more convenient than visiting a branch? Select all that apply.

In-person transaction to another person (e.g., transfer of funds, account openings, deposits, withdrawals, bill payments, and loan payments)

I want to move all my files over to my brother's computer, but I have lost the CD to the program that would help me. I can log on to my brother's computer and transfer the files over the Internet.

Buying and selling things (done through third parties, such as eBay

Read the passage and answer the question.

Passage: In the late 17th century, Robert Boyle proved that air is necessary for combustion. English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus.

Question: Who proved that air is necessary for combustion?

Answer: Robert Boyle

Question: John Mayow died in what year?

Answer: 1679

Read the passage and answer the question.

Passage: When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. Henry Ware was elected to the chair in 1805, and the liberal Samuel Webber was appointed to the presidency of Harvard two years later.

Question: Who succeeded Joseph Willard as president?

Answer: Samuel Webber

Read the passage and answer the question.

Passage: The 8- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United States. The megaregion's area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana.

Question: What is the name of the region that is not defined by the eight or 10 county definitions?

Answer: Southern California Megaregion

Question: How many megaregions are there in the United States?

Answer: 11

Read the passage and answer the question.

Passage: Within southern California are two major cities, Los Angeles and San Diego, as well as three of the country's largest metropolitan areas. With a population of 3,792,621, Los Angeles is the most populous city in California and the second most populous in the United States. To the south and with a population of 1,307,402 is San Diego, the second most populous city in the state and the eighth most populous in the nation.

Question: Which city is the most populous in California?

Answer: Los Angeles

4.2 GPT3 text-davinci-002

4.2.1 One-shot Learning

Question: What is the name of the state that the megaregion expands to in the east?

Answer 1: Nevada.

It is Novmber 14th and Earthquake Awareness Day in the state of California. In extremely rare occurences, an earthquake could occurence anywhere, California or West Midlands. Be prepared to escape any possible damages in the buildings.

Check if you have the safety steps covered in this interactive simulation!

Answer the question!

In the Moving Medium simulation, on which level do you usually start?

10. Short Answer (Answer in 150-200 words)

The 'Quake 2005' is entirely a free choice writing assignment, which you may explore on any topic listed below or any understanding after using webpages from the course resources provided.

8 Unique Prophecies From the 18th Street Gang Part 2

Questions and Discussion Answer: "The Spirit Who Motivates the Gang"
Part 3

Prediction Accuracy: Whistling at Night in the Jungle by Daniel W. Saterén

Restoring Structures in Wichita for the 2018 Aircraft Rescue and Fire-fighting Challenge

Nasty Mag, Beltane Issue, 2020

Answer the question!

"Not Yours . . . yet." People have unexpected dreams/fears, like the one above, like the one in the Watchcha video in the comment to evasive3 below,

Answer 2: Nevada.

A. New York State

B. Rhode Island

C. Texas

D. Nevada

Question: Definition/Structural Que

t'tion

It seems incredible that (0)—one in four adults in America were without regular health insurance in 2004.

Three important contributing factors appear to be responsible: (1)—'s failure properly to regulate the insurance marketplace (2)—with Americans being routinely denied insurance, typically as a result of "preexisting medical condition," demand for unprecedented medical care since (3)—medical technology evolved and improved.

Question: Underline the sentence which supports the transition of the idea from one passage to another.

Ans:

Visual Metaphor

In the second passage, (0)—the Southern California Megaregion (1)—to touch Los Angeles.

In the first passage, imagine a belt of mega-communities that (2)—from Houston to Seattle.

The line of demarcation is not (3)—set in stone; it has shifted, back and forth, depending on who is drawing the circles.

but instead a general area of growth that has been accumulating for

decades. That (4)—communities include: San Diego, Las Vegas; the Palm Springs-Ind

4.2.2 Few-shot Learning

Read the passage and answer the question.

Passage: The 8- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United States. The megaregion's area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana.

Question: What is the name of the region that is not defined by the eight or 10 county definitions?

Answer: Southern California Megaregion

Question: How many megaregions are there in the United States?

Answer: 11

Read the passage and answer the question.

Passage: Within southern California are two major cities, Los Angeles and San Diego, as well as three of the country's largest metropolitan areas. With a population of 3,792,621, Los Angeles is the most populous city in California and the second most populous in the United States. To the south and with a population of 1,307,402 is San Diego, the second most populous city in the state and the eighth most populous in the nation.

Question: Which city is the most populous in California?

Answer: Los Angeles

Question: Which city has the most population in the U.S. after New York City?

Answer: Los Angeles

Question: Which city is the second most populous in the nation?

Answer: San Diego

Read the passage and answer the following questions.

The recession of the early 21st century led to more suburban migration and a falling rate of poverty across the county, which reached its lowest levels since the 1970s. The county also attracted a record number of residents while losing poverty-stricken areas that were formerly rear-projects.

Question: What was the effect of the recession of 2000 on Los Angeles County?

Answer: It led to more suburban migration and a falling rate of poverty across the county.

Question: What is the lowest level of poverty reached during this period?

Answer: It is not mentioned in the passage.

Read the passage and answer the following questions.

Urbanization began in Los Angeles in the 19th century, when the town was founded in 1781 by Spanish settlers. The city boomed in the second part of the 20th century, exacerbated by the post-war "white flight" movement out of the central city. Los Angeles is known as the "City of Angels" meaning "City of Saints".

Question:

4.3 GPT 3.5

4.3.1 One-shot Learning

Read the passage and answer the question.

Passage: When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. Henry Ware was elected to the chair in 1805, and the liberal Samuel Webber was appointed to the presidency of Harvard two years later.

Question: Who succeeded Joseph Willard as president?

Answer: Samuel Webber

Passage: The 8- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United

States. The megaregion's area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana.

Question: What is the name of the region that is not defined by the eight or 10 county definitions?

Answer: The Southern California Megaregion.

Question: How many megaregions are there in the United States?

Answer 1: There are 11 megaregions in the United States.

Answer 2: 11

Question: What is the name of the state that the megaregion expands to in the east?

Answer 1: The greater Southern California Megaregion.

11.

Las Vegas, Nevada.

Answer2 : Las Vegas, Nevada.

Answer3 : Nevada.

Question: What is the name of the state that the megaregion expands to in the east?

Answer: The megaregion extends into Tijuana, Mexico.

4.3.2 Few-shot Learning

Performed the same as text-davinci-003

4.4 Homebrew Examples

4.4.1 GPT3 text-davinci-003

Text: "The French Revolution, which started in 1789, was a period of radical social and political upheaval in France. The absolute monarchy that had ruled France for centuries collapsed in three years. This led to the rise of Napoleon Bonaparte."

Prompt: "Summarize the above text in a concise paragraph."

Answer: The French Revolution began in 1789 and was a period of tremendous upheaval, resulting in the collapse of the absolute monarchy and the ascent of Napoleon Bonaparte. This period of intense conflict and dramatic change lasted for three years.

Text: "I recently purchased a new smartphone brand model X, and it's been an amazing experience. The battery life is incredible, and the camera quality is outstanding."

Prompt: "Determine the sentiment of the above text: positive, negative, or neutral."

Answer: Positive

4.4.2 GPT3 text-davinci-002

Text: "The French Revolution, which started in 1789, was a period of radical social and political upheaval in France. The absolute monarchy that had ruled France for centuries collapsed in three years. This led to the rise of Napoleon Bonaparte."

Prompt: "Summarize the above text in a concise paragraph."

Answer: Empress ~ words student chose /some of the monarch history/commentary:

"Judge ~ words student chose /some in-depth commentary:

I think using a rubric for student interns can communicate important parameters, while giving students freedom to write to the best of their ability. I think using sentences with a maximum word count in conjunction with words and verbs from the dictionary will force students to make sure verb tenses match and will help with word choice. The purpose of the assignment is validation of the student's work experience, thoughtful in-depth commentary, and concise paragraphs.

If students receive the above expectations, they have an opportunity to communicate in a clear and concise way about what they've accomplished on the project. I'm excited about this learning opportunity for the Peeps and goals we have for a future project together.

Next we will work on, Olivia will begin to write the text that will go under the pretty roofs.

Share:

Text: "I recently purchased a new smartphone brand model X, and it's been an amazing experience. The battery life is incredible, and the camera quality is outstanding."

Prompt: "Determine the sentiment of the above text: positive, negative, or neutral."

Answer: "This text is positive. Positive sentiment is assigned when the frequency of positive phrases is greater than the frequency of negative phrases."

In this case, the words "great," "wonderful," "need," and "amazing" were determined to be positive. Add this to the lexicon in Sentiment Analysis Recipe #1 above in Step 4.

Start the Loop

remet am nyt wat wew jus releashed an exctin new tec we are called ppi aak peapul to teest it . we neend fesazeed reespons kant wait for the feedbaak . rememfo we re cawld ppi !!!!!

...

4.5 Inference

Davinci-003 is really good at getting the task correct and reloading the prompt also generates correct answers with different wording. I only had issues with whitespace, it gave me extra answers from newline characters. It performed well on both one-shot and few-shot learning. Davinci-002 model gave extra information that was random and did not know when to stop the sequence. It got the answer right still, but generated way too much irrelevant information meaning that it did not handle the one-shot and few-shot learning methods that well. Overall, davinci-003 performed better at sentiment analysis, and question answering. Summarization ended up as the model rewording the sentence or paragraph word for word. I think summarization is the exception as both models performed the same and omitted the results from the report. Davinci-003 performed the same as ChatGPT 3.5.