



Santa's Digital Dilemma: Solving The Influx of Wishes with Language Model Magic

Author: Christoph Graczyk (ZIB)

Project: IOL

Challenge

Amidst the endless snow and twinkling stars of the North Pole, Santa Claus took a modern turn. To be more accessible to children globally, he set up a dedicated email inbox for their heartfelt Christmas wishes. However, crafting personalized responses to millions of emails seemed a stretch too far, even for Santa.

The elves, always eager to support Santa, convened an emergency council to solve the problem. The younger elves, like Elvin, advocated for the use of modern tech wonders, like Large Language Models. "Imagine," she exclaimed, "automating card-writing with the new computer we use for emails!" But Santa was hesitant, reluctant to lose the personal touch that defined his cards.

It was here that Elara, the oldest of the elves, suggested a middle ground. "Why not use something similar to a Bigram Model? It's the foundation of language generative tools. It can help select the introductory and concluding lines of the cards, based on our rich history of personalized cards. This way, Santa could still pen the main message. The model would only use Santa's own phrases as its vocabulary to generate the cards, instead of just the individual letters of the alphabet used in these models."

Elara's idea was simple. By analyzing past cards, they could predict how likely a phrase should appear given the previous phrases. This would retain the essence of Santa's messages while speeding up the process. To demonstrate, Elara delved into the archives, pulling out a random sample of 10,000 cards.

Each card Santa wrote typically had the same structure: an **opening line**, followed by his personal message, then a **closing wish** and a **goodbye phrase** before Santa's signature. Elara meticulously analyzed the samples from the archives. Following are some exemplary statistics for some specific phrases' frequency of usage

Opening lines:

- “As the winter wind whispers” - 220 occurrences.
- “Under the shimmering northern lights” - 180 occurrences.

Closing wishes: The choice of closing wish is depending only on the **given** opening lines:

- For cards that open with “As the winter wind whispers”:
 - Beginnings:
 - * “A snowflake echoing through the frosty air,” - 70 occurrences.
 - * “In every snowflake’s unique journey,” - 150 occurrences.
 - Endings:
 - * “a melody of joy and hope.” - 100 occurrences.
 - * “is the story of a thousand stars.” - 120 occurrences.
- For cards that open with “Under the shimmering northern lights”:
 - Beginnings:
 - * “As the fireplace crackles softly,” - 80 occurrences.
 - * “With the serenity of a winter’s night,” - 100 occurrences.
 - Endings:
 - * “may your heart be merry and light.” - 90 occurrences.
 - * “let warmth and comfort in your heart dwell.” - 90 occurrences.

Goodbye Phrases

- “From the winter wonderland” - 245 occurrences.
- “Yours in festive cheer” - 155 occurrences.

However, in his excitement, Elvin overlooked a critical aspect of the model. Unlike the actual Bigram Model, which would learn the probability of the phrases depending on the previous phrase, Elvin’s model was much simpler. His model selects phrases based on their overall frequency of usage with the additional criterion that the beginning of the closing wishes depend on the opening line. But, unlike Elara explained, his model treats the given endings of the closing wishes independently from the given beginnings of the closing wishes. This error could lead to the creation of unusual or even nonsensical combinations. For Santa the worst of them were the missing subject in the closing wish.

Question:

With the model based on the sample and considering Elvin’s omission, what is the probability that a card created by Santa Claus this year,

- starts with one of the given opening sentences,
- contains a grammatically incorrect closing wish where the subject is missing,
- and ends with one of the given farewell sentences?

Hint: In the imperative sentence "... let warmth and security reside in your heart," it can be assumed that a subject (implicitly) is given.

The two incorrect sentences are:

- "As the fireplace crackles softly, is the story of a thousand stars." ("What is the story?")
- "With the serenity of a winter's night, is the story of a thousand stars." ("What is the story?")

Possible answers:

1. 6.075%
2. 2.73375%
3. 0.0216%
4. 0.00972%
5. 0.243%
6. 0.436%
7. 0.78%
8. 10.45%
9. 3.34%
10. 0.89%

Project reference:

The IOL research lab is dedicated to exploring the intersection of mathematical optimization and machine learning, with a focus on developing innovative techniques for learning and optimization. By integrating these two fields, the IOL group aims to create new approaches to solving complex problems that leverage the strengths of both optimization and machine learning.