N-gram Language Models

Felix Borisov Timofei Simon

June 2, 2024

Abstract

This is a brief summary of your essay. It should be concise and informative.

Contents

1 Introduction

This is the introduction section where you provide background information on your topic and outline the structure of your essay.

2 Main Section 1 (FELIX)

2.1 Subsection 1.1

Here you can start discussing the details of your first main point. For example, algorithms are a fundamental part of computer science and understanding them is crucial for any software developer. As stated in [1], algorithms are essential for efficient problem solving in computing.

2.2 Subsection 1.2

Continue with further details and analysis related to your first main point.

3 Fortgeschrittene Konzepte und Techniken in N-Gramm-Modellen (Borisov Timofei)

3.1 Was ist un-seen N-Grams?

Da jeder Corpus begrenzt ist, fehlen darin zwangsläufig einige völlig akzeptable Wortfolgen. Das heißt, wir werden viele Fälle von vermeintlichen "zero probability n-grams" haben, die in Wirklichkeit eine Nicht-Null-Wahrscheinlichkeit haben sollten [2].

Betrachten wir die Wörter, die auf das Bigram aus Daniel Jurafskys Buch basieren. Ein Textkorpus zusammen mit ihrer Anzahl:

- denied the allegations: 5
- denied the speculation: 2
- denied the rumors: 1
- denied the report: 1

Aber nehmen wir an, unser Testsatz enthält Sätze wie:

- denied the offer
- denied the loan

P(offer—denied the) ist 0! P(loan—denied the) ist 0!

Diese Nullen bedeuten, dass wir die Wahrscheinlichkeit anderer Wortkombinationen stark unterschätzt haben, was die richtige Entscheidung der Anwendung, die auf unserem Modell läuft, stark beeinflussen könnte.

Das Problem heißt "Data Sparsity" [3]. Also kurz gesagt, das bedeutet, dass viele Daten in einem Datensatz fehlen oder auf Null gesetzt sind, sodass die meisten Zellen in einer Tabelle leer bleiben. Dies tritt häufig bei spärlichen Matrizen oder hochdimensionalen Datensätzen auf, wo nicht alle Elemente beobachtet oder erfasst werden. Was können wir also mit den Wörtern machen, die wir verwenden und die nicht in den Kontext unserer Trainingsdaten passen?

Die Methode zur Lösung dieses Problems heißt: Smoothing oder Discounting.

3.2 Smoothing Techniques

Further details and analysis related to your second main point.

3.3 Vergleich von N-Grammen und neuronalen Netzen

Further details and analysis related to your second main point.

4 Main Section 3 (SIMON)

4.1 Subsection 3.1

Discuss the details of your third main point.

4.2 Subsection 3.2

Further details and analysis related to your third main point.

5 Conclusion

Summarize the main points discussed in your essay and provide your final thoughts.

References

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms*, 3rd Edition, MIT Press, 2009.

[2] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd Edition, Pearson, 2023.

[3] Dremio, "Data Sparsity," `https://www.dremio.com/wiki/data-sparsity/`, Accessed: June 1, 2024.

[4] Author Name, "Article Title," *Journal Name*, Volume(Issue), pages, Year.

[5] Author Name, "Webpage Title," `http://example.com`, Accessed: Date.