CONTEXTO - A WORD ASSOCIATION GAME USING HNSW FOR ANN SEARCH

PROJECT PROPOSAL

Marco Lagos & Apoorv Walia

COMP 540: Statistical Machine Learning

Rice University

Fall 2023

Contents

ABSTRACT

This project proposal outlines the development of "contexto," an innovative word association game that harnesses the power of the Hierarchical Navigable Small Worlds algorithm (HNSW) for Approximate Nearest Neighbors (ANN) search. The goal is to create a dynamic and context-aware gaming experience, enriching word associations and fostering cognitive engagement.

Keywords Word Association Game · HNSW Algorithm · ANN Search · Dynamic Thinking · Context-Aware · Cognitive Engagement

1 Introduction

The primary challenge we aim to address in this project is the development of an engaging and educational word association game that harnesses the power of advanced techniques in Approximate Nearest Neighbors (ANN) search. Traditional word association games, while popular, often lack depth and context. By integrating the Hierarchical Navigable Small Worlds algorithm (HNSW) into our game design, we seek to imbue these games with a new level of dynamism and context-awareness. HNSW allows us to create a more immersive and interactive experience for players by dynamically connecting words based on their semantic relationships. This, in turn, enhances the player's ability to form meaningful associations between words, encouraging deeper cognitive engagement. The result is a word association game that not only entertains but also educates, fostering a deeper understanding of word relationships and improving vocabulary skills. By leveraging HNSW, we are poised to revolutionize the word association gaming experience, providing players with an unprecedented level of interactivity and cognitive enrichment.

2 Background

We will explore is the comparison of semantic spaces derived from word2vec embeddings and those built from lexical databases like WordNet and Moby Thesaurus. A study by Stanovová et al. (2018) [Reference 1] demonstrates how network science and graph theory tools can be effectively employed to compare these semantic spaces. Their work reveals interesting structural differences between human-built and machine-built semantic networks, providing insights into the contextual usage and perceived meanings of words. This research can inform our understanding of word associations and contribute to the design of more context-aware word association games.

Furthermore, we acknowledge the foundation laid by k-Nearest Neighbors (k-NN) as the precursor to more advanced ANN search techniques such as the Hierarchical Navigable Small Worlds algorithm (HNSW). While k-NN represents the traditional approach to nearest neighbor search, HNSW embodies a modern evolution of this concept. HNSW constructs a hierarchical graph structure, optimizing query times and scalability, which builds upon the principles of k-NN to provide enhanced efficiency and effectiveness in word association gaming. As we explore the nuances of word relationships and language diversity, we consider this evolutionary step from k-NN to HNSW as an integral part of our project's foundation.

3 Data

- **Common Crawl**: The Common Crawl dataset is a massive collection of web pages, freely available text corpora. https://commoncrawl.org/
- **Wikipedia Dump**: Wikipedia provides regular dumps of its entire content, which contain a vast amount of text from a wide range of topics. https://dumps.wikimedia.org/
- **Project Gutenberg**: Project Gutenberg offers a large collection of free eBooks, including literature classics, historical texts, and more. https://www.gutenberg.org/

4 Method

Certainly, let's expand on the "Method" section of your project proposal, specifically focusing on the algorithm selection and the use of the Hierarchical Navigable Small Worlds algorithm (HNSW) for Approximate Nearest Neighbors (ANN) search:

4.1 Algorithm Selection

The choice of the Hierarchical Navigable Small Worlds algorithm (HNSW) for ANN search is a pivotal decision in our project, driven by its robustness and efficiency in handling high-dimensional data, which aligns with the complex nature of word associations. HNSW represents a breakthrough in the field of ANN search, known for its exceptional speed and accuracy. Unlike traditional k-NN search algorithms, HNSW constructs a hierarchical graph structure, allowing for faster and more effective nearest neighbor queries.

4.2 HNSW Algorithm

At its core, HNSW works by partitioning data points into clusters and building a multi-level graph structure that connects these clusters. It optimizes the search process by guiding queries through this hierarchy, efficiently narrowing down the search space. HNSW's unique properties make it suitable for our word association game, as it can rapidly identify and link words with similar semantics, contributing to a more engaging and context-aware player experience.

4.3 Existing Implementations

While HNSW is a relatively advanced algorithm, there are existing implementations available in popular machine learning libraries, such as Faiss and NMSLIB. These libraries

provide efficient implementations of HNSW for large-scale ANN search tasks and offer various parameters for customization.

4.4 Our Approach

In our project, we plan to leverage these existing implementations of HNSW to streamline the ANN search process within the word association game. By incorporating well-established libraries, we can focus on the game's core design and the integration of word associations, rather than reinventing the wheel. We will provide URLs to reference implementations from these libraries to ensure transparency and facilitate replication of our work.

4.5 Modifications and Adaptations

To tailor HNSW specifically for the word association game, we anticipate making certain modifications and adaptations. These adaptations may include fine-tuning the algorithm's parameters to accommodate the unique characteristics of word associations. For instance, we may adjust the dimensionality of the data space, optimize the graph structure for word relationships, and implement custom scoring mechanisms to assess the relevance of associations.

Our goal is to enhance the algorithm's performance in capturing meaningful word associations while maintaining computational efficiency. These adaptations will be crucial in achieving the game's objectives of fostering engaging and context-aware associations among words, providing players with a rich and immersive experience."

This expanded section provides a more detailed explanation of why HNSW was chosen, how it works, and how existing implementations will be utilized and adapted for the word association game. It emphasizes the efficiency and effectiveness of HNSW in handling word associations and sets the stage for the implementation phase of the project.

5 Evaluation

5.1 Qualitative

We will gather user feedback through surveys, interviews, and user testing sessions. Participants will be asked to provide insights into their experience with the game, highlighting aspects such as engagement, enjoyment, and the educational value of word associations. Qualitative feedback will help us understand how the game enhances word association

experiences and whether it achieves our objectives of making associations more dynamic and context-aware.

5.2 Quantitative

Quantitatively, we will define performance metrics to measure various aspects of the game's performance. These metrics may include response time, accuracy of the algorithm's suggestions, and user engagement statistics such as session duration and repeat play. These quantitative measures will provide valuable insights into the game's efficiency and effectiveness in facilitating word associations.

5.3 Other metrics

If applicable, we will compare the results and performance of our "contexto" game with those of traditional word association games. Criteria for comparison may include the quality of associations formed, the speed of gameplay, and the overall user experience. This comparative analysis will help us determine the extent to which HNSW-enhanced word associations provide more insight into word associations.

6 Conclusion

This project endeavors to create 'contexto,' an innovative word association game that leverages the power of the Hierarchical Navigable Small Worlds algorithm (HNSW) for Approximate Nearest Neighbors (ANN) search. By merging advanced techniques in ANN search with engaging gameplay, our goal is to enrich word associations through the fusion of theory and practical implementation. We aim to demonstrate the potential of HNSW in enhancing word association games while contributing to the broader understanding of word relationships. We anticipate that this project will not only entertain but also educate, offering players a novel perspective on language and semantics.

7 References

- Veremyev, A., Semenov, A., Pasiliao, E.L. et al. Graph-based exploration and clustering analysis of semantic spaces. Appl Netw Sci 4, 104 (2019). https://doi.org/10.1007/s41109-019-0228-y
- Nebojša D. Grujić & Vladimir M. Milovanović (2023) Associative Word Relations in Natural Language Processing, Applied Artificial Intelligence, 36:1, https://doi.org/10.1080/08839514.2022.2034262

- Facebook AI Research. (Year, Month Day of the latest update). Faiss: A library for efficient similarity search. Faiss Documentation. Retrieved from https://faiss.ai/index.html
- Pinecone.io. (n.d.). HNSW - Hierarchical Navigable Small Worlds. Pinecone.io. Retrieved from https://www.pinecone.io/learn/series/faiss/hnsw/