Building a Vector Space Model for Vietnamese Using Word2Vec and FastText

Nguyen Viet Bac

July 2024

Abstract

In this paper, we present a comprehensive approach to building a vector space model for the Vietnamese language using Word2Vec and FastText algorithms. We utilize the Vietnamese Wikipedia corpus (viwiki) to train the models, which consists of 2,175,551 samples. The preprocessing steps include HTML tag removal, sentence splitting, word tokenization, and stopword removal. We provide a comparative analysis of the performance of Word2Vec and FastText and visualize the vector space for a set of selected words.

1 Introduction

Vector space models are essential for various natural language processing (NLP) tasks. This study focuses on developing such a model for Vietnamese, leveraging the extensive data available from Wikipedia. We use Word2Vec and FastText due to their effectiveness in capturing semantic meanings of words.

2 Data Preparation

The dataset used in this project is the Vietnamese Wikipedia dump dated 2024-02-01. It includes articles, templates, media/file descriptions, and primary meta-pages, provided in multiple bz2 streams, with 100 pages per stream.

3 Data Preprocessing

The preprocessing steps are crucial for cleaning and structuring the data appropriately for model training.

- **HTML Tag Removal:** We remove redundant HTML tags using regular expressions.
- **Sentence Splitting:** Sentences are split based on punctuation marks.

- **Word Tokenization:** Sentences are further tokenized into words.
- **Normalization:** We remove diacritics to simplify the words.
- **Stopword Removal:** Common stopwords are removed to focus on meaningful words.

4 Model Training

We trained two models: Word2Vec and FastText.

4.1 Word2Vec

Word2Vec generates word vectors based on the context of words in the corpus. However, it has limitations in handling out-of-vocabulary words.

4.2 FastText

FastText, an extension of Word2Vec developed by Facebook in 2016, addresses the limitations of Word2Vec by breaking down words into n-grams. For instance, "apple" is broken down into "app," "ppl," and "ple," and the word vector for "apple" is the sum of these n-grams. This approach handles rare words effectively.

5 Results and Discussion

5.1 Word2Vec Results

We tested the trained Word2Vec model to find the most similar words to "mat_troi" (sun). The results are as follows:

- anh_sang (light): 0.828
- trai_dat (earth): 0.768
- quy_dao (orbit): 0.760
- be_mat (surface): 0.756
- moc (rise): 0.754
- loi (core): 0.746
- vat_the (object): 0.743
- bau_troi (sky): 0.736
- ban_kinh (radius): 0.732
- thang_dung (upright): 0.730

5.2 Visualization

We visualized the vector space for the following words: toi (I), dat_nuoc (country), quoc_gia (nation), giau_sang (wealth), ngheo (poor), phu_nu (woman), con_gai (girl), dan_ong (man), con_trai (boy), tien_bac (money), thoi_gian (time). The results are depicted in Figure 1.

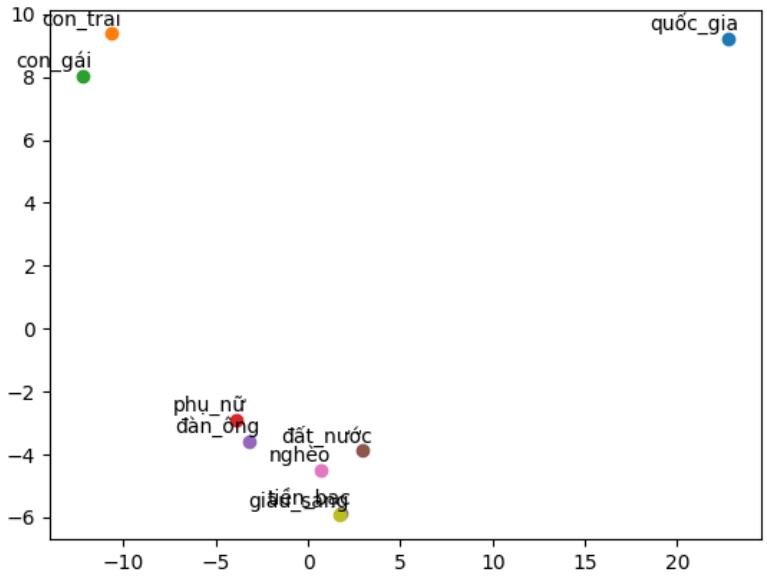

Figure 1: Vector space visualization for selected Vietnamese words.

6 Discussion

The distribution of the words in the vector space highlights several key insights:

6.1 Semantic Clustering

Words with similar meanings or those frequently used in similar contexts are positioned closer together. For example, terms related to gender and social roles such as "con_trai" (boy), "con_gai" (girl), "dan_ong" (man), and "phu_nu" (woman) are clustered together, reflecting their semantic similarity and frequent co-occurrence in discussions.

6.2 Distinct Concepts

Words representing distinct or less frequently related concepts are positioned farther apart. For instance, "quoc_gia" (nation) is isolated from other words, indicating its unique semantic context compared to the other terms in the dataset.

6.3 Contextual Associations

The visualization also shows how certain words are grouped based on their contextual usage. Words associated with economic status and social conditions such as "dat_nuoc" (country), "ngheo" (poor), and "giau_sang" (wealth) are positioned near each other, highlighting their frequent association in discussions about socioeconomic conditions.

7 Conclusion

The vector space model effectively captures the semantic relationships between different Vietnamese words. The visualization provides a clear and intuitive representation of how words are related to each other based on their meanings and contextual usage. This general overview showcases the ability of the word embeddings to encode meaningful linguistic patterns, offering valuable insights into the semantic structure of the Vietnamese language as represented in the Wikipedia corpus.

This study successfully demonstrates the creation of a vector space model for Vietnamese using Wikipedia data. FastText proves to be more effective in handling rare words compared to Word2Vec. The visualization highlights the semantic relationships between different words in Vietnamese.

References

[1] Pham Huu Quang: Building a vector space model for Vietnamese (2018).