# Finding Related Documents For Searching And Recommendation Using Word Embedding

Sairam Pillai, Ravi Javiya, Shivam Shah, Biren Sarvaiya
Guided by Professor Uday Yadav

## Abstract

The basic idea is to represent unstructured movie data (document) in a multidimensional vector space. A word embedding language model, which is trained on movie's data is proposed. The dataset includes movie plot, actors and reviews. This model would allow us to query the dataset in natural language instead of the regular SQL based query. With the help of a large dataset, this model will be able to tackle the problem of synonymy and polysemy in any language and help us retrieve the most relevant movies to the user query while being able to recommend related movies. Further our projects strengthens the fact that a word embedding model would work regardless of the language if provided enough data.

## Introduction

- A regular information retrieval system identifies the main characteristics of unstructured data to create a metadata. Entities from the natural language query are used to retrieve data from the database and the metadata is used to provide rankings to the retrieved data.
- But this system does not help in understanding synonyms or search the dataset in multiple contexts. For eg.Given the user query "the movie where a spy plays poker", the result was not the movie "Casino Royale(2013)" because the data for the movie had no word poker in it. It instead contained the hyponym of poker- "high stakes Texas Hold Em".
- Such a system also limits the query types. For eg. the query "entertaining movies" would not yield sensible results as the word "entertaining" could mean multiple things and would vary based on user preference.
- To overcome these particular obstacles we decided to train a word embedding model.

movie where a spy plays poker

Results:

| | |
|---|---|
| Casino Royale | The Spy Who Loved Me |
| The November Man | Rounders |

Fig 1. Expected searching results and working

## Word Embedding and Word2Vec

- Word Embedding vectors aims at quantifying and categorizing semantic similarities between linguistic items based on their distributional properties in large samples of language data.
- The underlying idea that "a word is characterized by the company it keeps"
- Vector space models represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points.
- Word2vec [1] is a computationally-efficient predictive model for learning word embeddings from raw text. Some interesting results from training derived from the Google News dataset:

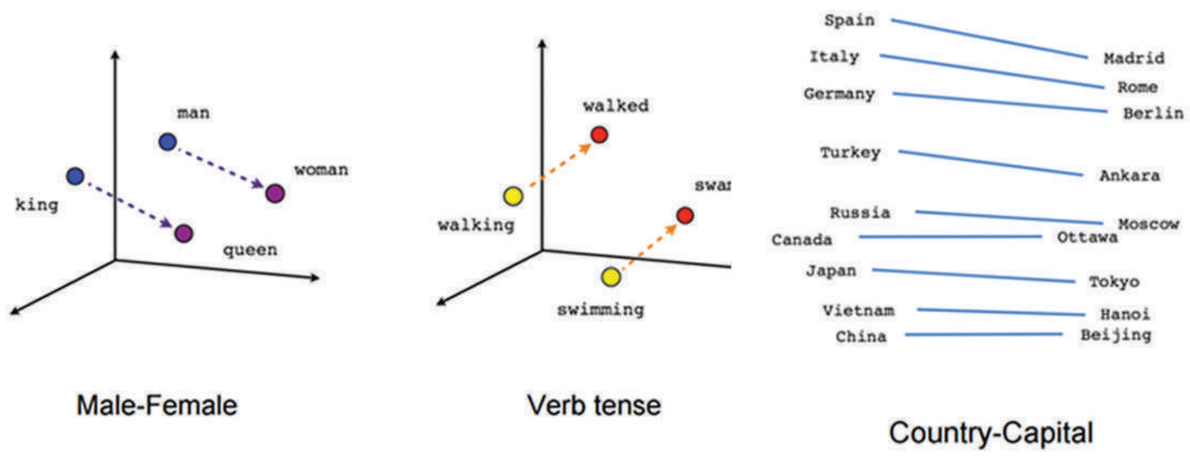



Fig 2. Visualizing embeddings using t-SNE plotted vectors

## Architecture and Implementation

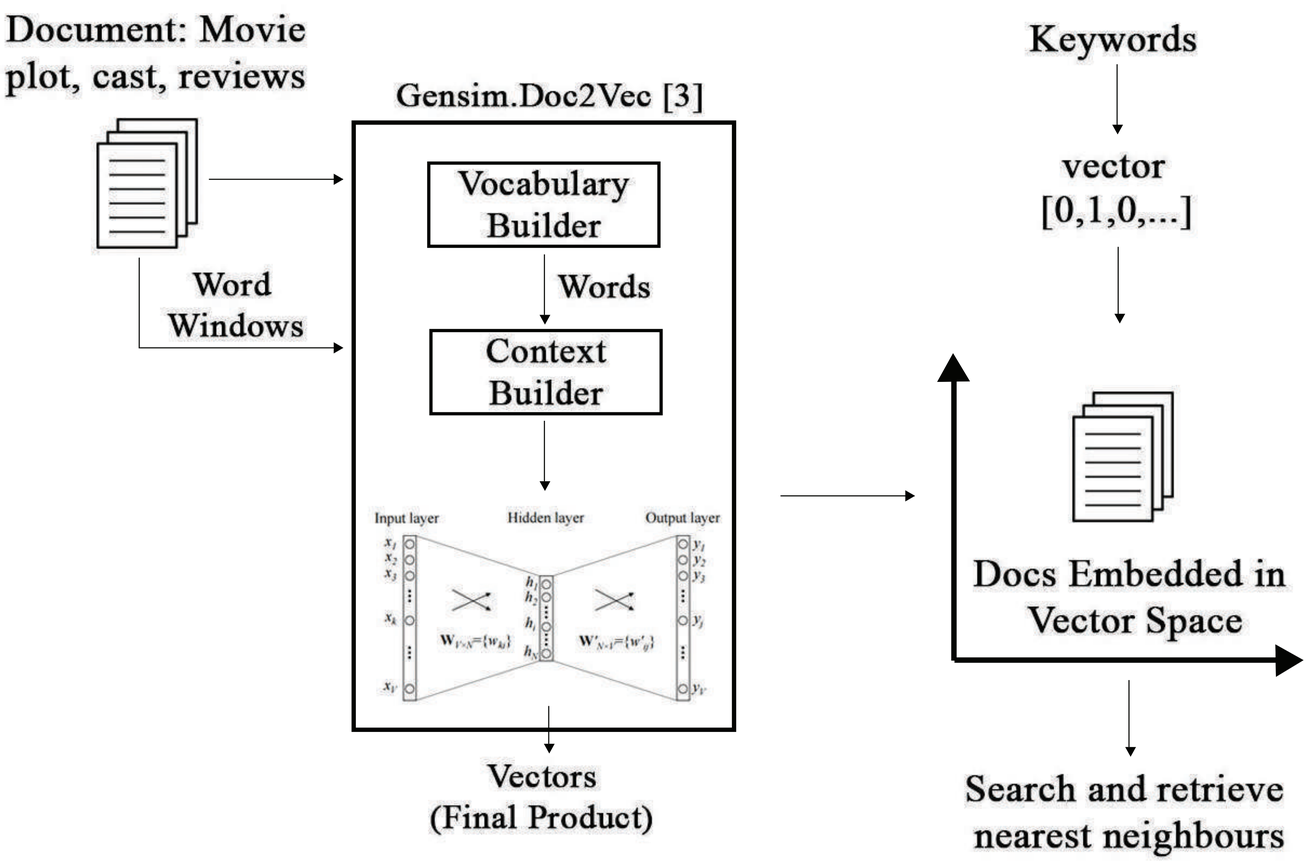



Fig 3. Result for the query "spy poker"

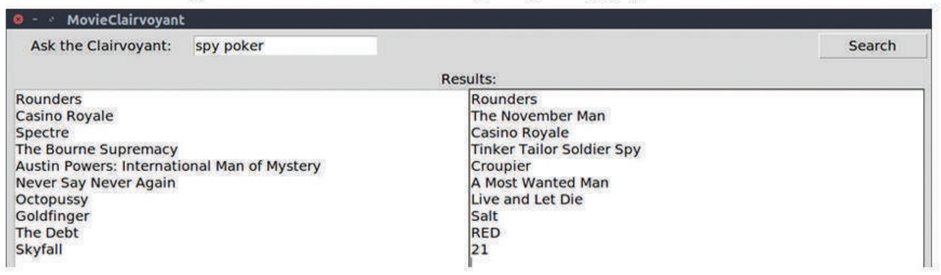
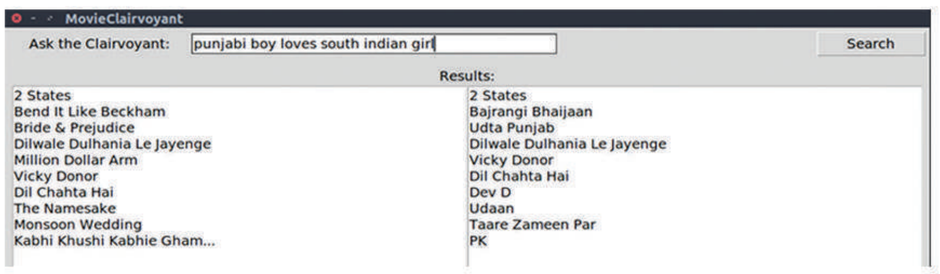



Fig 4. Result for the query "punjabi boy loves south indian girl"




## Conclusion

With our final model using Word Embeddings, we have been able to represent the required plots in many contexts and capture the distributional semantics on a 300 vector word space to enable efficient searching and retrieval of related documents in acceptable time. We hope to use this approach to make a scalable model and recommend movies based on search query and get enough data to derive unexplored genre combinations and plot points.

## References

[1] Distributed Representations of Words and Phrases and their Compositionality, Mikolov et al, 2013
[2] Distributed Representations of Sentences and Documents, Le et al, 2014
[3] Doc2Vec - https://radimrehurek.com/gensim/models/doc2vec.html
[4] https://www.tensorflow.org/tutorials/word2vec
[5] word2vec Parameter Learning Explained, Ron Xing, 2016

Acknowledgement
We sincerely thank our faculty members for providing us the guidance and resources to fulfil this project.