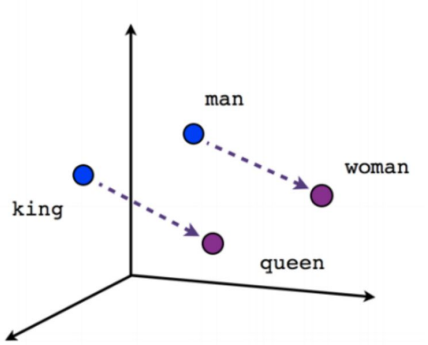
Word Embedding

Machine Learning Project Presentation

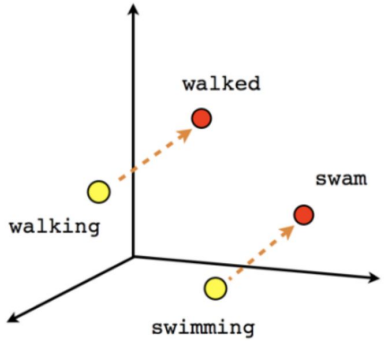
J.Jeyanthasingam	140272V
N.Kavirajan	140302P
P.Paralogarajah	140431J
K.Suthagar	140611L

Word Embedding

- Word embeddings represent the “meaning” of a word as a read-valued vector.
- Semantic relationships are often preserved on vector operations
- Comes into action with the advent of artificial neural networks

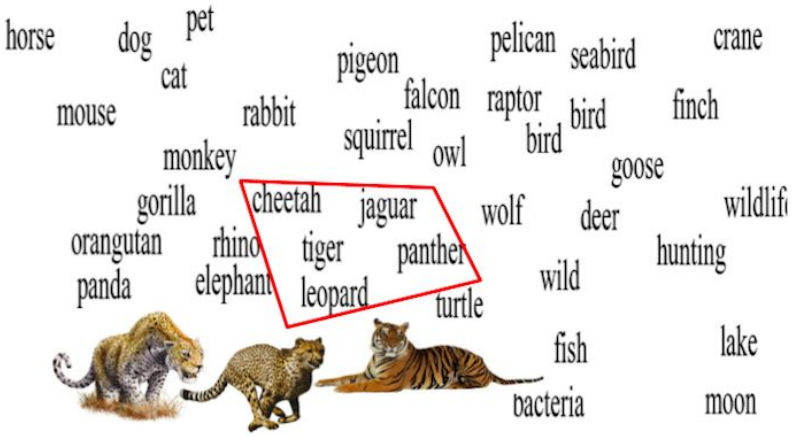


Male-Female

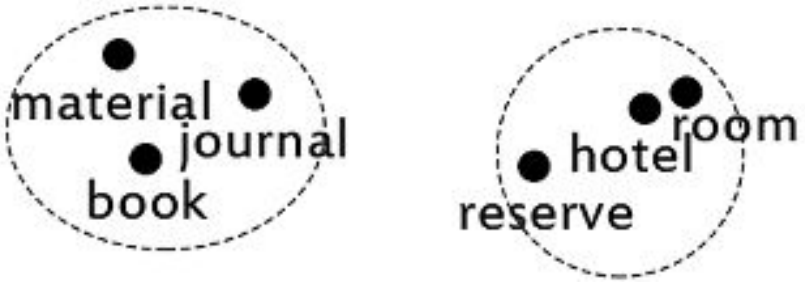


Verb tense

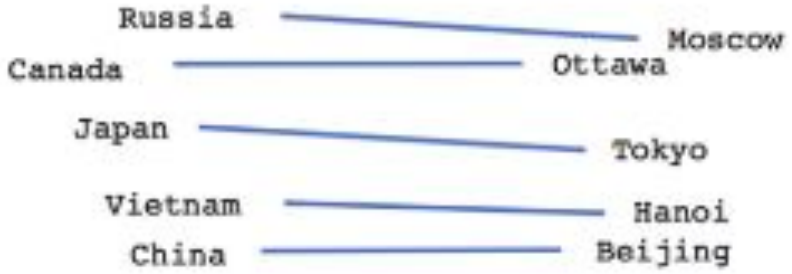
- **Word embedding is an unsupervised learning**, mainly due to their generalization power.
- Neural language model is trained on a large corpus and the output of the network is used to learn word vectors.



Word Similarity



Similar words embedded closely



Country and Capital

Methodology

1. Collecting corpus
2. Preprocessing
3. Model training
4. Evaluation

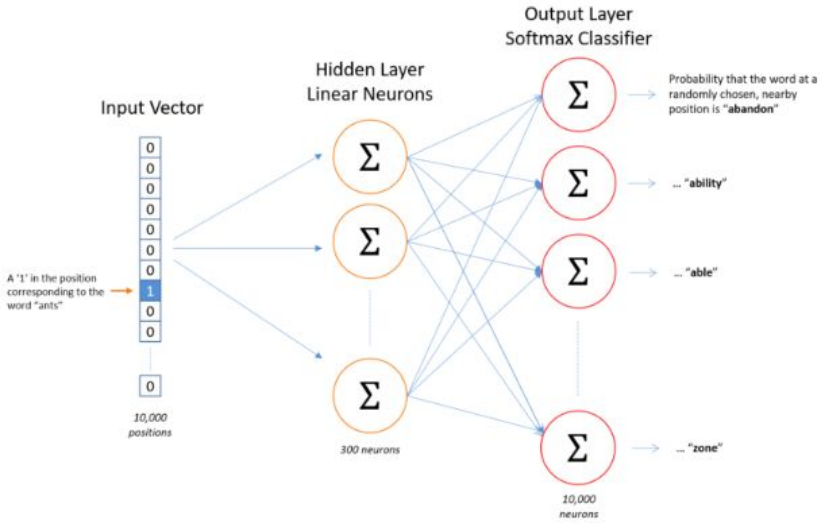
1. Pre-processing Techniques

- Special characters and phrases removals
- Tokenization
- Stop word removals
- POS tagging
- Lemmatization

2. Model Training Skip-gram model

The skip-gram neural network model is simple compare to CBOW model.

We have to train the neural network to tell us the probability for every word in our vocabulary of being the “nearby word” that we chose.



- Weights representing the connections are learnt through backpropagation.

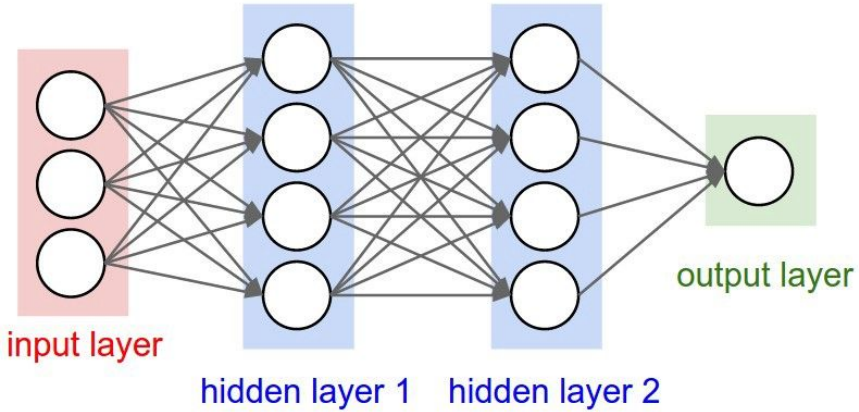
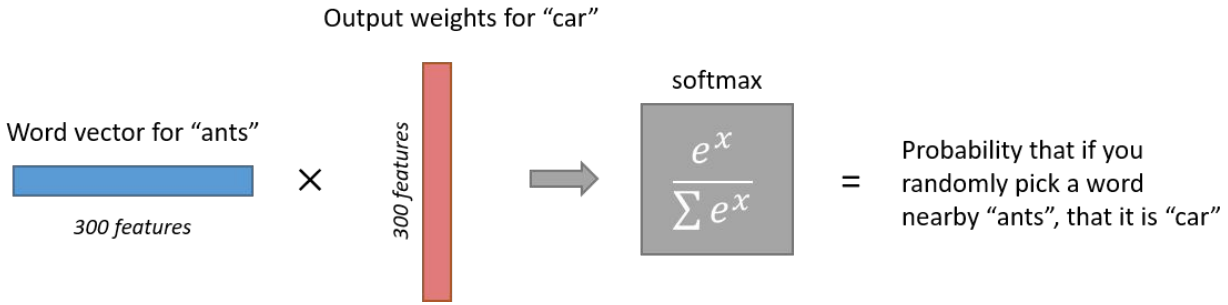


Illustration for nearby word



3. Evaluations

Subjective Evaluations

Words	Nearby words in vector space obtained from our model
america_N	○ usa_N, obama_N, newyork_N, visa_N, …
book_V	○ hotel_N, room_N, reservation_N, bus_N, …
book_N	○ buy_V, cover_N, book_V, school_N, story_N,...

Objective Evaluation

Using WS-353 dataset

Epocs	Window Size	Negative samples	Accuracy
1	5	0	14.171
1	5	5	17.895
1	8	0	16.225
5	8	5	32.589
15	8	5	39.789
30	**8**	**5**	**44.148**