greatlearning
Learning for Life

Natural Language Processing

Word embeddings

greatlearning
Learning for Life

Topics covered

- Sparse vs. dense vectors
- Word Embeddings
  - Skip-gram model
  - CBOW model
- Applications
  - POS tagging
  - Named Entity Recognition(NER)

greatlearning
Learning for Life

Session Agenda

- Understanding word embeddings
- Semantics in word embeddings
- Word embeddings model
- Application of word embeddings

Word Representations: Sparse vs. Dense

- Term-document matrix or Term-term matrix
  - Given a fixed vocabulary, we count the number of times each word occurs in a document for all documents. This matrix is the term-document matrix.
  - We count the number of times a each word pair occurs in the document for a given vocabulary, resulting matrix is the term-term matrix.
- Tf-Idf vectors
  - Tf-idf is similar to term-document matrix with each word occurrence count divided by inverse document matrix.
- One-hot encoding of words.
- Above representations of documents are sparse since most of the elements in the matrix will be zero.
- These representations do not take into account individual word relationships.

Word Embeddings: Word2Vec

- Word2Vec introduces an approach to learn word embeddings from the data.
- Word embeddings learned from this approach are dense and encode semantics between words.
- One-hot encodings are orthogonal to each other therefore words like king and queen have zero similarity. Word2Vec on the other hand provides us with dense embeddings that understand the similarity between king and queen.
- There are two methods to train your own Word2Vec model.
    - Skip-gram
    - CBOW

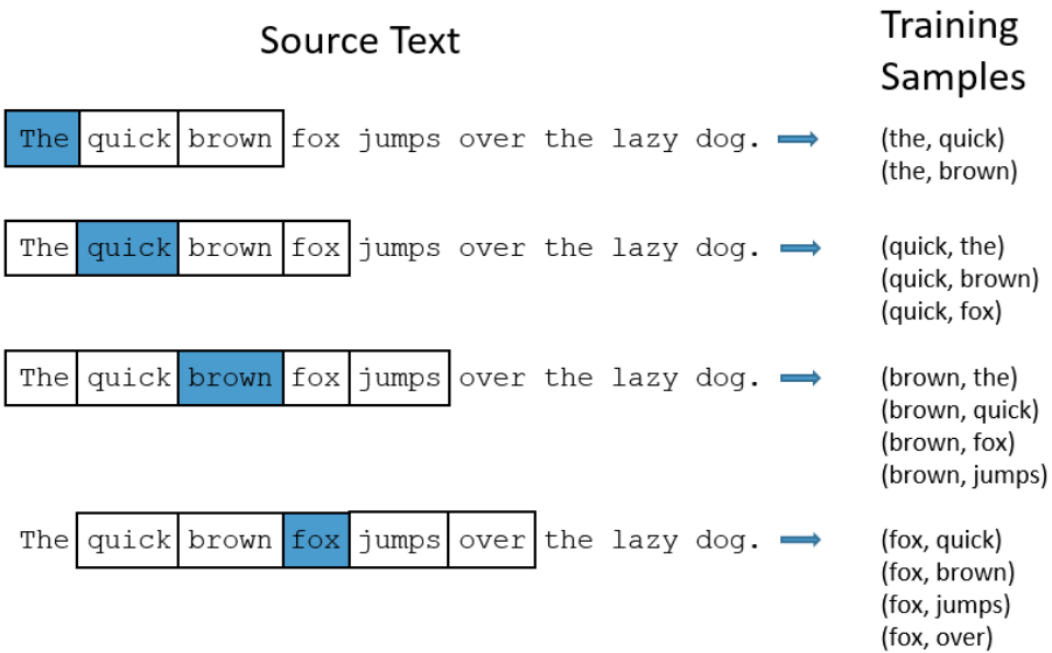
Skip-gram Model

- We'll train our model on a fake task described below,
  - Given a specific word in the middle of a sentence (the input word), look at the words nearby and pick one at random. The network is going to predict the probability for every word in our vocabulary of being the "nearby word" that we chose.
  - "Nearby" is defined on the basis of a window size.

CBOW Model

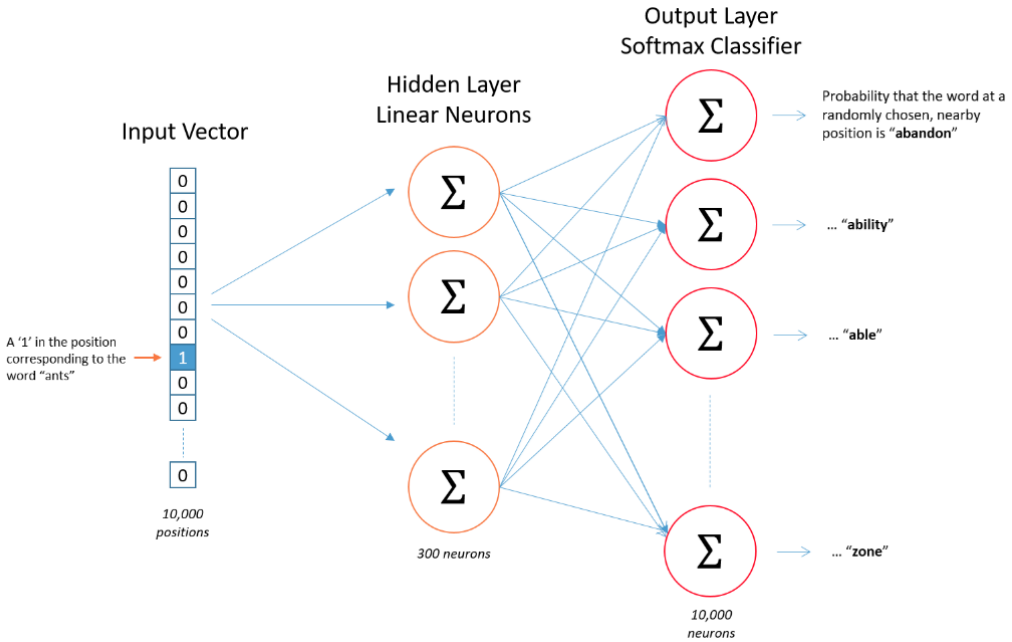
- We'll train our model on a fake task described below,
  - Given nearby words to the input words, the network is going to predict the probability of input word.

Training Sample Generation



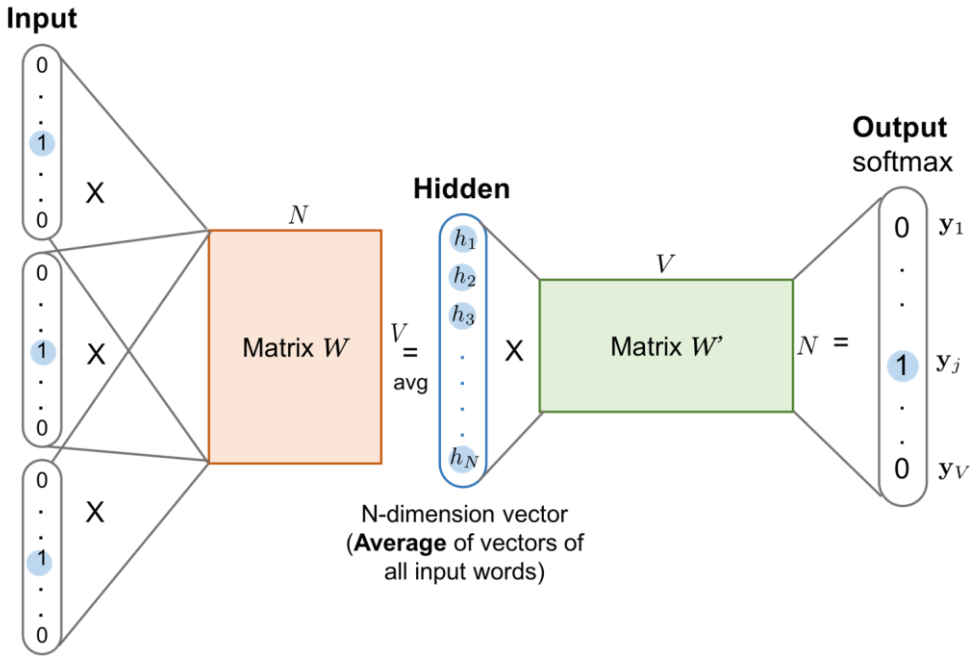
Source: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Skip-gram Model



Source: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

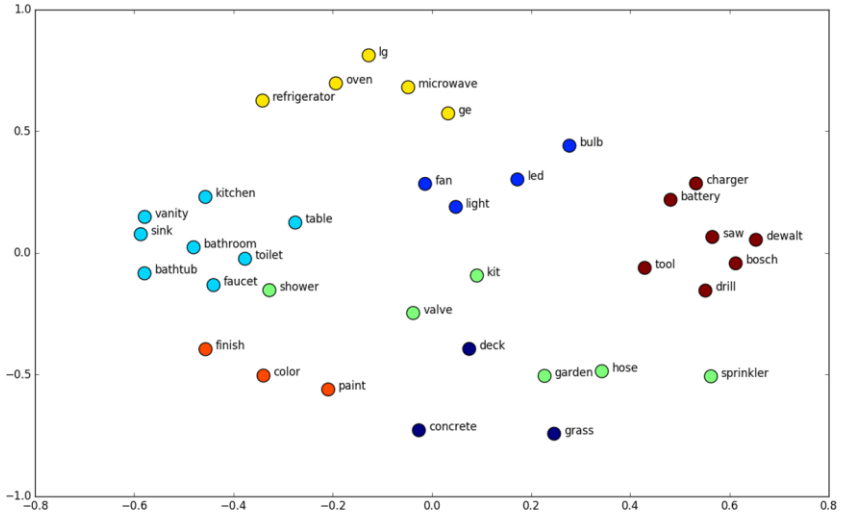
CBOW Model



Source: https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html

Word Semantics

- Words belonging to same super-category are close to each other in word2vec space.
- This similarity in word2vec space encodes the semantic relationships between words.



Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

POS Tagging

Parts of Speech(POS) are specific lexical categories to which words are assigned, based on their syntactic context and role. Usually, words can fall into one of the following major categories:

- Adjective
- Adverb
- Noun
- Verb

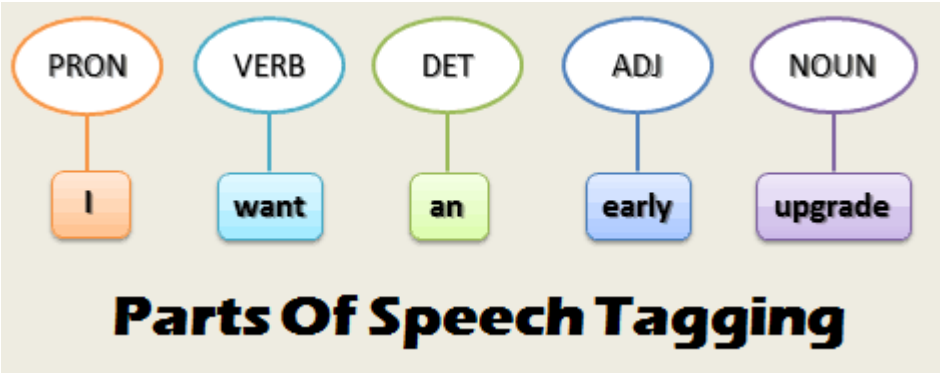


Image credits -

https://www.thinkinfi.com/2018/10/extract-custom-entity-using-nltk-pos.html

NER - Named Entity Recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Image credits - http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/

NER - Named Entity Recognition

- Classifies a text into predefined categories or real world object entities
- Takes a string of text(sentence or paragraph) as input and identifies relevant nouns(people, places, and organizations) that are mentioned in that string.

Uses

- Classify or categorize contents by getting the relevant tags
- Improve search algorithms
- For content recommendations
- For info extraction
- Check out - https://spacy.io/api/annotation#named-entities

Thank you!

Happy Learning :)