What Are Word Embeddings?

A word embedding is a learned representation for text where words that have the same meaning have a similar representation.

It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems.

One of the benefits of using dense and low-dimensional vectors is computational: the majority of neural network toolkits do not play well with very high-dimensional, sparse vectors. … The main benefit of the dense representations is generalization power: if we believe some features may provide similar clues, it is worthwhile to provide a representation that is able to capture these similarities.

— Page 92, Neural Network Methods in Natural Language Processing, 2017.

Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network, and hence the technique is often lumped into the field of deep learning.

Key to the approach is the idea of using a dense distributed representation for each word.

Each word is represented by a real-valued vector, often tens or hundreds of dimensions. This is contrasted to the thousands or millions of dimensions required for sparse word representations, such as a one-hot encoding.

associate with each word in the vocabulary a distributed word feature vector … The feature vector represents different aspects of the word: each word is associated with a point in a vector space. The number of features … is much smaller than the size of the vocabulary

— A Neural Probabilistic Language Model, 2003.

The distributed representation is learned based on the usage of words. This allows words that are used in similar ways to result in having similar representations, naturally capturing their meaning. This can be contrasted with the crisp but fragile representation in a bag of words model where, unless explicitly managed, different words have different representations, regardless of how they are used.

There is deeper linguistic theory behind the approach, namely the "*distributional hypothesis*" by Zellig Harris that could be summarized as: words that have similar context will have similar meanings. For more depth see Harris' 1956 paper "Distributional structure".

Fuente: https://machinelearningmastery.com/what-are-word-embeddings/

Algunos tutoriales copados Word Embeddings:
https://www.kaggle.com/c/word2vec-nlp-tutorial#description
https://es.coursera.org/lecture/nlp-sequence-models/learning-word-embeddings-APM5s
https://rare-technologies.com/doc2vec-tutorial/
https://rare-technologies.com/word2vec-tutorial/
:

Recursos varios
enhttp://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
https://medium.com/@adityathakker/introduction-to-word2vec-how-it-works-453ab2fb0721

Codigo:
https://github.com/isohyt/gensim/tree/e55e4cc0e430f6175309198356428e90de1ffdd4/docs/notebooks
https://radimrehurek.com/gensim/models/word2vec.html
https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb (Doc2Vec)
https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-IMDB.ipynb

Mas general en NLP(cosas bastante copadas para arrancar con tutoriales):
https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72
https://markroxor.github.io/gensim/tutorials/index.html

Posiblemente esto sea útil: LDA2Vec (Topic Modeling)
https://www.datacamp.com/community/tutorials/lda2vec-topic-model

https://saravananthirumuruganathan.wordpress.com/2012/01/10/detecting-mixtures-of-genres-in-movie-dialogues/ (LDA en peliculas)

COSAS PARA MIRAR DE LSA
http://mccormickml.com/2016/03/25/lsa-for-text-classification-tutorial/
http://www.datascienceassn.org/sites/default/files/users/user1/lsa_presentation_final.pdf
http://lsa.colorado.edu/papers/dp1.LSAintro.pdf