summary & mcq

summary

* **Word2Vec:** A 2013 natural language processing technique that represents words as numerical vectors. Cosine similarity between these vectors indicates semantic similarity. It uses a two-layer neural network trained on a large text corpus.

* **Word2Vec Applications:** Detects synonyms, suggests words for incomplete sentences, identifies word relationships (e.g., man-boy=woman-girl), and handles singular/plural forms.

* **Word2Vec Model Types:**
    * **CBOW (Continuous Bag-of-Words):** Predicts a word based on its surrounding context words. Faster than Skip-gram but less effective with infrequent words. Uses a hybrid of Bag-of-Words and N-gram techniques.
    * **Skip-gram:** Predicts surrounding context words given a word. Weights nearby words more heavily. Better for infrequent words and requires text generation.

* **Bag-of-Words (BOW):** A simplified text representation where word order is ignored, and words are represented as a multiset (bag) of their frequencies. Used as a feature in document classification.

* **N-grams:** Contiguous sequences of *n* items (letters, words, etc.) from a text. Examples include unigrams (single words), bigrams (pairs of words), trigrams (three-word sequences), etc.

* **CBOW Implementation Details:** Uses an embedding matrix and can be applied to article spinning (rephrasing articles while avoiding plagiarism).

* **Skip-gram Implementation Details:** Specifies a maximum word relationship size (determining the context window). Uses an embedding matrix (either randomly initialized or from a library like SpaCy) and one-hot encoding for input. A neural network predicts subsequent words given a single word input.

* **CBOW vs. Skip-gram:** CBOW is faster, while Skip-gram performs better with infrequent words. The choice depends on specific needs and dataset characteristics. A practical example of building a Skip-gram model from scratch without using libraries is given in the text.

mcqs

Here are 5 multiple-choice questions based on the provided text:

1. **Word2Vec is best described as:**
   a) A single-layer neural network for text classification.
   b) A technique for representing words as vectors based on semantic similarity.
   c) A method for directly measuring the frequency of words in a text corpus.
   d) An algorithm solely focused on identifying grammatical errors in text.

2. **What is the key difference between CBOW and Skip-gram in Word2Vec?**
   a) CBOW predicts the context words given a target word, while Skip-gram predicts the target word given the context words.

b) CBOW uses a single-layer neural network, while Skip-gram uses a multi-layer network.
c) CBOW is only suitable for large corpora, while Skip-gram works well with smaller datasets.
d) CBOW focuses on frequent words, while Skip-gram excels at identifying infrequent words.

3. **In the context of Word2Vec's CBOW model, what does BOW represent?**
   a) Backpropagation Over Words
   b) Bag of Words
   c) Binary Output Weights
   d) Best Of Words

4. **Which of the following is NOT a typical application of a trained Word2Vec model?**
   a) Detecting synonyms.
   b) Suggesting words to complete a sentence.
   c) Identifying the grammatical role of each word in a sentence.
   d) Finding analogies like "man-boy = woman-girl".

5. **In the Skip-gram model, what is the purpose of specifying a "maximum size of word relatedness"?**
   a) It determines the size of the vocabulary.
   b) It defines the number of hidden layers in the neural network.
   c) It sets the size of the context window considered around the target word.
   d) It limits the number of iterations during the training process.

Answer Key:

1. b)
2. a)
3. b)
4. c)
5. c)