Homework #8 Problem 1 - Roshan Poudel

What are some common natural language tasks that can be performed by deep learning?

- **Text classification** - Categorizing text into predefined categories such as spam/not spam, positive/negative, or topic-based categories.
- **Content filtering** - Filtering negative/unwanted content from the given text.
- **Sentiment analysis** - Analyzing text to determine the writer's attitude or emotion towards a subject.
- **Language modeling** - Predicting the likelihood of a sequence of words occurring in a given language.
- **Translation** - Translating text from one language to another
- **Summarization** - Condensing a larger text into a shorter, more concise summary.

Describe the three basic text preparation techniques.

1. **Standardization:** Text standardization is the process of converting text to a standard format that is consistent across the dataset. This includes removing punctuation and special characters, converting all characters to lowercase, and correcting spelling errors. Normalization can also involve stemming or lemmatization, which is the process of reducing words to their base form, such as converting "running" to "run" or "ran".
2. **Tokenization:** Tokenization is the process of breaking up text into smaller units to be vectorized. These small chunks of text are called tokens. This can be done in three different ways:
    - Word-level tokenization—Where tokens are space-separated (or punctuation-separated) substrings.
    - N-gram tokenization—Tokens are groups of N consecutive words.
    - Character-level tokenization—Where each character is its own token.
3. **Indexing:** Indexing/ is the process of assigning tokens their unique index to later represent text as a numerical vector that can be used as input to machine learning models. This involves converting each token to a numerical value, such as a count or a weight that reflects its importance in the context of the text. One common technique for vectorization is bag-of-words, where each token is assigned a count or frequency, and the resulting vector represents the frequency of each token in the text.

What are the advantages/disadvantages of "bag of words" vs. "word sequence" input types?

Advantages of "bag of words" representation:

- Since the "bag of words" model only considers the frequency of each word, it can quickly process large amounts of data, making it useful for applications that require processing large amounts of text, such as sentiment analysis or topic modeling.

- The "bag of words" model is insensitive to the order of words in a sentence, which can be an advantage in certain contexts, such as text classification, where the specific order of words may not be as important as their presence or absence.

Disadvantages of "bag of words" representation:

- The "bag of words" model does not take into account the context in which a word appears, which can result in losing important semantic information.
- The "bag of words" model treats each word as independent, which means it is unable to capture the sequential relationship between words, such as the order in which they appear.

Advantages of "word sequence" representation:

- Captures contextual and sequence information
- The "word sequence" model can perform better than the "bag of words" model on tasks such as language modeling or speech recognition.

Disadvantages of "word sequence" representation:

- The "word sequence" model is more complex to implement and can be more computationally expensive than the "bag of words" model.
- The "word sequence" model takes into account the order of words in a sentence, which can be a disadvantage in certain contexts where the order of words is not as important.
- The "word sequence" model requires more data than the "bag of words" model to train effectively, which can be a limitation in applications with limited data availability.

Describe and list the advantages of using pre-trained word embeddings in sequence models

Pretrained word embeddings are fixed-size vectors that represent words learned from a large corpus of text. The advantages of using them in sequence models include the:

- Reduced data requirements
- Improved generalization
- Reduced computational requirements
- Improved performance on specific tasks.

They can save time and resources, improve performance, and enhance the generalization ability of the model.