CSCE 771: Computer Processing of Natural Languages
Prof. Biplav Srivastava, Fall 2024

Quiz 2 / Instructions

- This is a programming heavy quiz. Code has to be placed in your Github in a directory of your GitHub called "Quiz2" with sub-dir for code, data and doc. Code will have your source code, data will have any input or output generated, and doc will have a .pdf of this file (called Quiz2-CSCE771-answers.pdf) along with any answers
- Complete quiz by Thursday, Oct 3, 2024 by uploading .pdf to Blackboard.
- [Optional] send email to by sending an email to to both Instructor biplav.s@sc.edu and TA vrawte@mailbox.sc.edu
- Total points = 20 + 40 + 15 + 25 = 100
- Obtained =

Student Name:

Rasoul Nikmehr

Objective: Understand AutoEncoder (AE) concepts

Question 1: Implement an AutoEncoder

[10 + 10 = 20 points]

An AE is a model which can produce its inputs. A sample code is given at:

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l11-nn-dl/AutoEncoder%20Sequence%20using%20LSTM.ipynb

Implement it for

- Example 1: numeric array [10 points] and
- Example 2: character array [10 points]

Objective: Understand word representation concepts – vector size fixed

Question 2: Implement Word2Vec and Glove

[20 + 20 = 40 points]

We discussed word representation methods where vector length is derived from data (like TF-IDF) v/s those where vector length is given as input. Word2Vec and Glove are examples of the latter, and so are other more recent embedding methods.

Read about:

a) Glove from: https://nlp.stanford.edu/projects/glove/

b) Word2Vec from: https://jalammar.github.io/illustrated-word2vec/

Now do the notebooks:

a) https://github.com/biplav-s/course-nl/blob/master/l7-language/code/Word%20embedding%20with%20Gensim.ipynb
b) https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l13-llm-quiz/Glove%20usage.ipynb

Question 3: [5 + 5 + 5 = 15 points]

a) What is the benefit of using a counting based vector representation like TF-IDF over a learning based representation like Word2Vec? [5 points]

(b) What are the advantages of character-based over word-based representation? What is the disadvantage [5 + 5 points]

Question 4: Word2Vec on your resume

[15 + 10 = 25 points]

(a) Take your latest resume (must be more than 1 page). Create a word2vec representation for it using genism and print statistics of embeddings.

(b) Visualize the embedding using PCA.

Question3(a):
Benefits of TF-IDF:
Simple and easy to interpret.Highlights important, document-specific terms.

Ideal for document classification and information retrieval.

Emphasizes rare but significant words.

Benefits of Word2Vec:
Captures semantic relationships between words based on context.

Effective for tasks like semantic similarity and analogy detection.

Provides deeper insights into word meanings and relationships.Suitable for context-sensitive language tasks.

Question3(b):
Advantages of Character-based Representation:
Handles Unknown Words: Can represent any word, including those not seen during training, by breaking them into characters.
Good for Complex Languages: Works well with morphologically rich languages by capturing word formation patterns.
Resilient to Misspellings: More robust to spelling errors and word variations.
Smaller Vocabulary: Reduces the need for a large word-level vocabulary, making it efficient in languages with large vocabularies.

Disadvantages of Character-based Representation:
Slower Training: Requires more time and resources due to longer sequences.
Challenges with Long-Distance Dependencies: May struggle to capture relationships between distant parts of a sentence.
Less Semantic Information: Characters carry less meaning than whole words, so it may require more data to achieve rich representations.