COMP 6721 Applied Artificial Intelligence (Winter 2023)

Assignment #4: Natural Language Processing

Due: 11:59PM, April 23rd, 2023

Note Your will be submitting two separate files from this assignment as follows:

(a) **One(1) .pdf file:** containing answers to Question as well as reported results from coding you develop.

(b) **One(1) .zip folder:** containing all developed Python codes including a README.txt file on explaining how to run your code.

Theoretical Questions

Question 1 Consider a toy example from a language that is made specifically to appreciate Artificial Intelligence. This language is simple, it uses only 7 words of the English language. Base all the next questions on the following corpus:

"Artificial Intelligence is what I like. What I like is Artificial Intelligence. What is Artificial Intelligence? What is like Artificial Intelligence? Nothing is like Artificial Intelligence. Artificial Intelligence is like nothing."

You can ignore case distinctions and sentence boundaries when answering the following questions.

(a) What are the values of the following: P(Intelligence|Artificial), P(nothing|is), P(is like), and P(Intelligence is).

(b) Build a bigram language model based on the given training corpus. Show the frequencies and probabilities for each bigram.

(c) Re-build your bigram using 0.5 smoothing. Show the frequencies and probabilities of each bigram.

(d) Using each of the previously built bigrams (with and without smoothing), determine which of the following statements is more probable. Show your work.

"Artificial Intelligence is nothing"

"I Artificial like Intelligence"

"I like Artificial Intelligence"

Question 2 Consider the following training set for disambiguating the sense of the word *orange*.

Sentence	**Sense**
This orange is very tasty.	Sense1
The fruit basket has an orange, an apple, a banana, and a pineapple.	Sense1
Maroon, Blue, and Orange are my favorite colors.	Sense2
Do you have these pants in orange or black?	Sense2
The sun had just begun to set, casting a warm, orange glow over the horizon, signaling the end of another day.	Sense2
The citrusy scent of freshly squeezed orange juice wafted through the air, making my mouth water with anticipation.	Sense1
I recommend you try our fresh orange juice if you do not like the apple juice.	Sense1
The glow of this orange shirt is unreal!	Sense2
Both orange juice and apple juice are tasty, but I prefer apples!	Sense2
The colours of the rainbow are: red, orange, yellow, green, blue, indigo, and violet.	Sense1

Consider the following list of stop words:

a, an, and, are, as, at, be, for, from, in, is, it, if, my, of, or, our, so, that, this, the, to, was, you

Use a Naive Bayes approach to determine the sense of the given statements. Use the following:

- A context window of ± 3 words,
- a vocabulary of size 33,
- smoothing with the value of 0.5, and
- stop-word removal

Calculate the score of each possible sense and find the most probable sense of the word *orange* in each of the following sentences:

(a) *Orange is my favorite color because it is so bright and cheerful.*

(b) *From our breakfast menu I recommend omelettes which come with orange juice.*

(c) *The tropical fruit salad was filled with chunks of pineapple, mango, orange, and apple, all tossed in a tangy citrus dressing.*

Question 3 In this question, you are tasked with building a CBOW Word2Vec model, using the following sentence for training:

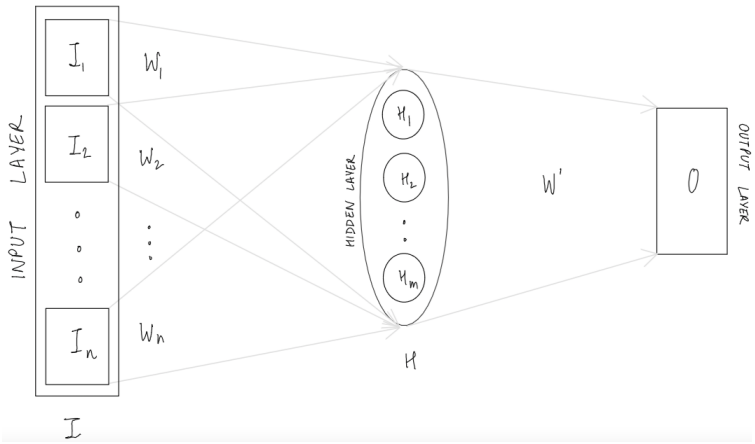
"Artificial Intelligence can replace humans in several jobs and applications".

You need to use word embeddings of dimension 3, a context window of size 4 (two words before and two words after the target word), and a vocabulary containing only the above sentence.

(a) How many training instances can be generated using the sentence above? List these instances showing the training data and the labels.

(b) Define the vocabulary of the problem and apply one-hot encoding to each of the words. Represent the training data and the labels using the one-hot encoded words.

(c) Using the standard network given below, what will be the values of n and m? What will be the sizes for I_i, W_i (for each $1 \leq i \leq n$), W', and O.



(d) Assume that we have the following weight vectors:

$$W = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 2 & 1 \\ 3 & 1 & 2 \\ 3 & 1 & 2 \\ 2 & 2 & 2 \\ 1 & 3 & 2 \\ 1 & 3 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 1 & 3 \end{bmatrix}$$

$$W' = \begin{bmatrix} 1 & 2 & 4 & 3 & 1 & 2 & 3 & 3 & 2 & 3 \\ 4 & 5 & 2 & 3 & 1 & 2 & 4 & 6 & 7 & 2 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 1 \end{bmatrix}$$

Trace the first and second feed forward passes in the network and show the values propagated all the way to the output layer. Apply a softmax function to the output layer

(e) Calculate the errors after the first pass and second feed forward passes.

Implementation Questions

Question 1 Implement a network model to compute word embeddings as per the details given in Question 3 of the Theoretical Questions. Use the aforementioned sentence as the training example, with the vocabulary containing only the words in the given sentence. Use the same word embeddings and context window size as Question 3. You should do the following:

- Define your training sentence, and compute the vocabulary and the training instances (trigrams).
- Define your language model network. You have to use the same architecture as Question 3.
- define a loss function, an optimizer, and a learning rate, and train your model. Plot the training loss.

Question 2 Consider the following paragraph from the TV show Breaking Bad:

"*Who are you talking to right now? Who is it you think you see? Do you know how much I make a year? I mean, even if I told you, you wouldn't believe it. Do you know what would happen if I suddenly decided to stop going into work? A business big enough that it could be listed on the NASDAQ goes belly up. Disappears! It ceases to exist without me. No, you clearly don't know who you're talking to, so let me clue you in. I am not in danger, Skyler. I am the danger. A guy opens his door and gets shot and you think that of me? No. I am the one who knocks!*"

(a) Using the spacy library, extract and print named entities from the given text.

(b) Similarly to Question 1, implement a network model to compute the word embeddings. Pre-process the input and train the model. Use a context size of 2 (trigrams). You are free to design your own architecture for the network model such that the loss is minimized. Plot the training loss.

Bonus Question In this question, you are required to build a CNN for a task of text classification. The aim is to build a classifier that is able to differentiate between fake and real news based on a given input text. The dataset for this problem can be found on Kaggle. Below are some general instructions to be followed:

- You need to follow the general text pre-processing steps. This includes cleaning, tokenization, embedding, etc.
- You should use Word2Vec embeddings in your solution.
- You are free to used any library for text pre-processing and embedding. However, you can only use PyTorch for deep learning related functions.

- Since you need to use CNNs, an input sentence should be converted into a 2D array/tensor.
- You can choose to use an existing CNN architecture (with or without transfer learning), or build your own CNN.
- Use a train-val-test split of 70-10-20.
- The metric to assess the proposed methods will be the classification accuracy, so aim to maximize this metric.

This question awards a bonus of 40% of the assignment grade to the final course grade.