# SMU MSDS 7337 Mid-Term Exam Summer 2020

# Short Essay Responses

(25 pts each, 300-500 words each).

1. Select one career or industry that makes use of applied NLP.
   1. Explain generally how that field or career utilizes NLP.
   2. Explain at least some methods of NLP that are very likely to be used in the career or industry you selected.
   3. Give at least one specific example of a use case for NLP within the chosen field, and explain how the problem or situation is (or could be) improved by applying NLP.
2. Choose one of the “trade-offs” in NLP that was covered in the asynchronous materials for this course.
   1. Explain the trade-off in general terms. Define the two choices.
   2. Explain the benefits and weaknesses of each side of the trade-off. Include at least one benefit and one weakness of each.
   3. Describe a work-situation that would make one of the choices in the trade-off much better, in terms of practical outcomes for you and your stakeholders on a project.

# NLP Networks

(50 pts).

I have a vocabulary of 10 words assigned the following indexes (in a dictionary):

{“the”: 0

“quick”: 1

“brown”: 2

“fox”: 3

“jumped”: 4

“over”: 5

“fence”: 6

“under”: 7

“car” : 8

“did”: 9 }

I have a network that classifies a sentence as a question or a statement. 0 means statement, 1 indicates a question.

I give you the following code as the network:

# truncate and pad input sequences

max\_sent\_length = 8

X\_train = sequence.pad\_sequences(X\_train, maxlen=max\_sent\_length)

X\_test = sequence.pad\_sequences(X\_test, maxlen=max\_sent\_length)

embedding\_vec\_length = 75

model = Sequential()

model.add(Embedding(top\_words, embedding\_vecor\_length, input\_length=max\_sent\_length))

model.add(LSTM(115, return\_sequences=True))

model.add(RNN(95))

model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary\_crossentropy', optimizer='adam', metrics=['accuracy'])

Draw/Make a diagram of this network using an input sequence of “the car jumped over the fence”

Assumptions:

The sequence tokens are words, split by whitespace.

You may label a cell by its type—there is no need to show the inner connections of the LSTM cell. (A quick reminder—LSTM has 4 sets of gates/weights, but all those gates/weights have the same size matrix—that size is what I am after!)

1. Label each block and step by input/sequence step. Compute the dimensions of the weight for all steps. All inputs must be labeled by dimension. Include your original word ENCODING (notice not vector!) as input. You may omit bias
2. Write the initial vector form of the input sequence using only 1s and 0s
3. Find the average Glove Word Vector of your the input sequence (Spacy uses Glove vectors!)
4. Find the nearest word (in the above dictionary) to answer #3
5. What is the difference between the W(weight) matrix of the first LSTM sequence at time/sequence 0 and at time/sequence 5. How do you know this?
6. What is missing in the above code—something important is not determined and based on that, there are some minor adjustments or additions that need to be made Make a logical determination of what that missing piece of info should be based on the info given here and what additions or adjustments are necessary.

Example Diagram (Obviously you would fill in the “?”):

Dense Layer 1 (100 neurons), Wdim = (?, ?)

Dense Layer 2 (300 neurons), Wdim = (?, ?)

Classifier Dense Layer (300 neurons), Wdim = (?, ?)

Input, dimensions = (1 , 29)

(1,29)

(?,?)

(?,?)

(1,300)

2.

# Question 1

## 1A

**Explain generally how that field or career utilizes NLP.**

A couple of places where NLP is being used in eCommerce is for **comparison shopping** and for **extracting product features** from the product description.

**Comparison Shopping**: When a customer comes to an eCommerce website, they may type a general description of the product that they are looking for. By presenting only the most relevant products to the customer, the eCommerce company can potentially lock in the customer's decision faster and close the sale of the product. If the customer has to wade through irrelevant search results to get to what they want instead, the company might end up losing this customer. NLP can be used to aid this comparison process by finding the closest matching products to the users search query.

**Extracting Product Features from Product Description**: In today's world where customers may have short attention span, vendors selling products usually misuse the product description field where they not only add the product description but a lot of the product features as well. This is done so that they can catch the customers attention faster. We can however, use NLP to extract the feature information from the product description to make an accurate catalog of the products being sold and this information can also use used to feed the feature filters that show up on the comparison shopping page.

## 1B.

**Explain at least some methods of NLP that are very likely to be used in the career or industry you selected.**

**Comparison Shopping**: One of the NLP methods that this might use is sentence similarity. Here sentence refers to the product description. We may choose to train a model (to get the word vectors or embeddings) using the vocabulary of the products in our eCommerce website. Using these word vectors, we can compute how similar the product descriptions are to what the user has typed in the search box. While there may be various methods to extract the single "vector" corresponding to the product description, we would start with a very basic method of taking the average of the word vectors in the product descripton and the average of the word vectors in the users search and compare the 2 to find the closest matching products.

**Extracting Product Features from Product Description**: The method most likley to be used in this case could be sentence segmentation. Before we extract the exact features out of the product description, we would want to break up the product descriptions into the sections that describe each individual feature. For this we may have to first create a training dataset that indicates when a feature description starts and when it ends in the product description. We can then train a model to recognize this and this trained model can eventually be used to segment the product description into the constituent features of the product.

## 1C.

**Give at least one specific example of a use case for NLP within the chosen field, and explain how the problem or situation is (or could be) improved by applying NLP.**

Specifically, lets look at the example of **Comparison Shopping** to see how this could be improved by using NLP. Lets say that the customer is specifically searching for an older version (version 1) of a camera on the eCommerce website because it is cheaper than the latest version (version 2). The eCommerce company may have many vendors selling both versions of these cameras. However, the eCommerce company may not specifically have a field for "product verison" for this product so the vendors may choose to add the product version to the description section itself. However the word vector for version 1 may be different from the word vector for version 2 so when we do the "product similarity" comparison using the word vectors, we will most likely end up presenting version 1 as the top search results to the customer since that would be closest to the users search query. Alternately, if we had not used this method, we might have presented version 2 as the top result (since that is what more customer might be searching for at this point in time), but that would not have been relevant to this customer. In the worst case, if version 1 did not show up on the first page of the search, the customer may have gone to a different eCommerce website thinking that version 1 is not offered at this website.

# Question 2

Choose one of the “trade-offs” in NLP that was covered in the asynchronous materials for this course.

* 1. Explain the trade-off in general terms. Define the two choices.
  2. Explain the benefits and weaknesses of each side of the trade-off. Include at least one benefit and one weakness of each.
  3. Describe a work-situation that would make one of the choices in the trade-off much better, in terms of practical outcomes for you and your stakeholders on a project.

## 2A

In terms of trade-offs in NLP, the one that resonates with me is **Feature Learning vs. Feature Engineering**.

In simple terms, Feature Learning means that using (a lot of) data, train a machine learning or deep learning algorithm to automatically figure out the “important features” needed to get the best metrics possible. In this approach, we rely on the machine to figure out (mine) the right features from a large corpus of data.

On the contrary, Feature Engineering relies on a Subject Matter Expert (SME) to define what the right set of features should be. This method relies on the SMEs expertise and intuition to define these features.

## 2B

In terms of pros and cons of both approaches, Feature Learning is useful when we have a lot of labelled data already available in which case, we can feed this into our algorithm quickly (maybe with a little bit of data cleaning and EDA). On the other hand, if large amounts of labelled data are not available, the algorithms will most likely fail to recognize the right set of “engineered features” from the data. Also, this approach tends to alienate the SMEs since they feel that the decisions are being made by a “black box” algorithm and these decisions may not always be explainable or make sense to these experts.

On the contrary, the pros of “Feature Engineering” by a SME is that the experts feel somewhat in control and have a vested interest in the success of the work. This can lead to better buy in from management teams. In addition, in many cases, large amounts of labelled may not be available and in these cases, hand engineering of the important features by an SME would be a logical choice. The cons of feature engineering that we must find a cooperative SME who is willing to work with the data scientist/machine learning engineer in order to define these features.

## 2C

The domain that I work in is Analog Circuit Design. This field is considered highly specialized and Subject Matter Experts (SMEs) consider this to be more of an “art” form than science. Hence any notion of “automation” is frowned upon by a large section of the population.