

# Assignment4

January 9, 2022

## 1 Assignment 4: Language Processing with RNN-Based Autoencoders

**Deadline:** Sunday, June 15th, by 9pm.

**Submission:** Submit a PDF export of the completed notebook as well as the ipynb file.

In this assignment, we will practice the application of deep learning to natural language processing. We will be working with a subset of Reuters news headlines that are collected over 15 months, covering all of 2019, plus a few months in 2018 and in a few months of this year.

In particular, we will be building an **autoencoder** of news headlines. The idea is similar to the kind of image autoencoder we built in lecture: we will have an **encoder** that maps a news headline to a vector embedding, and then a **decoder** that reconstructs the news headline. Both our encoder and decoder networks will be Recurrent Neural Networks, so that you have a chance to practice building

- a neural network that takes a sequence as an input
- a neural network that generates a sequence as an output

This assignment is organized as follows:

- Question 1. Exploring the data
- Question 2. Building the autoencoder
- Question 3. Training the autoencoder using *data augmentation*
- Question 4. Analyzing the embeddings (interpolating between headlines)

Furthermore, we'll be introducing the idea of **data augmentation** for improving of the robustness of the autoencoder, as proposed by Shen et al [1] in ICML 2020.

[1] Shen, Tianxiao, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. "Educating text autoencoders: Latent representation guidance via denoising." In International Conference on Machine Learning, pp. 8719-8729. PMLR, 2020.

```
[1]: import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim

import matplotlib.pyplot as plt
import numpy as np
import random
from math import floor
```

## 1.1 Question 1. Data (20 %)

Download the files `reuters_train.txt` and `reuters_valid.txt`, and upload them to Google Drive.

Then, mount Google Drive from your Google Colab notebook:

```
[4]: from google.colab import drive
drive.mount('/content/gdrive')

train_path = '/content/gdrive/My Drive/Deep_Learning_Course/data/reuters_train.
→txt' # Update me
valid_path = '/content/gdrive/My Drive/Deep_Learning_Course/data/reuters_valid.
→txt' # Update me
```

Mounted at `/content/gdrive`

As we did in some of our examples (e.g., training transformers on IMDB reviews) will be using PyTorch's `torchtext` utilities to help us load, process, and batch the data. We'll be using a `TabularDataset` to load our data, which works well on structured CSV data with fixed columns (e.g. a column for the sequence, a column for the label). Our tabular dataset is even simpler: we have no labels, just some text. So, we are treating our data as a table with one field representing our sequence.

```
[ ]: import torchtext.legacy.data as data

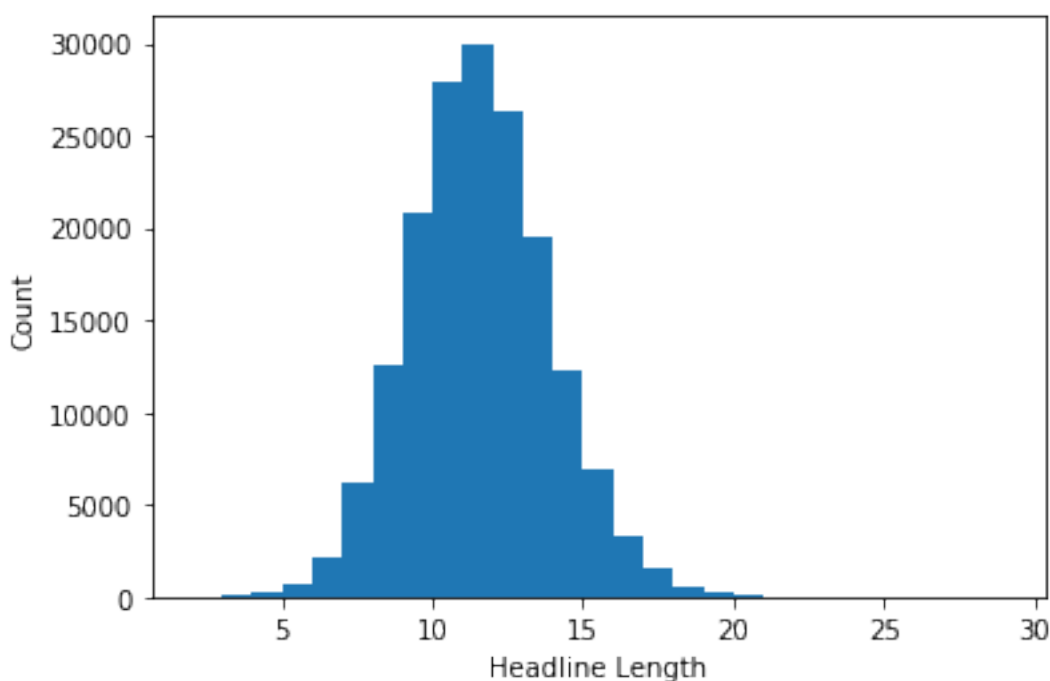
# Tokenization function to separate a headline into words
def tokenize_headline(headline):
    """Returns the sequence of words in the string headline. We also
    prepend the "<bos>" or beginning-of-string token, and append the
    "<eos>" or end-of-string token to the headline.
    """
    return ("<bos> " + headline + " <eos>").split()

# Data field (column) representing our *text*.
text_field = data.Field(
    sequential=True,          # this field consists of a sequence
    tokenize=tokenize_headline, # how to split sequences into words
    include_lengths=True,     # to track the length of sequences, for
→batching
    batch_first=True,         # similar to batch_first=True used in nn.RNN
→demonstrated in lecture
    use_vocab=True)           # to turn each character into an integer index
train_data = data.TabularDataset(
    path=train_path,          # data file path
    format="tsv",             # fields are separated by a tab
    fields=[('title', text_field)]) # list of fields (we have only one)
```

### 1.1.1 Part (a) -- 5%

Draw histograms of the number of words per headline in our training set. Excluding the <bos> and <eos> tags in your computation. Explain why we would be interested in such histograms.

```
[ ]: # Include your histogram and your written explanations
length_list = []
for headline in train_data:
    length_list.append(len(headline.title[1:-1]))
plt.hist(length_list, bins=max(length_list)-min(length_list))
plt.xlabel('Headline Length')
plt.ylabel('Count')
plt.show()
```



#### Write your explanation here:

During inference, we can observe the decoded sentence length and decide if it is reasonable based on its length as well, e.g., if we decode a sentence with length 200, this is not a reasonable result based on the training data distribution and we can ignore this result. The histogram can further help us to validate that the sentences are not too long ~ [1 - 30] words per sentence. The observation that the sentences are not too long encourages us to think that RNN has the capabilities to solve this problem.

In addition, from the sentence length, we can have some intuition about the latent space size.

### 1.1.2 Part (b) -- 5%

How many distinct words appear in the training data? Exclude the <bos> and <eos> tags in your computation.

```
[ ]: # Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values

words = set()
for sentence in train_data:
    for word in sentence.title:
        words.add(word)
# We remove the <bos> and <eos> from the length of distinct words
len_words = len(words) - 2
print(f'Number of distinct words {len_words}')
```

Number of distinct words 51298

### 1.1.3 Part (c) -- 5%

The distribution of *words* will have a long tail, meaning that there are some words that will appear very often, and many words that will appear infrequently. How many words appear exactly once in the training set? Exactly twice? Print these numbers below

```
[ ]: # Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
from collections import Counter
words_dict = {word: 0 for word in words}
for sentence in train_data:
    for word in sentence.title:
        if word == '<eos>' or word == '<bos>':
            continue
        else:
            words_dict[word] += 1
count_words = Counter(words_dict)
res = [1 for word, length in count_words.items() if length == 1]
print(f'Number of words that appear only 1 times is - {sum(res)}')
res = [1 for word, length in count_words.items() if length == 2]
print(f'Number of words that appear only 2 times is - {sum(res)}')
```

Number of words that appear only 1 times is - 19854

Number of words that appear only 2 times is - 7193

### 1.1.4 Part (d) -- 5%

We will replace the infrequent words with an <unk> tag, instead of learning embeddings for these rare words. torchtext also provides us with the <pad> tag used for padding short sequences for batching. We will thus only model the top 9995 words in the training set, excluding the tags <bos>, <eos>, <unk>, and <pad>.

What percentage of total word count(whole dataset) will be supported? Alternatively, what percentage of total word count(whole dataset) in the training set will be set to the <unk> tag?

```
[ ]: # Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
```

```

# Exclude the words <eos> and <bos> from the counter
try:
    count_words.pop('<bos>')
except KeyError:
    pass
try:
    count_words.pop('<eos>')
except KeyError:
    pass

# Calculate how many words we are going to use
numberOfUsedWords = sum([appear for word, appear in count_words.
    ↪most_common(9995)])
numberOfTotalWords = len(sorted(count_words.elements()))
print(f'Percentile of words are used for training {numberOfUsedWords/
    ↪numberOfTotalWords * 100}%')
print(f'Percentile of words are excluded from training {100 - numberOfUsedWords/
    ↪numberOfTotalWords * 100}%')

```

Percentile of words are used for training 93.97857393100142%

Percentile of words are excluded from training 6.021426068998579%

The torchtext package will help us keep track of our list of unique words, known as a **vocabulary**. A vocabulary also assigns a unique integer index to each word.

```

[:]: # Build the vocabulary based on the training data. The vocabulary
# can have at most 9997 words (9995 words + the <bos> and <eos> token)
text_field.build_vocab(train_data, max_size=9997)

# This vocabulary object will be helpful for us
vocab = text_field.vocab
print(vocab.stoi["hello"]) # for instances, we can convert from string to ↵
    ↪(unique) index
print(vocab.itos[10])      # ... and from word index to string

# The size of our vocabulary
vocab_size = len(text_field.vocab.stoi)
# Here are the two tokens that torchtext adds for us:
print(vocab.itos[0]) # <unk> represents an unknown word not in our vocabulary
print(vocab.itos[1]) # <pad> will be used to pad short sequences for batching

```

0  
on  
<unk>  
<pad>

## 1.2 Question 2. Text Autoencoder (40%)

Building a text autoencoder is a little more complicated than an image autoencoder like we did in class. So we will need to thoroughly understand the model that we want to build before actually building it. Note that the best and fastest way to complete this assignment is to spend time upfront understanding the architecture. The explanations are quite dense, but it is important to understand the operation of this model. The rationale here is similar in nature to the seq2seq RNN model we discussed in class, only we are dealing with unsupervised learning here rather than machine translation.

## 2 Architecture description

Here is a diagram showing our desired architecture:

There are two main components to the model: the **encoder** and the **decoder**. As always with neural networks, we'll first describe how to make **predictions** with of these components. Let's get started:

The **encoder** will take a sequence of words (a headline) as *input*, and produce an embedding (a vector) that represents the entire headline. In the diagram above, the vector  $\mathbf{h}^{(7)}$  is the vector embedding containing information about the entire headline. This portion is very similar to the sentiment analysis RNN that we discussed in lecture (but without the fully-connected layer that makes a prediction).

The **decoder** will take an embedding (in the diagram, the vector  $\mathbf{h}^{(7)}$ ) as input, and uses a separate RNN to **generate a sequence of words**. To generate a sequence of words, the decoder needs to do the following:

1. Determine the previous word that was generated. This previous word will act as  $\mathbf{x}^{(t)}$  to our RNN, and will be used to update the hidden state  $\mathbf{m}^{(t)}$ . Since each of our sequences begin with the <bos> token, we'll set  $\mathbf{x}^{(1)}$  to be the <bos> token.
2. Compute the updates to the hidden state  $\mathbf{m}^{(t)}$  based on the previous hidden state  $\mathbf{m}^{(t-1)}$  and  $\mathbf{x}^{(t)}$ . Intuitively, this hidden state vector  $\mathbf{m}^{(t)}$  is a representation of *all the words we still need to generate*.
3. We'll use a fully-connected layer to take a hidden state  $\mathbf{m}^{(t)}$ , and determine *what the next word should be*. This fully-connected layer solves a *classification problem*, since we are trying to choose a word out of  $K = \text{vocab\_size}$  distinct words. As in a classification problem, the fully-connected neural network will compute a *probability distribution* over these vocab\_size words. In the diagram, we are using  $\mathbf{z}^{(t)}$  to represent the logits, or the pre-softmax activation values representing the probability distribution.
4. We will need to *sample* an actual word from this probability distribution  $\mathbf{z}^{(t)}$ . We can do this in a number of ways, which we'll discuss in question 3. For now, you can imagine your favourite way of picking a word given a distribution over words.
5. This word we choose will become the next input  $\mathbf{x}^{(t+1)}$  to our RNN, which is used to update our hidden state  $\mathbf{m}^{(t+1)}$ , i.e., to determine what are the remaining words to be generated.

We can repeat this process until we see an <eos> token generated, or until the generated sequence becomes too long.

### 3 Training the architecture

While our autoencoder produces a sequence, computing the loss by comparing the complete generated sequence to the ground truth (the encoder input) gives rise to multiple challenges. One is that the generated sequence might be longer or shorter than the actual sequence, meaning that there may be more/fewer  $\mathbf{z}^{(t)}$ s than ground-truth words. Another more insidious issue is that the **gradients will become very high-variance and unstable**, because **early mistakes will easily throw the model off-track**. Early in training, our model is unlikely to produce the right answer in step  $t = 1$ , so the gradients we obtain based on the other time steps will not be very useful.

At this point, you might have some ideas about "hacks" we can use to make training work. Fortunately, there is one very well-established solution called **teacher forcing** which we can use for training: instead of *sampling* the next word based on  $\mathbf{z}^{(t)}$ , we will forget sampling, and use the **ground truth  $\mathbf{x}^{(t)}$**  as the input in the next step.

Here is a diagram showing how we can use **teacher forcing** to train our model:

We will use the RNN generator to compute the logits  $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}$ . These distributions can be compared to the ground-truth words using the cross-entropy loss. The loss function for this model will be the sum of the losses across each  $t \in \{1, \dots, T\}$ .

We'll train the encoder and decoder model simultaneously. There are several components to our model that contain tunable weights:

- The word embedding that maps a word to a vector representation. In theory, we could use GloVe embeddings, as we did in class. In this assignment we will not do that, but learn the word embedding from data. The word embedding component is represented with blue arrows in the diagram.
- The encoder RNN (which will use GRUs) that computes the embedding over the entire headline. The encoder RNN is represented with black arrows in the diagram.
- The decoder RNN (which will also use GRUs) that computes hidden states, which are vectors representing what words are to be generated. The decoder RNN is represented with gray arrows in the diagram.
- The **projection MLP** (a fully-connected layer) that computes a distribution over the next word to generate, given a decoder RNN hidden state. The projection is represented with green arrows

#### 3.1 Part (a) -- 20%

Complete the code for the AutoEncoder class below by:

1. Filling in the missing numbers in the `__init__` method using the parameters `vocab_size`, `emb_size`, and `hidden_size`.
2. Complete the forward method, which uses teacher forcing and computes the logits  $\mathbf{z}^{(t)}$  of the reconstruction of the sequence.

You should first try to understand the `encode` and `decode` methods, which are written for you. The `encode` method bears much similarity to the RNN we wrote in class for sentiment analysis. The `decode` method is a bit more challenging. You might want to scroll down to the `sample_sequence` function to see how this function will be called.

You can (but don't have to) use the `encode` and `decode` method in your forward method. In either case, be careful of the input that you feed into either `decode` or to `self.decoder_rnn`. Refer

to the teacher-forcing diagram. **bold text** Notice that `batch_first` is set to `True`, understand how deal with it.

```
[ ]: class AutoEncoder(nn.Module):
    def __init__(self, vocab_size, emb_size, hidden_size):
        """
        A text autoencoder. The parameters
        - vocab_size: number of unique words/tokens in the vocabulary
        - emb_size: size of the word embeddings  $x^{(t)}$ 
        - hidden_size: size of the hidden states in both the
                      encoder RNN ( $h^{(t)}$ ) and the
                      decoder RNN ( $m^{(t)}$ )
        """
        super().__init__()
        self.embed = nn.Embedding(num_embeddings=vocab_size,
                                   embedding_dim=emb_size)
        self.encoder_rnn = nn.GRU(input_size=emb_size,
                                   hidden_size=hidden_size,
                                   batch_first=True)
        self.decoder_rnn = nn.GRU(input_size=emb_size,
                                   hidden_size=hidden_size,
                                   batch_first=True)
        self.proj = nn.Linear(in_features=hidden_size,
                               out_features=vocab_size)

    def encode(self, inp):
        """
        Computes the encoder output given a sequence of words.
        """
        emb = self.embed(inp)
        out, last_hidden = self.encoder_rnn(emb)
        return last_hidden

    def decode(self, inp, hidden=None):
        """
        Computes the decoder output given a sequence of words, and
        (optionally) an initial hidden state.
        """
        emb = self.embed(inp)
        out, last_hidden = self.decoder_rnn(emb, hidden)
        out_seq = self.proj(out)
        return out_seq, last_hidden

    def forward(self, inp):
        """
        Compute both the encoder and decoder forward pass
        given an integer input sequence inp with shape [batch_size,
        ↪ seq_length],
```



*with  $inp[a,b]$  representing the (index in our vocabulary of) the  $b$ -th  $\rightarrow$ word of the  $a$ -th training example.*

*This function should return the logits  $z^{(t)}$  in a tensor of shape  $[batch\_size, seq\_length - 1, vocab\_size]$ , computed using *teaching*  $\rightarrow$ forcing\*.*

*The  $(seq\_length - 1)$  part is not a typo. If you don't understand why we need to subtract 1, refer to the teacher-forcing diagram above.*

```
"""
x = self.encode(inp)
new_inp = inp[:, :-1]
output, last_hidden = self.decode(new_inp, hidden=x)
return output
```

### 3.1.1 Part (b) -- 10%

To check that your model is set up correctly, we'll train our autoencoder neural network for at least 300 iterations to memorize this sequence:

```
[ ]: headline = train_data[42].title
input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).long().unsqueeze(0)
```

We are looking for the way that you set up your loss function corresponding to the figure above. Be careful of off-by-one errors here.

Note that the Cross Entropy Loss expects a rank-2 tensor as its first argument (the output of the network), and a rank-1 tensor as its second argument (the true label). You will need to properly reshape your data to be able to compute the loss.

```
[ ]: model = AutoEncoder(vocab_size, 128, 128)
optimizer = optim.Adam(model.parameters(), lr=0.001)
criterion = nn.CrossEntropyLoss()

for it in range(300):
    z = model(input_seq)
    optimizer.zero_grad()
    # We remove the <bos> from the begin of sentence
    loss = criterion(torch.squeeze(z), torch.squeeze(input_seq[:, 1:]))
    loss.backward()
    optimizer.step()

    if (it+1) % 50 == 0:
        print("[Iter %d] Loss %f" % (it+1, float(loss)))
```

```
[Iter 50] Loss 0.086394
[Iter 100] Loss 0.025433
[Iter 150] Loss 0.016252
[Iter 200] Loss 0.011460
```

```
[Iter 250] Loss 0.008538
[Iter 300] Loss 0.006679
```

### 3.1.2 Part (c) -- 4%

Once you are satisfied with your model, encode your input using the RNN encoder, and sample some sequences from the decoder. The sampling code is provided to you, and performs the computation from the first diagram (without teacher forcing).

Note that we are sampling from a multi-nomial distribution described by the logits  $z^{(t)}$ . For example, if our distribution is [80%, 20%] over a vocabulary of two words, then we will choose the first word with 80% probability and the second word with 20% probability.

Call `sample_sequence` at least 5 times, with the default temperature value. Make sure to include the generated sequences in your PDF report.

```
[ ]: def sample_sequence(model, hidden, max_len=20, temperature=1):
    """
    Return a sequence generated from the model's decoder
    - model: an instance of the AutoEncoder model
    - hidden: a hidden state (e.g. computed by the encoder)
    - max_len: the maximum length of the generated sequence
    - temperature: described in Part (d)
    """
    # We'll store our generated sequence here
    generated_sequence = []
    # Set input to the <BOS> token
    inp = torch.Tensor([text_field.vocab.stoi["<bos>"]]).long()
    for p in range(max_len):
        # compute the output and next hidden unit
        output, hidden = model.decode(inp.unsqueeze(0), hidden)
        # Sample from the network as a multinomial distribution
        output_dist = output.data.view(-1).div(temperature).exp()
        top_i = int(torch.multinomial(output_dist, 1)[0])
        # Add predicted word to string and use as next input
        word = text_field.vocab.itos[top_i]
        # Break early if we reach <eos>
        if word == "<eos>":
            break
        generated_sequence.append(word)
        inp = torch.Tensor([top_i]).long()
    return generated_sequence

# Your solutions go here
hidden = model.encode(input_seq)
for iter in range(5):
    print(25*'=' + ' Iteration ' + str(iter) + ' ' + 25*'=' )
    print(sample_sequence(model, hidden))
```

```
===== Iteration 0 =====
['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
```

```

===== Iteration 1 =====
['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
===== Iteration 2 =====
['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
===== Iteration 3 =====
['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
===== Iteration 4 =====
['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']

```

### 3.1.3 Part (d) -- 6%

The multi-nomial distribution can be manipulated using the temperature setting. This setting can be used to make the distribution "flatter" (e.g. more likely to generate different words) or "peakier" (e.g. less likely to generate different words).

Call `sample_sequence` at least 5 times each for at least 3 different temperature settings (e.g. 1.5, 2, and 5). Explain why we generally don't want the temperature setting to be too **large**.

```

[ ]: # Include the generated sequences and explanation in your PDF report.
for temperature in [1.5,2,5]:
    print(25*('=' + ' Temperature ' + str(temperature) + ' ' + 25*('='))
    for iter in range(5):
        print(f'{str(iter)} {sample_sequence(model, hidden,
        →temperature=temperature)}')

```

```

===== Temperature 1.5 =====
0) ['informal', 'cyclone', 'zambian', 'president', 'swears', 'in', 'new',
'chosen', 'in', 'invasion', 'bors', 'wide', 'zarif', 'morales']
1) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'in', 'at',
'argentina', 'new', 'army', 'fundraising', 'default', '_num_pound', 'clues',
'cautions', 'stanley', 'chief']
2) ['zambian', 'traveling', 'swears', 'in', 'suspicion', 'yanks', 'library',
'final']
3) ['zambian', 'headlines', 'climbs', 'swears', 'in', 'new', 'army', 'tiktok',
'athlete', '_num_second', 'versum']
4) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
===== Temperature 2 =====
0) ['renews', 'notorious', 'interfax', 'defenses', 'dilemma', 'herd', 'fury',
'push', 'displays', 'wpp', 'authorities', 'taste', 'dips', 'your', 'gain',
'specialist', 'tick', 'positive', 'letter']
1) ['notch', 'cobalt', 'army', 'layoffs', 'populist', 'snc-lavalin', 'close',
'lady', 'paribas', '_num_spotify', 'cells', 'lyles', 'in-house', 'swears',
'six-year', 'computer', 'zero', 'commuter', 'impact', 'kohli']
2) ['_num_facebook', 'exits', 'found', 'osaka', 'dialog', 'current', 'palo',
'widening', 'absence', 'full-year', 'new', 'jose', 'plead', 'employees',
'enough']
3) ['bogota', 'refugee', 'rekindles', 'northwestern', 'chief', 'telecom',
'antibiotic', 'defies', 'confirmed', 'differences', 'endo', 'restaurants',
'grace', 'opener', 'toward', '-official', 'manufacturers', 'blackouts',

```

```

'forces', 'caught']
4) ['reshuffles', 'netease', 'cheap', 'bath', 'beer', 'new', '_num_-dish',
'anz', 'prove', 'treaty', 'shaw', 'bias', 'parting', 'rouhani', 'president',
'cape', 'hand', 'in', 'new', 'king']
===== Temperature 5 =====
0) ['belgian', 'lender', 'astronauts', 'neutrality', 'tasnim', 'proof',
'gbagbo', 'ireland', 'royal', 'rooney', 'abn', 'widening', 'manufacturers',
'contain', 'francis', 'waning', 'legislative', 'bottles', 'gunman', 'second']
1) ['ex-aide', 'schumer', 'nixon', 'opening', 'arkansas', 'prayer', 'fca-
renault', 'respond', 'procedures', 'troop', 'tightens', 'institutions',
'recede', 'convictions', 'chargers', 'park', 'major', 'past', 'who', 'chief']
2) ['abbas', 'putting', 'audit', 'sails', 'divided', 'billionaire', 'sika',
'mortgage', 'words', 'united', 'holy', 'toronto', 'lunar', 'more', 'forever',
'rowe', 'sec', 'wife', 'manufacturers', 'pats']
3) ['warrant', 'vtb', 'nebraska', 'blind', 'catalonia', 'bit', 'exhibition',
'ftc', 'summer', 'respect', 'announces', 'takeda', 'chief', 'stress', 's.',
'contained', 'factors', 'afghanistan', 'pressed', 'sberbank']
4) ['bladder', 'electrek', '_num_-california', 'resort', 'barnier', 'chocolate',
'activity', 'broad', 'alphabet', 'ailing', 'oslo', 'egg', 'strategic', 'sport',
'markets-stock', 'aside', 'hall', 'sovereignty', 'vindman', 'berlin']

```

The temperature value determine how flatten our distribution. To demonstrate it let's consider two extreme cases where the temperature  $t$  tends to zero or to infinity. The probability to sample the  $i$ 'th word from the vocabulary is  $\Pr(w_i) = \frac{\exp(\frac{z_i}{t})}{\sum_{j=1}^n \exp(\frac{z_j}{t})}$ , where  $z_i$  is the  $i$ 'th element in the logit vector  $z$ . While  $t$  tends to infinity, for any  $i \in \{k\}_{k=1}^n$ ,  $\Pr(w_i) = \frac{1}{\sum_{j=1}^n 1} = \frac{1}{n}$ . While  $t$  tends to zero, let's denote by  $i^* \in \{i\}_{i=1}^n$ , the index that satisfies  $z_{i^*} = \max\{z_i\}_{i=1}^n$  (for simplicity we will assume that there is a unique such  $i^*$ ).  $\Pr(w_i) = \frac{\exp(\frac{z_i}{t})}{\exp(\frac{z_{i^*}}{t}) + \sum_{j=1, j \neq i^*}^n \exp(\frac{z_j}{t})}$ . By dividing the numerator and the denominator by  $\exp(\frac{z_{i^*}}{t})$ , we get  $\Pr(w_i) = \frac{\exp(\frac{z_i - z_{i^*}}{t})}{\exp(0) + \sum_{j=1, j \neq i^*}^n \exp(\frac{z_j - z_{i^*}}{t})}$ . Note that because  $z_j - z_{i^*} < 0$  for  $j \neq i^*$  then, as  $t$  tends to zero, the denominator tends to 1. If  $i \neq i^*$  then,  $z_i - z_{i^*} < 0$ , hence while  $t$  tends to zero the numerator tends to zero and therefore,  $\Pr(w_i) = 0$ . If  $i = i^*$  then,  $z_i - z_{i^*} = 0$ , hence the numerator is equal to 1. Therefore  $\Pr(w_i) = 1$ . The conclusion is that for  $t$  that tends to zero,  $\Pr(w_i) = \delta[i - i^*]$  where  $\delta$  denotes the Kronecker delta function. Therefore, we conclude that if the temperature is high, the model is more likely to return a words which according to the model are not the most likely with higher probability. In the extreme case when the chosen temperature is very high, The predicted word is simply random.

### 3.2 Question 3. Data augmentation (20%)

It turns out that getting good results from a text auto-encoder is very difficult, and that it is very easy for our model to **overfit**. We have discussed several methods that we can use to prevent overfitting, and we'll introduce one more today: **data augmentation**.

The idea behind data augmentation is to artificially increase the number of training examples by "adding noise" to the image. For example, during AlexNet training, the authors randomly cropped  $224 \times 224$  regions of a  $256 \times 256$  pixel image to increase the amount of training data. The

authors also flipped the image left/right. Machine learning practitioners can also add Gaussian noise to the image.

When we use data augmentation to train an *autoencoder*, we typically only add the noise to the input, and expect the reconstruction to be *noise free*. This makes the task of the autoencoder even more difficult. An autoencoder trained with noisy inputs is called a **denoising auto-encoder**. For simplicity, we will *not* build a denoising autoencoder today.

### 3.2.1 Part (a) -- 5%

We will add noise to our headlines using a few different techniques:

1. Shuffle the words in the headline, taking care that words don't end up too far from where they were initially
2. Drop (remove) some words
3. Replace some words with a blank word (a <pad> token)
4. Replace some words with a random word

The code for adding these types of noise is provided for you:

```
[ ]: def tokenize_and_randomize(headline,
                                drop_prob=0.1, # probability of dropping a word
                                blank_prob=0.1, # probability of "blanking" out a
→word
                                sub_prob=0.1, # probability of substituting a word
→with a random one
                                shuffle_dist=3): # maximum distance to shuffle a
→word
    """
    Add 'noise' to a headline by slightly shuffling the word order,
    dropping some words, blanking out some words (replacing with the <pad>
→token)
    and substituting some words with random ones.
    """
    headline = [vocab.stoi[w] for w in headline.split()]
    n = len(headline)
    # shuffle
    headline = [headline[i] for i in get_shuffle_index(n, shuffle_dist)]

    new_headline = [vocab.stoi['<bos>']]
    for w in headline:
        if random.random() < drop_prob:
            # drop the word
            pass
        elif random.random() < blank_prob:
            # replace with blank word
            new_headline.append(vocab.stoi["<pad>"])
        elif random.random() < sub_prob:
            # substitute word with another word
            new_headline.append(random.randint(0, vocab_size - 1))
```

```

        else:
            # keep the original word
            new_headline.append(w)
        new_headline.append(vocab.stoi['<eos>'])
    return new_headline

def get_shuffle_index(n, max_shuffle_distance):
    """ This is a helper function used to shuffle a headline with n words,
    where each word is moved at most max_shuffle_distance. The function does
    the following:
        1. start with the *unshuffled* index of each word, which
           is just the values [0, 1, 2, ..., n]
        2. perturb these "index" values by a random floating-point value between
           [0, max_shuffle_distance]
        3. use the sorted position of these values as our new index
    """
    index = np.arange(n)
    perturbed_index = index + np.random.rand(n) * 3
    new_index = sorted(enumerate(perturbed_index), key=lambda x: x[1])
    return [index for (index, pert) in new_index]

```

Call the function `tokenize_and_randomize` 5 times on a headline of your choice. Make sure to include both your original headline, and the five new headlines in your report.

```

[:]: # Report your values here. Make sure that you report the actual values,
      # and not just the code used to get those values

text = train_data[50].title
text = ' '.join(text)
print(f'Original: {text}')
print(100* '-')
for iter in range(5):
    # Argument the sentence - we remove the <bos> and <eos> because the function
    # →insert them manually
    output = tokenize_and_randomize(text[1:-1])
    # Convert to string from integer representation
    output = [vocab.itos[elem] for elem in output]
    # Join the sentence
    output = ' '.join(output)
    print(f'{str(iter)} {output}')

```

Original: <bos> vehicle ploughs into new year 's crowd in tokyo <eos>

-----

- 0) <bos> vehicle <unk> into year new <pad> tokyo crowd in <unk> <eos>
- 1) <bos> <unk> vehicle <unk> patrols new year cats santos <pad> tokyo <unk> <eos>
- 2) <bos> haftar <unk> <unk> grip new year 's in crowd tokyo <unk> <eos>

- 3) <bos> vehicle <unk> year 's recommended crowd in both solved <eos>
- 4) <bos> <pad> vehicle <unk> museum new 's year crowd <unk> tokyo in <eos>

### 3.2.2 Part (b) -- 8%

The training code that we use to train the model is mostly provided for you. The only part we left blank are the parts from Q2(b). Complete the code, and train a new AutoEncoder model for 1 epoch. You can train your model for longer if you want, but training tend to take a long time, so we're only checking to see that your training loss is trending down.

If you are using Google Colab, you can use a GPU for this portion. Go to "Runtime" => "Change Runtime Type" and set "Hardware acceleration" to GPU. Your Colab session will restart. You can move your model to the GPU by typing `model.cuda()`, and move other tensors to GPU (e.g. `xs = xs.cuda()`). To move a model back to CPU, type `model.cpu`. To move a tensor back, use `xs = xs.cpu()`. For training, your model and inputs need to be on the *same device*.

```
[ ]: device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")

[ ]: def train_autoencoder(model, batch_size=64, learning_rate=0.001, num_epochs=10):
    optimizer = optim.Adam(model.parameters(), lr=learning_rate)
    criterion = nn.CrossEntropyLoss()
    train_loss = []
    iterVec = []
    iterCounter = 0
    for ep in range(num_epochs):
        # We will perform data augmentation by re-reading the input each time
        field = data.Field(sequential=True,
                           tokenize=tokenize_and_randomize, # <--
                           include_lengths=True,
                           batch_first=True,
                           use_vocab=False, # <-- the tokenization
                           pad_token=vocab.stoi['<pad>'])
        # This BucketIterator will handle padding of sequences that are not of
        dataset = data.TabularDataset(train_path, "tsv", [('title', field)])
        train_iter = data.BucketIterator(dataset,
                                         batch_size=batch_size,
                                         sort_key=lambda x: len(x.title), # to
                                         repeat=False)
        for it, ((xs, lengths), _) in enumerate(train_iter):
            xs = xs.to(device)
            # Fill in the training code here
            z = model(xs).reshape(-1, vocab_size)
            labels_without_bos = xs[:,1:].reshape(-1)
            optimizer.zero_grad()
```

```

        # We remove the <bos> from the begin of sentence
        loss = criterion(z, labels_without_bos)
        loss.backward()
        optimizer.step()
        if (it+1) % 100 == 0:
            print("[Iter %d] Loss %f" % (it+1, float(loss)))
            train_loss.append(float(loss))
            iterVec.append(iterCounter)
            iterCounter += 1

        # Optional: Compute and track validation loss
        # val_loss = 0
        # val_n = 0
        # for it, ((xs, lengths), _) in enumerate(valid_iter):
        #     zs = model(xs)
        #     loss = None # TODO
        #     val_loss += float(loss)

# Include your training curve or output to show that your training loss is
→trending down
    return iterVec, train_loss

```

```

[ ]: def plot_learning_curve(iters, train_losses):
    """
    Plot the learning curve.
    """
    plt.title("Learning Curve: Loss per Iteration")
    plt.plot(iters, train_losses, label="Train")
    plt.xlabel("Iterations")
    plt.ylabel("Loss")
    plt.show()

```

```

[ ]: model = AutoEncoder(vocab_size, 128, 128).cuda()
    iters, losses = train_autoencoder(model, num_epochs=1)
    plot_learning_curve(iters[:50], losses[:50])

```

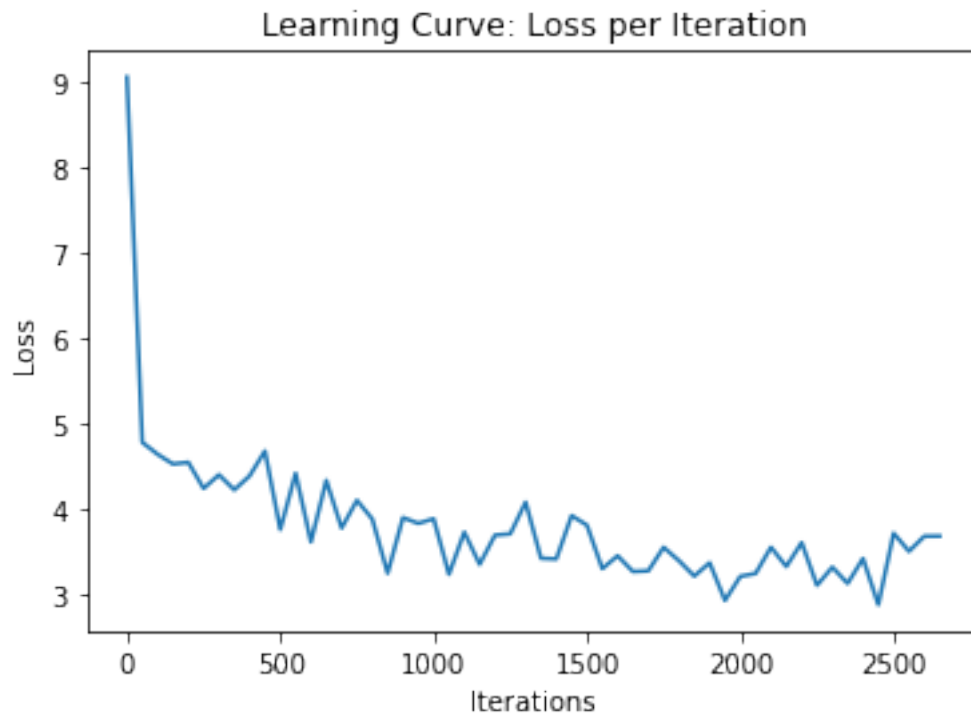
```

[Iter 100] Loss 4.165724
[Iter 200] Loss 3.849206
[Iter 300] Loss 3.761531
[Iter 400] Loss 4.483603
[Iter 500] Loss 4.468360
[Iter 600] Loss 4.188010
[Iter 700] Loss 3.539526
[Iter 800] Loss 3.943288
[Iter 900] Loss 3.356971
[Iter 1000] Loss 3.703237
[Iter 1100] Loss 3.556163
[Iter 1200] Loss 3.111730

```



```
[Iter 1300] Loss 3.944893
[Iter 1400] Loss 3.100752
[Iter 1500] Loss 3.306914
[Iter 1600] Loss 3.217526
[Iter 1700] Loss 3.001801
[Iter 1800] Loss 3.277736
[Iter 1900] Loss 3.417384
[Iter 2000] Loss 3.496868
[Iter 2100] Loss 3.330262
[Iter 2200] Loss 3.337307
[Iter 2300] Loss 3.323949
[Iter 2400] Loss 3.326774
[Iter 2500] Loss 3.466325
[Iter 2600] Loss 3.750400
```



### 3.2.3 Part (c) -- 7%

This model requires many epochs (>50) to train, and is quite slow without using a GPU. You can train a model yourself, or you can load the model weights that we have trained, and available on the course website (AE\_RNN\_model.pk).

Assuming that your AutoEncoder is set up correctly, the following code should run without error.

```
[ ]: model = AutoEncoder(10000, 128, 128)
checkpoint_path = '/content/gdrive/My Drive/Deep_Learning_Course/AE_RNN_model.
↳pk' # Update me
model.load_state_dict(torch.load(checkpoint_path))
```

```
[ ]: <All keys matched successfully>
```

Then, repeat your code from Q2(d), for `train_data[10].title` with temperature settings 0.7, 0.9, and 1.5. Explain why we generally don't want the temperature setting to be too **small**.

```
[ ]: # Include the generated sequences and explanation in your PDF report.

headline = train_data[10].title
input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).unsqueeze(0).long()
hidden = model.encode(input_seq)
orgTitle = ' '.join(headline)
print(f'Original) {orgTitle}')
# Include the generated sequences and explanation in your PDF report.
for temperature in [0.7, 0.9, 1.5]:
    print(25* '=' + ' Temperature ' + str(temperature) + ' ' + 25* '=')
    for iter in range(5):
        print(f'{str(iter)} {sample_sequence(model, hidden,
↳temperature=temperature)}')
```

Original) <bos> wall street rises , limps across the finish line of a turbulent year <eos>

===== Temperature 0.7 =====

```
0) ['wall', 'street', 'rises', ',', 'limps', 'across', 'the', 'finish', 'line',
'of', 'a', 'turbulent', 'year']
1) ['wall', 'street', 'rises', ',', 'limps', 'across', 'the', 'finish', 'line',
',', 'next', 'election', 'post']
2) ['wall', 'street', 'rises', ',', 'limps', 'open', 'sentence', ',', 'second',
'olympics', '<pad>', 'praise', 'solar']
3) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', '$', '<pad>',
'highway', 'election', 'surge']
4) ['wall', 'street', 'rises', ',', 'limps', 'form', 'atlantia', 'after', 'to',
'<pad>', 'credits', ':', 'nazi']
```

===== Temperature 0.9 =====

```
0) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', ',', 'pay',
'market', 'after', 'gabon']
1) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', '"s", 'ways',
'season', ':', '<pad>', 'political']
2) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', '$', 'top',
'iran', 'azerbaijan', 'signs']
3) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', 'of',
'sciences', 'election', 'four']
4) ['wall', 'street', 'rises', ',', 'limps', 'across', 'the', 'finish', 'line',
'of', 'a', 'turbulent', 'year']
```

===== Temperature 1.5 =====

```

0) ['wall', 'street', 'rises', ',', 'limps', 'hazard', 'to', 'consent', 'as',
    'stake', '<pad>', 'peaking', 'ahead']
1) ['wall', 'street', 'rises', ',', 'limps', 'behind', 'home', 'title', 'avoid',
    'to', 'confident', 'this', 'year']
2) ['wall', 'street', 'st.', 'positive', '<unk>', 'abandoned', "'s", 'pulls',
    'challenging', '<pad>', 'back', '<pad>', 'found', 'post']
3) ['wall', 'street', 'comcast', 'basis', ',', 'initial', 'iaaf', 'alarms',
    "'s", 'rally', 'indonesia', 'deal', 'royal']
4) ['wall', 'street', 'rises', ',', 'limps', 'across', 'starmer', 'after',
    'look', '<pad>', 'three', 'in', 'modestly']

```

**Write your explanation here:** As we showed in section (2d), when the temperature decreases, the distribution tends to a delta function over the maximal logit value. If for example there are one or more logits that are close to the maximal logit value, while the temperature is very small, the model will almost always return the word that corresponded to the maximal logit value. In this case, the model will probably miss other reasonable words with logit values that are close to the maximal logit value. We also would like to insert randomness in the model prediction since there is a sense of randomness in the human language, where several sentences may start the same while ending up totally different. For example, There are many beautiful women, There are many cold-blood killers.

### 3.3 Question 4. Latent space manipulations (20%)

In parts 2-3, we've explored the decoder portion of the autoencoder. In this section, let's explore the **encoder**. In particular, the encoder RNN gives us embeddings of news headlines!

First, let's load the **validation** data set:

```

[ ]: valid_data = data.TabularDataset(
    path=valid_path,          # data file path
    format="tsv",             # fields are separated by a tab
    fields=[('title', text_field)]) # list of fields (we have only one)

```

#### 3.3.1 Part (a) -- 4%

Compute the embeddings of every item in the validation set. Then, store the result in a single PyTorch tensor of shape [19046, 128], since there are 19,046 headlines in the validation set.

```

[ ]: # Write your code here
# Show that your resulting PyTorch tensor has shape `[19046, 128]`
embeddingList = []
# move the model to the device
model.to(device)
for element in valid_data:
    headline = element.title
    input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).unsqueeze(0).
    →long().to(device)
    hidden = model.encode(input_seq).cpu()
    embeddingList.append(hidden.squeeze())

```

```
[ ]: embeddingTensor = torch.stack(embeddingList,0)
      embeddingTensor.shape
```

```
[ ]: torch.Size([19046, 128])
```

### 3.3.2 Part (b) -- 4%

Find the 5 closest headlines to the headline `valid_data[13]`. Use the cosine similarity to determine closeness. (Hint: You can use code from assignment 2)

```
[ ]: # Write your code here. Make sure to include the actual 5 closest headlines.
      embeddings = embeddingTensor.detach().numpy()
      norms = np.linalg.norm(embeddings, axis=1)
      embeddings_norm = (embeddings.T / norms).T
      similarities = np.matmul(embeddings_norm, embeddings_norm.T)
```

```
[ ]: (19046, 19046)
```

```
[ ]: def find_similarity(data,HeadlineIdx,similarityMat):
      similarities_vec = similarityMat[HeadlineIdx,:]
      most5sim = similarities_vec.argsort()[-6:][::-1][1:]
      similarities_headlines = [' '.join(data[idx].title) for idx in most5sim]
      return similarities_headlines, similarities_vec[most5sim]
```

```
[ ]: def print_similarity(idx):
      headline = valid_data[idx].title
      orgTitle = ' '.join(headline)
      print(f'Original {orgTitle}')
      similarHeadlines,similarityVec = find_similarity(valid_data,idx,similarities)
      for cnt, vals in enumerate(zip(similarHeadlines,similarityVec)):
          headline,similarityVal = vals
          similarityVal = floor((similarityVal) * 10**5) / 10**5
          print(f'{cnt}) {headline} -- Similarity of {similarityVal}')
```

```
[ ]: print_similarity(13)
```

```
Original) <bos> asia takes heart from new year gains in u.s. stock futures <eos>
0) <bos> italy 's salvini loses aura of invincibility in emilia setback <eos> --
Similarity of 0.93087
1) <bos> saudi , russia look to seal deeper output cuts with oil producers <eos>
-- Similarity of 0.93055
2) <bos> eu orders quarantine for staff who traveled to northern italy <eos> --
Similarity of 0.92985
3) <bos> update _num_-italy 's prime minister says new government will bicker
less <eos> -- Similarity of 0.92876
4) <bos> portugal 's moura pays tribute to cod fishermen at milan fashion close
<eos> -- Similarity of 0.92808
```

### 3.3.3 Part (c) -- 4%

Find the 5 closest headlines to another headline of your choice.

```
[ ]: print_similarity(10)
```

```
Original) <bos> trump invites congressional leaders to border security briefing
- source <eos>
0) <bos> trump ask congressional leaders to return for shutdown talks on friday
: lawmaker <eos> -- Similarity of 0.93873
1) <bos> trump to receive update on border security bill : official <eos> --
Similarity of 0.93129
2) <bos> trump told russians in _num_ not concerned by meddling -wapo <eos> --
Similarity of 0.92995
3) <bos> trump re-election campaign targeted by iran-linked hackers : sources
<eos> -- Similarity of 0.92706
4) <bos> trump picks herman cain for fed , news reports say <eos> -- Similarity
of 0.92681
```

### 3.3.4 Part (d) -- 8%

Choose two headlines from the validation set, and find their embeddings. We will **interpolate** between the two embeddings like we did in the example presented in class for training autoencoders on MNIST.

Find 3 points, equally spaced between the embeddings of your headlines. If we let  $e_0$  be the embedding of your first headline and  $e_4$  be the embedding of your second headline, your three points should be:

$$e_1 = 0.75e_0 + 0.25e_4$$

$$e_2 = 0.50e_0 + 0.50e_4$$

$$e_3 = 0.25e_0 + 0.75e_4$$

Decode each of  $e_1$ ,  $e_2$  and  $e_3$  five times, with a temperature setting that shows some variation in the generated sequences. Try to get a logical and cool sentence (this might be hard).

```
[ ]: WordIdx = [13,17]
headlines = [valid_data[idx].title for idx in WordIdx]
input_seqs = torch.stack([torch.Tensor([vocab.stoi[w] for w in headline]).
    →long() for headline in headlines])
model = model.cpu()
e = model.encode(input_seqs).squeeze(0).detach()

e1 = 0.75*e[0]+0.25*e[1]
e2 = 0.5*e[0]+0.5*e[1]
e3 = 0.25*e[0]+0.75*e[1]

encs=[e1,e2,e3]
temp=1.5
temperature = 1.5
print("temperature:",temperature)
```

```

es = ["e1", "e2", "e3"]
for idx, headline in enumerate(headlines):
    print(f'Original {idx} {" ".join(headline)}')

for e_index, e in enumerate([e1, e2, e3]):
    print(es[e_index], " decodings:")
    e = e.unsqueeze(0).unsqueeze(0)
    for i in range(5):
        print(sample_sequence(model, e, temperature=temperature))

```

temperature: 1.5

Original 0) <bos> asia takes heart from new year gains in u.s. stock futures  
<eos>

Original 1) <bos> indian manufacturing growth slows in december despite price  
cuts : pmi <eos>

e1 decodings:

['update', 'palm', 'serena', 'flat', 'slowest', 'u.s.', 'for', 'day', 'decade',  
'inbev', 'than']  
 ['markets', 'services', 'demonstrate', 'first', 'lazard', 'for', 'after',  
'crude', 'tumble', '\_num\_', 'wall']  
 ['asia', 'takes', 'crosshairs', 'sector', 'says', 'santos', 'march', 'tumble',  
'of', 'biggest', 'powder']  
 ['apple', 'ended', 'in', 'jet', 'prices', '"s', 'three', 'weekly', 'after',  
'<pad>', 'rivals']  
 ['inbev', 'who', 'slumps', 'cities', 'biggest', 'of', 'week', 'steadies',  
'almost', 'line', 'weak']

e2 decodings:

['recycling', ';', 'nearly', 'under', 'of', 'high', 'compound', 'decade',  
'bolsonaro', 'as', 'likely']  
 ['indian', 'lining', 'lower', 'bayer', 'coast', 'growth', 'in', '<pad>', 'fuel',  
'november', 'growth']  
 ['indian', 'miners', 'trading', 'emissions', 'in', 'central', 'quarter', 'oil',  
'european', '\_num\_', 'suncor']  
 ['emerging', '\_num\_', 'criticism', 'drops', 'in', 'top', 'three', 'inbev',  
'output', 'stocks', 'nice']  
 ['antitrust', 'nearly', 'up', 'concerns', 'death', 'investments', 'tanker',  
'takeover', 'investment', 'deal', 'thursday']

e3 decodings:

['indian', '%', 'takeover', 'january', 'in', '%', 'despite', 'to', 'aramco',  
'shifting', 'sources']  
 ['utah', 'output', '\_num\_', 'asian', 'in', 'hughes', '<pad>', 'global',  
'january', 'strong', 'seat']  
 ['indian', 'asia', 'pistons', 'tame', 'in', 'next', ';', 'nov', 'more', ';',  
'netherlands']  
 ['indian', 'down', 'revenue', 'anti-semitism', 'first', 'in', 'investments',  
'to', 'fx', 'plants', ':']  
 ['indian', 'income', 'max', 'in', 'fall', 'weeks', 'dec', 'chances', 'fall',

```
'oil', 'past']
```

### Generating PDF format

```
[12]: !apt-get install texlive texlive-xetex texlive-latex-extra pandoc  
      !pip install py pandoc
```

```
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
pandoc is already the newest version (1.19.2.4~dfsg-1build4).  
texlive is already the newest version (2017.20180305-1).  
texlive-latex-extra is already the newest version (2017.20180305-2).  
texlive-xetex is already the newest version (2017.20180305-1).  
0 upgraded, 0 newly installed, 0 to remove and 37 not upgraded.  
Requirement already satisfied: py pandoc in /usr/local/lib/python3.7/dist-  
packages (1.7.2)
```

```
[13]: from google.colab import drive  
      drive.mount('/content/gdrive')
```

Drive already mounted at /content/gdrive; to attempt to forcibly remount, call `drive.mount("/content/gdrive", force_remount=True)`.

```
[16]: !cp gdrive/My\ Drive/Colab\ Notebooks/Assignment4.ipynb ./  
      !jupyter nbconvert --to=pdf "Assignment4.ipynb"
```

```
[NbConvertApp] Converting notebook Assignment4.ipynb to pdf  
[NbConvertApp] Support files will be in Assignment4_files/  
[NbConvertApp] Making directory ./Assignment4_files  
[NbConvertApp] Making directory ./Assignment4_files  
[NbConvertApp] Writing 115026 bytes to ./notebook.tex  
[NbConvertApp] Building PDF  
[NbConvertApp] Running xelatex 3 times: [u'xelatex', u'./notebook.tex',  
'-quiet']  
[NbConvertApp] Running bibtex 1 time: [u'bibtex', u'./notebook']  
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no  
citations  
[NbConvertApp] PDF successfully created  
[NbConvertApp] Writing 138391 bytes to Assignment4.pdf
```