Week 1

1. According to the Zipf Law approximation, approximately what frequency (express it has a percent) would the 3rd most popular word in a generic piece of text appear?

The 3rd most popular word in a generic piece of text appear at `0.1/3 = 3.33%`.

2. True or False: what is considered a stopword changes depending on the business context and dataset you are working with. If true, provide an example. If false, explain why it is false.

This can be either. Usually stop words in English may contain "a","the","is","are" and etc. They are commonly used to eliminate words that are so commonly used that they carry very little useful information. However, not in every situation, the stop word are all the same. For example, "not" is considered a stop word by many packages, and it can be a stop word if you are only counting the frequency of the topics that customers mentioned; however, it cannot be deleted directly when you are testing the sentiment of the context as it will change attitudes.

3. Explain to someone who does not understanding regular expressions what the following pattern is matching for: `b[a-z]+ly`.

It matches and return any string that starts with 'b', followed by at least one letter from a to z, and ends with `ly`*. For example* `badly`.

Week 2

1. Which of the encodings (ASCII, UTF-8, latin-1) will be able to encode this text: 사업

UTF-8. The characters have codepoints 49324 and 50629. ASCII can only represent the first 128 code points, and latin-1 can only represent the first 256 codepoints.

2. True or False: a Chinese character will take more space to encode than the character a.

True. Chinese characters have much, much higher Unicode codepoints than the character a, which is codepoint `97`. So UTF-8 will only need 1 byte to represent `a` but likely multiple bytes to represent Chinese characters.

3. Provide an example of how stemming can improve recall.

For example, to search for `athlete`, stemming improves recall by changing `athlete`, `athletes` and `athletics` to `athlet`, so that TP increases and recall increases (while TP+NF is constant).

Week 3

Calculate for the following documents:

```
Doc 1: He ate the food.

Doc 2: He liked the meal.
```

1. IDF: 2
2. TF for doc 1: 0
3. TF for doc 2: 1

Make sure to note that these answers depend on the TF and IDF functions!

Week 4

1. Which pair of words from the list of four below would have the closest similarity score using word2vec? Explain why based on your understanding of word2vec.

- happy
- hoppy
- cheerful
- derecha (Spanish word for right)

`happy` and `cheerful` will have the closest similarity since they can be substituted in the same context for each other with very little change in meaning. `hoppy` and `happy` would not have any similarity - this is not edit distance / fuzzy match. And `derecha` would not work and likely be an `OOV` token since most word2vec implementations are trained on single language corpus.

2. Identify all the named entities in the following document:

> Obama will return to the White House for the first time as Democrats look ahead to midterm elections

`Obama`, `White House`, `Democrats`, potentially `midterm elections`.

3. Write a named or unnamed capture group to extract email address' user names (the part before the `@`)

`r'(\w+)@\w+\.(?:com|net)'`

Week 5

1. You can use SVD to perform topic modelling by decomposing it into eigenvalues and eigenvectors.

False. SVD decompose matrix into document matrix, strength matrix and terms matrix while PCA decompose matrix into eigenvalues and eigenvectors.

2. The values within an HMM's emission matrix represent the likelihood of seeing the word given a target class and each row should sum to 1.

False. In our example, the columns should sum to 1.

3. The word sack would have different word2vec embedding value based on the context it is used in.

False. The value for `sack` will never change. `word2vec` produces static embeddings.

4. One of the benefits of reducing dimensions using SVD or PCA is improved feature explainability.

False. Both SVD and PCA project the original features into lower dimensional spaces that are linear combinations of the original features. Therefore the final reduced dimensions are not immediately human-interpretable.

Week 6

1. What is the difference between the regex patterns `C.*l` and `C.*?l`? Provide an example of how it is different.

`C.*l` returns greedy results while `C.*?l` returns lazy results. For text example like 'Cooloolool', the former returns 'Cooloolool' while the later returns 'Cool'.

2. What does this regex intended to do? `(?<=YUCHEN: )([\w @;\n\.,]+)(?=\nMIKE: |$)`.

It using positive lookaheads and lookbehinds to capture all text that is sent by Yu Chen that is ultimately followed by text from Mike.

3. True or False. The longer the sequence of a document, the more difficult it is for an RNN to model.

True. For an RNN model, the input has to be in a fixed length, as a result of which, longer sequence of a document would usually ends in loss of information. However, if the RNN hyperparameter is already set, then longer sequence of a document does not cause any additional parameters, since RNN generates same parameters w.r.t. all sequences.

4. True or False. Word embeddings are always trained via word2vec skipgram architecture neural networks.

False. They can be trained via CBOW, or completely different neural network architectures (see , or not even using neural networks (GloVe).

Week 7

1. True or False: A similar sized transformer model can be trained faster than an RNN. Explain why/why not. (Tom's lecture)

Generally true, transformer models perform all the sequence calculations at once, instead of sequentially like an RNN (they perform backpropagation through each layer of the transformer architecture, instead of through time).

2. True or False: BERT uses static embeddings that don't change based on the context. Whether true/false, explain why with an example. (Tom's lecture)

False - BERT uses attention mechanism, which will dynamically adapt the final embedding for a token depending on the context. Ex. "The bat made contact with the baseball" vs. "The bats in the cave are mammals", bat will have same embeddings in word2vec but same in BERT. Make sure there is an example.

3. What is the most interesting/rewarding thing you learned about in this course?

4. What is the difference between an encoder and a decoder in NLP? (Matt's lecture, a bit in Tom's lecture)

Encoder takes raw text, tokenizes, lookups up embeddings, outputs a hidden state that is a representation of the text. Decoder takes a hidden state vectors and turns them into raw text (the final output).

5. What is a potential business use case for NLG? (Matt's lecture)

Chatbots, question/answer services, text summarization, etc.