Overview

A. Industry applications of natural language processing

Natural language processing (NLP) is one of the most important tasks in industry, and it makes heavy use of machine learning. NLP deals with anything related to using machines to process and understand human text/speech (i.e. "natural language"). Many companies, including giants like Google and Amazon, have teams dedicated to NLP research and application in products.

B. How NLP works

In this course, we focus more on the text aspect of NLP, since speech-related tasks will utilize many disciplines outside of machine learning (e.g. signal processing, linguistics, etc.). With written text, it is much simpler to process the data into a feasible input for a machine learning model.

Text data is normally processed with relation to a *vocabulary*. The vocabulary just is the set of unique words that appear across all the text in the *corpus* (the set of documents used to train the model). There are many different ways to process text data, but in this course you'll learn how to convert sentences/documents into *embedding vectors*.

After processing the text data, we feed it into a particular type of neural network called a *recurrent neural network* (RNN). RNNs are great for dealing with sequential data like text, and in this course you'll be using the long short-term memory (LSTM) variation of RNNs.

The LSTM model can be adjusted to perform various NLP tasks, ranging from text classification to text generation. For different NLP tasks, the model will have different outputs.

C. What will this course provide?

After taking this course, you'll be able to process text data, train different LSTM models on the data, and use models to perform a variety of NLP tasks. Specifically, you will be able to:

- Process documents of text into embedding vectors
- Build a variety of different LSTm models for tasks ranging from text classification to text generation and machine translation

Introduction

An overview of natural language processing and word embeddings.

In this section, you will learn about using word embeddings to give numeric vector representations of words. Word embeddings are an extremely important part of natural language processing, since they allow us to use text with deep learning models.

A. What is NLP?

Natural language processing (NLP) encompasses any task related to machines dealing with natural language, i.e. human spoken or written language. Hence, it is one of the most important fields for machine learning and artificial intelligence. Tasks such as translating between languages, speech recognition, text analysis, and automatic text generation all fall under the scope of NLP. Without NLP we wouldn't have voice assistants like Siri and Alexa, or even search engines such as Google and Bing.

While machine learning for NLP is the focus of this course, not all NLP tasks require machine learning. For example, search engines rely largely on non-ML algorithms to find the most relevant documents or web pages for a given search query. Nevertheless, machine learning is becoming more and more widespread in NLP, and this trend will likely continue for many years to come.

B. Using text data

Natural language deals with two main categories of data: spoken or written data. While spoken language data is heavily used when building conversational agents like Siri and Alexa, written data (i.e. text data) is much more prevalent in industry NLP tasks. Raw text data is almost always unusable in NLP applications. The data is basically just a mass of strings without any real meaning for a machine to process. It is up to the engineer to first convert the raw text data into usable machine data, which can then be used as input for NLP algorithms.


```
{'off': 179, 'was': 20, 'like': 116, "isn't": 64, 'criminal': 163, 'falls': 122, 'late': 112, 'collide': 157,
 'passion': 184, 'bulk': 83, 'have': 193, 'many': 224, 'solid': 151, '1987': 79, 'guys': 49, 'credited': 107,
 'lost': 153, "could've": 186, 'recalled': 80, 'it': 5, 'least': 73, 'kenneth': 85, 'traditional': 190, 'betw
een': 161, "that's": 134, 'you': 140, 'often': 120, 'without': 21, 'contact': 169, 'my': 19, 'expressing': 22
1, "it's": 40, 'our': 175, 'car': 136, 'ships': 154, 'in': 6, 'making': 39, 'another': 127, 'secretary': 162,
 'to': 12, 'he': 34, 'at': 72, 'lots': 52, 'different': 155, 'wanted': 173, 'star': 65, 'though': 197, 'movie
s': 211, 'its': 124, 'feeling': 227, 'rml': 185, 'beautiful': 216, 'polluted': 58, 'for': 150, 'age': 128, 'd
eeply': 141, 'nolte': 218, 'is': 7, 'minutes': 192, 'glad': 105, 'from': 18, 'any': 167, 'directed': 28, 'dec
ide': 172, 'depths': 57, 'carradine': 68, 'once': 74, 'done': 187, 'duds': 203, 'type': 130, 'ever': 212, 'se
ries': 198, 'fun': 63, 'glenn': 31, 'pulls': 139, 'bit': 206, 'john': 67, 'great': 47, 'result': 103, 'makes'
: 26, 'simply': 182, 'rises': 56, 'very': 131, 'how': 101, 'toward': 170, 'vonneguty': 219, 'awful': 102, 'in
trigued': 149, 'and': 2, 'library': 202, 'blockbuster': 200, 'tv': 114, 'netflix': 199, 'just': 138, 'his': 9
0, 'portraying': 92, 'effective': 132, 'faith': 205, 'consummate': 183, 'interview': 76, "you'd": 60, 'footag
e': 88, 'strock': 27, 'leave': 117, 'hartford': 86, 'restored': 204, 'tony': 66, '100': 191, 'heroine': 177,
"couldn't": 171, 'book': 10, 'let': 146, 'every': 119, 'fangoria': 78, 'the': 1, 'does': 24, 'greatest': 210,
 'sexual': 158, 'had': 82, 'been': 35, 'main': 142, 'credit': 99, 'end': 33, 'night': 230, 'an': 75, 'which':
 36, 'women': 51, 'but': 9, 'characters': 32, 'thing': 133, 'palpable': 165, 'face': 125, 'job': 220, 'lake':
 59, 'would': 22, 'received': 97, 'thin': 207, 'made': 213, 'chase': 137, 'of': 4, 'films': 225, 'eisley': 25
, 'after': 46, 'reading': 44, 'this': 41, 'engage': 226, 'almost': 164, 'featuring': 89, 'so': 38, 'most': 54
, 'tough': 48, 'curious': 188, 'why': 115, 'only': 29, 'nick': 217, 'caper': 180, 'on': 37, 'into': 42, 'moth
er': 229, 'one': 209, 'swearing': 53, 'that': 8, 'hero': 176, 'reminder': 126, "i'm": 195, 'hardly': 166, 'ha
s': 108, 'alone': 147, 'entertainment': 121, 'hours': 152, 'children': 91, 'particularly': 94, 'even': 43, 's
omehow': 84, 'worth': 70, 'flat': 123, 'lee': 215, 'viewing': 71, 'film': 11, 'monster': 17, 'their': 45, 'th
ink': 61, 'cover': 145, 'feel': 110, 'be': 62, 'mid': 111, 'pull': 178, 'probably': 104, 'herbert': 81, 'back
ground': 118, 'as': 15, 'feelings': 222, 'two': 16, 'unconvincing': 55, 'local': 201, 'or': 181, 'sexy': 50,
```


Example of processed text data.

In the following chapters you'll be introduced to an easy and efficient way to process raw text, and then you'll use the processed text to run a machine learning algorithm.