Natural Language Processing

Section One: Overview & Representation

NLP definition

- Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language. [1]

- Natural language processing (NLP) is the discipline of building machines that can manipulate human language — or data that resembles human language — in the way that it is written, spoken, and organized. [2]

NLP History

1- 1964 ElIZA was created, designed to imitate a psychiatrist using **reflection techniques** (no computer understanding)

2- 1980s Several complicated statistical models were created, this was led by IBM

3- 1990 **N-Grams** have become useful, recognizing and tracking clumps of linguistic data, numerically.

4- In 1997, LSTM **recurrent neural net (RNN) models** were introduced, and found their niche in 2007 for voice and text processing. [3]

String Representation: Machine Representation

String					
Character ⇩					
S	t	r	i	n	g
1001	1002	1003	1004	1005	1006

Letter	ASCII Code	Binary
a	097	01100001

String Representation: one-hot encoding

Assume a sentence:
- I love cats not dogs
- I agree

Word	Encoding
I	[1,0,0,0,0,0]
love	[0,1,0,0,0,0]
cats	[0,0,1,0,0,0]
not	[0,0,0,1,0,0]
dogs	[0,0,0,0,1,0]
I	[1,0,0,0,0,0]
agree	[0,0,0,0,0,1]

String Representation: one-hot encoding cont.

In a dataset for text classification:

f1	f2	f3	f4	f5	f6	f7	Label	Text Represented
1	1	0	0	0	0	0	1	[I, love]
0	1	1	1	1	0	0	0	[love, cats, not]
1	1	0	0	0	1	0	1	[I, love, I, dogs]
1	0	0	0	0	0	1	1	[I, dogs, agree, agree]

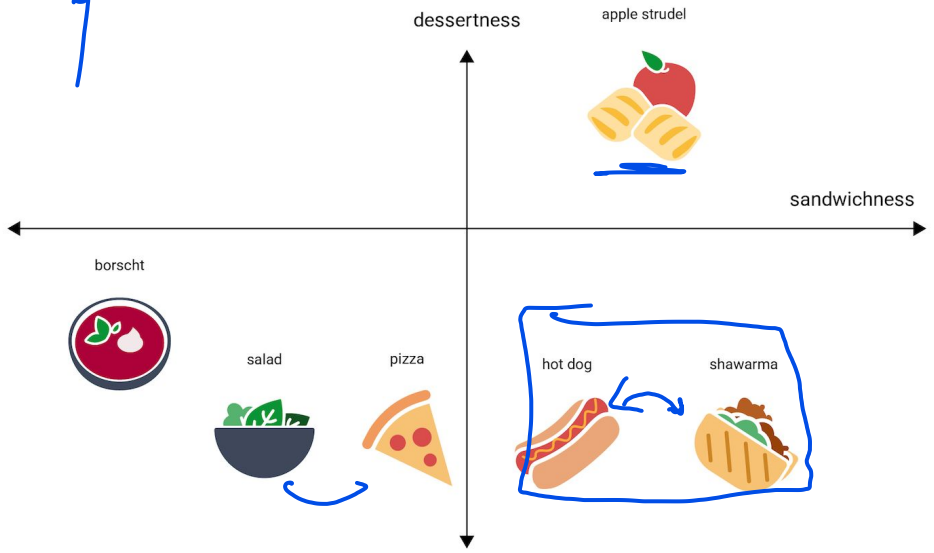
String Representation: Bag of Words

It's just one-hot encoding but with an extra step.
Which is counting how much each word is repeated.

f1	f2	f3	f4	f5	f6	f7	Label	Text Represented
1	1	0	0	0	0	0	1	[I, love]
0	1	1	1	1	0	0	0	[love, cats, not]
2	1	0	0	0	1	0	1	[I, love, I, dogs]
1	0	0	0	0	0	2	1	[I, dogs, agree, agree]

String Representation: Embedding Space

- An embedding is a vector representation of data, in our case it's words/tokens. This representation is reflection of its meaning.
- The "meaning" is its relative position of it in the embedding space (vector space of embeddings) to similarly semantic words.



String Representation: Embedding Space

Assume a sentence:
- I love cats not dogs
- I agree

Word	Encoding
I	[0.1, -0.2, 0.4, 0.8, -0.1, 0.3]
love	[0.5, 0.7, -0.3, 1.1, 0.4, -0.2]
cats	[0.8, -0.5, 0.6, 0.3, -0.7, 0.9]
not	[0.2, -0.4, 0.1, 0.5, 0.3, -0.6]
dogs	[0.7, -0.8, 0.2, 0.9, -0.3, 0.6]
I	[0.1, -0.2, 0.4, 0.8, -0.1, 0.3]
agree	[0.6, 0.9, -0.1, 1.3, 0.2, -0.4]

String Representation: Embedding Space cont.

In a dataset for text classification:

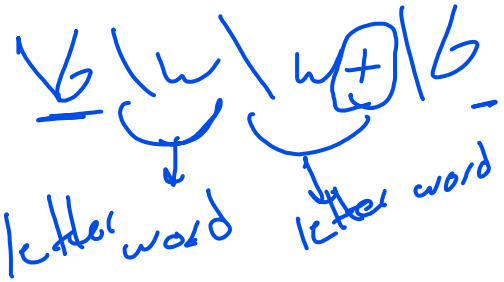
- We represent a sentence by pooling the embedding.
- Pooling is summing the vectors then dividing it by the number of vectors.

f1	f2	f3	f4	f5	f6	Label	Text Represented
0.3	0.25	0.05	0.95	0.15	0.05	1	[I, love]
0.5	-0.067	0.13	0.63	0	0.03	0	[love, cats, not]
0.35	-0.125	0.175	0.9	-0.025	0.25	1	[I, love, I, dogs]
0.5	0.2	0.1	1.075	0	0.025	1	[I, dogs, agree, agree]

Comparing representation methods

Aspect	One-Hot Encoding	Bag of Words	Embedding Vectors
Vector size	Large (vocab size)	Large (vocab size)	Small (50-300 dims)
Semantic understanding	None	None	High
Memory usage	Poor	Poor	Good
Word relationships	No	No	Yes (similar words close)
Training needed	No	No	Yes
Best for	Simple classification	Document classification	Modern NLP tasks

Coding



1- Create one-hot encoding from scratch

2- Create Bag of Words embedding from scratch

3- Use and explore Embedding Space in Word2Vec and Glove

4- Bonus: plotting the vectors representing words after processing them using PCA.

NLP is the study of lang