5.08 Word Vectors

word2vec

- Word2vec is a function **that accepts a text corpus as an input** and returns a set of vectors (also known as embeddings) where each vector is a numeric representation of a given word.

- Each value in the vector represents a context (surrounding) word of the given word.

- Widely used in recommender systems, sentiment analysis, and content labelling.

Key Parameters of Word2Vec

- **vector_size**
  - size of word vector. Words included in vector must be within window size
- **window**
  - number of words before and after the focus word that will be considered as context
- **min_count**
  - number of times word must appear in corpus in order to be included in vector

Important - Model will only be trained on words that meet min_count so model may not learn all words as a result

Word2vec Example – Word Vector

Data Set Row 1	I am the **King** of United Kingdom window words window words
Data Set Row 2	The new **King** was appointed in 2022 window words window words

Word Vector for "King" using params below

vector_size = 10

window_size = 3

min_count = 1

I	am	the	of	United	Kingdom	new	was	appointed	in
1	1	2	1	1	1	1	1	1	1

After Normalization

```
[0.2773501, 0.2773501, 0.5547002, 0.2773501, 0.2773501, 0.2773501, 0.2773501, 0.2773501, 0.2773501, 0.2773501]
```

Word2vec Example – Sentence Vector

I										
am										
the										
King										
of										
United										
Kingdom										

Row (Sentence) Vector for "Data Set Row 1" using Average
Take the average of all values in vectors of all the words row of data

Row Vector										

Importance of word2vec

- The structure of language has a lot of valuable information in it. The way we organize our text/speech tells us a lot about what things mean.

- By using Machine Learning to "learn" about the structure and content of language, our models can now organize concepts and learn the relationships among them.

- E.g. we did not explicitly tell the computer what "dog" or "puppy" or "cat" or "kitten" actually mean. But by learning from how words are being used in proximity, our model can quantify the relationship among these entities!

word2vec

- The idea is that we can use the position of words in sentences (i.e. see which words were commonly used together) to understand their relationships *(like how we previously selected features for regression based on correlation)*

  - If "dog" and "puppy" are used near one another a lot, then it suggests that there may be some sort of relationship between them.

  - If "cat" and "dog" are used near similar words a lot (i.e. "pet"), then it suggests that there may be some sort of relationship between them.