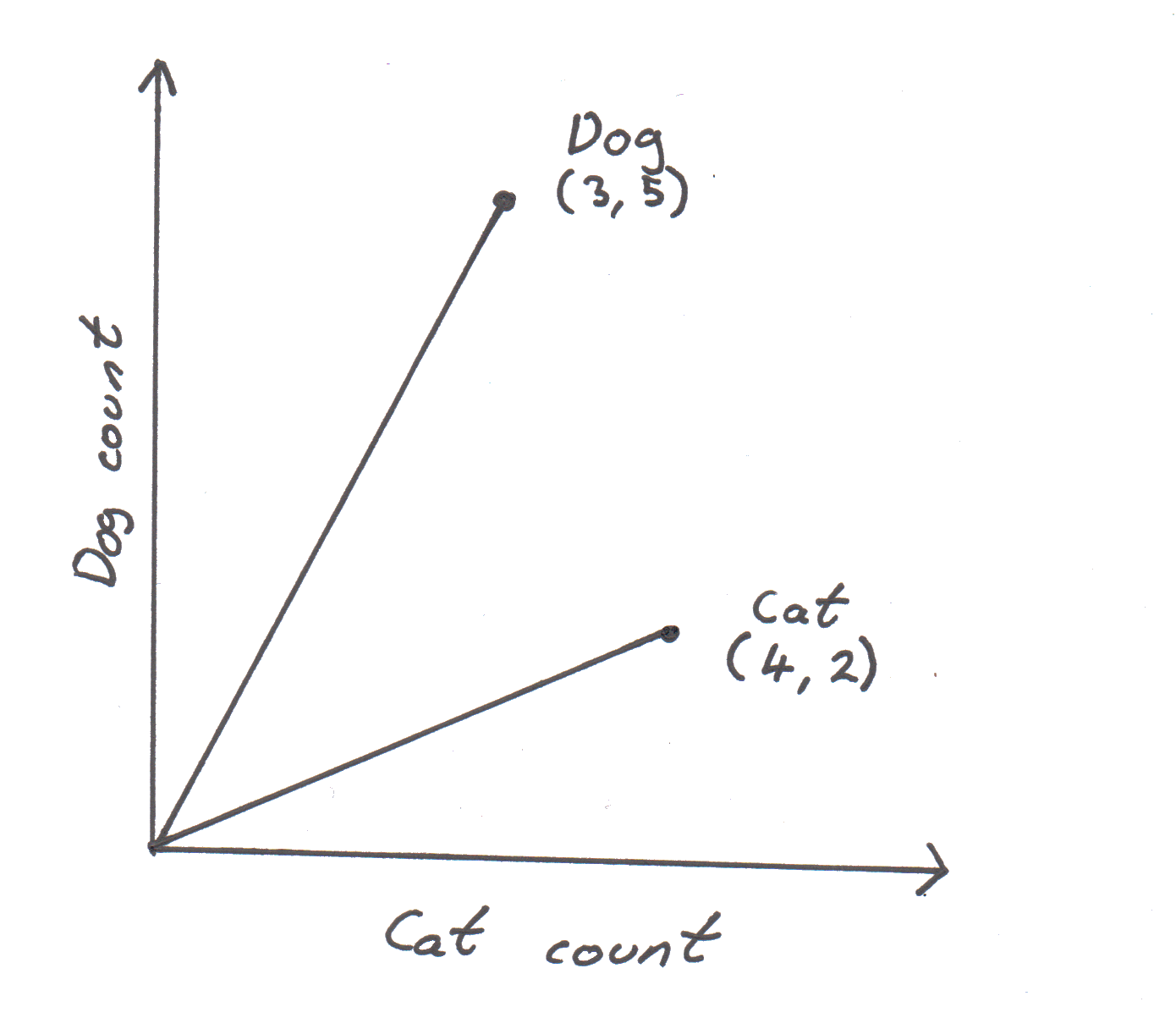
Data Collection and Cleaning Report

# The Algorithm

The first step was to find the number of times each keyword appeared in each set of articles. This data could then be transformed into a vector for each keyword, with each axis representing a keyword and the value the number of occurrences in that keyword’s articles. Using this we can find the cosine of the angle between any two of these vectors to find the semantic distance between these words.

Shape

Description automatically generated with low confidence

Θ

If we had two words (for example dog and cat) we can easily plot a graph to show this process. In this example there were five articles each for dog and cat, and the number of times each word appeared in the articles was recorded. These items can be plotted onto a graph of count in each set of articles and the angle (Θ) between the two lines from the origin calculated. The cosine of this is the semantic distance between the two words. To adapt this for more words you can increase the number of axes endlessly.

The distance between any two vectors can be calculated using the formula:

Where A is the first keywords vector, B is the second keywords vector and S is the semantic distance between the two keywords.

In this scale a distance of one means the words are semantically identical and zero means they are not similar at all.

# Advantages

An important part of the algorithm is that how often a word is used does not affect its semantic distance to other words as a word being used more just causes its vector to increase in magnitude and not change angle. This stops much more common words being unfairly favoured compared to less common ones that might be more similar to the word in question.

Another advantage is that second degree relationships can be taken into account. This means how often two words appear in other words article affects the angle between them. This adds another layer to the analysis to produce the distances as many of these keywords could be used in similar instances and may be very similar to each other, but due to the nature of reporting might not be relevant to a story, for example there might be sperate news stories about cats and dogs, which don’t mention the other animal, leading to a big distance between the words. However, both are pets, and pet articles would decrease this distance between the two words, giving a better result.

A good feature is that any keyword is guaranteed to get a distance of one from a keyword to itself as it has the same vector as itself without having to manually change this.

# The Charts

The first two charts so the semantic distances between the keywords just using the data from the BBC articles, the second two are for Wikipedia articles and the last two are the average between the two. The heatmaps are an easy way to see which words are close to each other and allow for all the data to be seen in one graph. The chord graphs on the other hand allow for a much more visual representation where the distances are determined by the thickness of the chords. This also makes it easy to see which words are close to the most other words due to their longer arc length.

Chart

Description automatically generated with medium confidenceChart, radar chart

Description automatically generated

Chart

Description automatically generated with medium confidence

Chart, radar chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Chart, radar chart

Description automatically generated