Interquartile range is a data dispersion measure which is just the 3rd quartile value minus the 1st quartile value. It an improvement over using the range as it is better able to identify outliers and gives us some idea of how the data is distributed throughout the range.

The median is a measure of central tendency which tells us the middle value in a sorted list. This has advantages over using a simple average as it is less susceptible to outliers and is better at describing data with “tails”.

Cosine similarity is a, well, similarity measure. It is the cosine of the angle between two non-zero vectors. It has a range of [-1, 1], with 1 being perfectly similar and -1 being perfectly opposite. This will probably be a useful metric later, when we begin identify how similar different documents are.