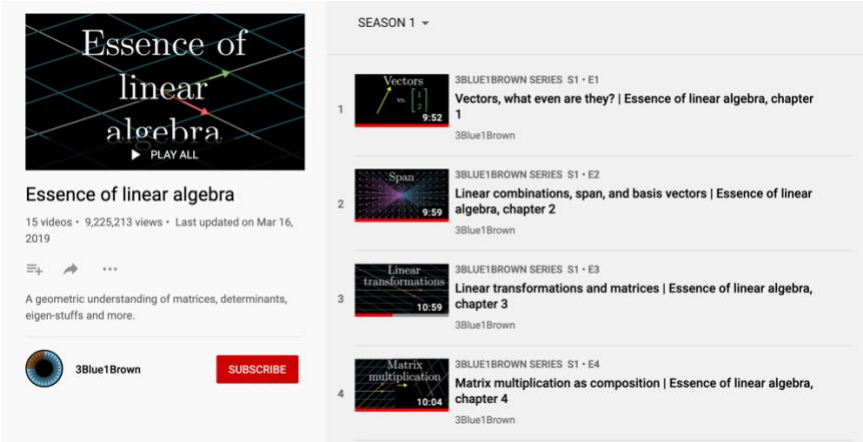
WEEK 3 – CONNECTING THE DOTS

1. Linear Algebra



   As we will be working on a task related to Text Processing, so it is extremely important to get familiar with fundamentals of Linear Algebra. Essence of Linear Algebra channel by 3Blue1Brown is amazing. Remember that it is more important to understand that intuition behind Vectors and Matrices, Matrix multiplication, Eigen Space etc. than doing the actual Matrix multiplication or matrix decomposition, the computer would do the latter for you.

   Please note that the total duration of all the videos is less than 2.5 hrs (Episode/Chapter 10, 11, 12 are optional).

   Solid foundation here will make you easily understand Word Vectors, Embeddings etc.

2. Read more on how scikit-handles text data, however you are free to use any other framework https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html, https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

3. For more in depth introduction to TF-IDF (Term Frequency – Inverse Document Frequency), Word Vectors etc. kindly refer to the attached chapter from Daniel Jurafsky open source book on Speech and Language Processing. Dan is a professor of Linguistics and Computer Science at Stanford University. One of the best resources to get a great conceptual understanding of

Natural Language Processing along with applications is his open book that can be accessed here: https://web.stanford.edu/~jurafsky/slp3/

4. [A few thought provoking questions, please remember that the core here is Linear Algebra and the above videos will help you in getting a solid foundation]

    a) Is your TF-IDF matrix a sparse matrix?
    b) What are word vectors?
    c) How can you do dimensionality reduction of your TF-IDF matrix in scikit?
    d) What does lower or latent dimensions mean here for our word vectors? Is it something similar to word embeddings?
    e) How can we capture local context information i.e. what is the drawback of term-document matrix? What are word context vectors?

    Please note that once you are able to think through above questions/scenarios it would be very easy to understand things like Word2Vec or Glove or even BERT (with little more effort ☺)