Singular Value Decomposition is a method of generating a series of rank one matrices (matrices with only one linearly independent column) that sum to approximate a matrix. It will work with a matrix of any shape.

The first step is finding three specific matrices, U, Σ, and V that decompose the given matrix A such that

A = UΣVT

This is very similar to the A = PDP-1 decomposition, where P is the horizontal concatenation of the eigenvectors of A, and D is a diagonal matrix of the associated eigenvalues.

Since we cannot find the eigenvectors of a nonsquare matrix, we instead find the eigenvectors of two matrices that are derived from A and are square and symmetrical no matter A’s shape: AAT and ATA. If A is an MxN matrix, the sizes of these matrices will be NxN and MxM, respectively.

The nonzero eigenvalues of AAT and ATA are the same. Their associated eigenvectors ui and vi are related by ui = Avi. The columns of U, an NxN matrix, are the eigenvectors of AAT in order of decreasing associated eigenvalues, and the columns of V, an MxM matrix, are the associated eigenvectors of ATA. Rather than the eigenvalues, Σ is a diagonal matrix of the singular values of A, ‘filled in’ with zeros so it has size MxN. The singular value .

The best rank 1 approximation of A can then be calculated as u1v1T. The best rank 2 approximation is u1v1T +u2v2T , and so on.

HOW WE USED IT

We used a Python script to generate an adjacency matrix (such as the one below for English) showing how frequently one letter followed another in the words of a large sample text. This matrix is conveniently square.

[insert colorized adjacency matrix here]

Maybe explain plotting and leave this stuff for the bottom right corner?

We then found U, Σ, and V for the adjacency matrix. In a first-rank approximation, all the rows are scalar multiples of v1T and all the columns are scalar multiples of u1. Therefore, the ith element of u1 increases with the frequency that the ith letter in the alphabet precedes other letters, and the ith element of v1 increases with the frequency that the ith letter in the alphabet follows other letters. Plotting these two values on the Cartesian plane as (u1, v1) shows both the relative frequency of the letters, with more common letters landing further from the origin, and the frequency of those letters beginning or ending a word, with letters occurring mostly at the beginning of words landing further from the x-axis and letters occurring mostly at the end of words landing further from the y-axis.

Second-rank:

The second singular values of the matrix are able to show what letters are most commonly preceded by a vowel or a consonant. To identify a vowel the orognal matrix is looked at and the average frequency for how often a letter occurs before or after another letter is found. The vowels are the letters that surpass the average frequency per letter for the most letters. For example if the average frequency for a letter to occur before B is 61 then the frequency for A to occur before B (628) is significantly higher. A has a higher frequency than the average for being in front of a letter for 18 letters, significantly more than half. A similar pattern can be seen for all vowels. After using this method to identify what is considered a vowel and a consonant the second singular value is able to sort what type of letter a letter is preceded and followed by as seen in the Second Singular Values figure. If a letter is more commonly preceded by a vowel it is more negative on the x-axis and more positive on the y-axis. In english vowels are typically preceded and followed by consonats and thus show up

The second singular values of the matrix are able to show what letters are most commonly preceded by a vowel or a consonant. To identify a vowel the orognal matrix is looked at and the average frequency for how often a letter occurs before or after another letter is found. The vowels are the letters that surpass the average frequency per letter for the most letters. For example if the average frequency for a letter to occur before B is 61 then the frequency for A to occur before B (628) is significantly higher. A has a higher frequency than the average for being in front of a letter for 18 letters, significantly more than half. A similiar pattern can be seen for all vowels. After using this method to identify what is considered a vowel and a consonant the second singular value is able to sort what type of letter a letter is preceded and followed by. In English and German vowels are typically preceded and followed by consonants and thus show up in quadrant II of the secondary singular value graph. A letter like h, which in english is often preceded by t and followed by e, shows up in quadrant I. Consonants show up in quadrant IV as they are preceded and followed by vowels and letters that are preceded by vowels and followed by consonants show up in quadrant III.