Data Science Assignment 4

Nike Marie Pulow – Henri Paul Heyden
stu239549 – stu240825

1 PCA, SVD

1. We search for matrices U, Σ, V with $X = U\Sigma V^T$ so that the columns of U and V are eigenvectors of XX^T and X^TX respectively.
   We'll begin by computing U:

$$\begin{aligned} XX^T &= \begin{bmatrix} 4 & 2 \\ \sqrt{2} & 2\sqrt{2} \\ -\sqrt{2} & -2\sqrt{2} \end{bmatrix} \times \begin{bmatrix} 4 & \sqrt{2} & -\sqrt{2} \\ 2 & 2\sqrt{2} & -2\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} \end{aligned}$$

2.

2 TF-IDF

1. Vocabulary: ["fast", "car", "highway", "road", "bike", "wheel"]

2. term frequencies:

t	$d=1$	$d=2$	$d=3$	$d=4$
fast	0.2	0.4	0.2	0
car	0.4	0.4	0	0.25
highway	0.2	0	0.2	0
road	0.2	0	0.4	0
bike	0	0.2	0	0.25
wheel	0	0	0.2	0.5

3. document frequency:

t	$\#d$ containing t	IDF
fast	3	$\log \frac{4}{3}$
car	3	$\log \frac{4}{3}$
highway	2	$\log \frac{4}{2} = \log 2$
road	2	$\log(2)$
bike	2	$\log(2)$
wheel	2	$\log(2)$

4. calculating TF-IDF vectors:

t	$d=1$	$d=2$	$d=3$	$d=4$
fast	0.025	0.05	0.025	0
car	0.05	0.05	0	0.031
highway	0.06	0	0.06	0
road	0.06	0	0.120	0
bike	0	0.06	0	0.075
wheel	0	0	0.06	0.151

Which gives us the following vectors for each d:

$r(d = 1) = (0.025, 0.05, 0.06, 0.06, 0, 0)$
$r(d = 2) = (0.05, 0.05, 0, 0, 0.06, 0)$
$r(d = 3) = (0.025, 0, 0.06, 0.12, 0, 0.06)$
$r(d = 4) = (0, 0.031, 00, 0.075, 0.151)$