Assignment 5A

Nolan McCaffery and Daniel Rosenbaum

15 March 2019

Part A: Warm Up

1. Read!

2.

   (a) co-occurence vectors for `b, d`

   $$b = [a : 4,\ b : 0,\ c : 4,\ d : 0,\ e : 0,\ f : 0,\ g : 0,\ h : 0]$$

   $$d = [a : 2,\ b : 0,\ c : 2,\ d : 0,\ e : 2,\ f : 0,\ g : 2,\ h : 0]$$

   (b) IDFs for each word

   ```
   idf(a) = log(6/3) = 0.301
   idf(b) = log(6/2) = 0.477
   idf(c) = log(6/3) = 0.301
   idf(d) = log(6/2) = 0.477
   idf(e) = log(6/2) = 0.477
   idf(f) = log(6/1) = 0.778
   idf(g) = log(6/2) = 0.477
   idf(h) = log(6/1) = 0.778
   ```

   (c) TF-IDF vectors

   $b = [a : 1.20411998,\ b : 0,\ c : 1.20411998,\ d : 0,\ e : 0,\ f : 0,\ g : 0,\ h : 0]$

   $d = [a : 0.60205999,\ b : 0,\ c : 0.60205999,\ d : 0,\ e : 0.95424251,\ f : 0,\ g : 0.95424251,\ h : 0]$

   (d) length normalized TF-IDF co-occurence vectors for `b, d`

   $b = [a : 0.70710678,\ b : 0,\ c : 0.70710678,\ d : 0,\ e : 0,\ f : 0,\ g : 0,\ h : 0]$

   $d = [a : 0.37731249,\ b : 0,\ c : 0.37731249,\ d : 0,\ e : 0.59802615,\ f : 0,\ g : 0.59802615,\ h : 0]$

   (e) final distance between words `b` and `d` using `TF-IDF` and `EUCLIDEAN`
   `sim(b, d)` $= 0.9658$