CSI 772
Homework 1
Bryan Adams

Problem 1

1a

Both resulting matrixes are identical. This makes since because we are only apply transformations to the data in a different order. Both IDF and word count normalization depends on the original data and not the previous calculation, therefore, the order does not matter which order you apply. See Appendix A for python code used to normalize the data.

1b

Article "tmnt raph" is the closest to "tmnt mike", reference Table 1 for distance calculations.

Table 1: Euclidean distance between given document and the document named “tmnt mike”

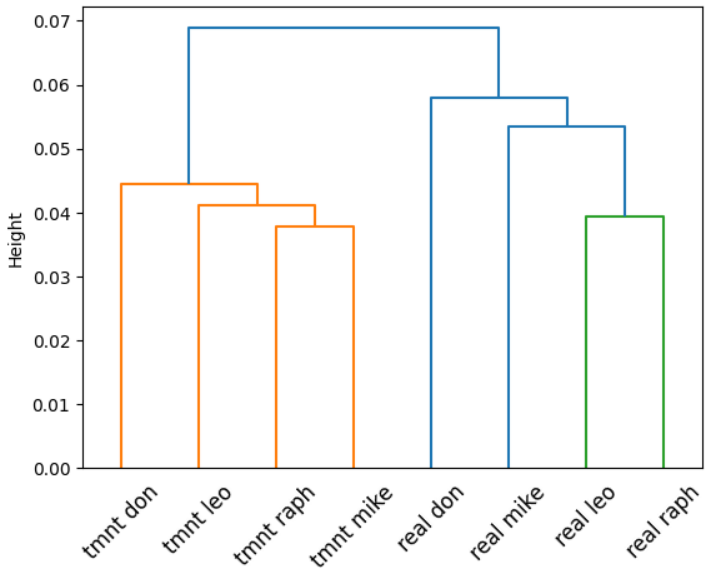
Article	Euclidean distance to "tmnt mike"
tmnt leo:	0.0412
tmnt raph:	0.0378
tmnt don:	0.0420
real leo:	0.0489
real raph:	0.0449
real mike:	0.0475
real don:	0.0638

1c

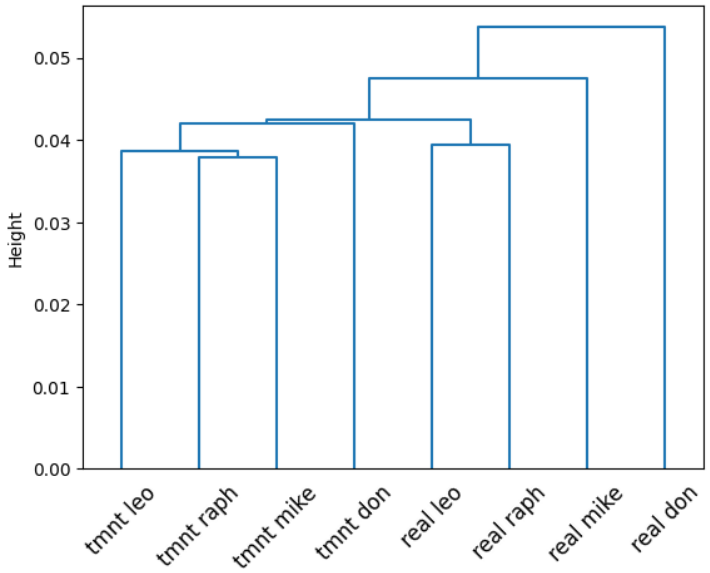
The distance matrix is found in Table 2. When running hierarchical agglomerative clustering, complete linkage provides the best clusters when $K = 2$ because it will cluster the Teenage Mutant Ninja Turtles articles together and the artist articles together, reference Figures 1a and 1b for the dendograms.

Table 2: Distance matrix for documents

	tmnt leo	tmnt raph	tmnt mike	tmnt don	real leo	real raph	real mike	real don
tmnt leo	0.0000	0.0386	0.0412	0.0444	0.0425	0.0496	0.0648	0.0671
tmnt raph	0.0386	0.0000	0.0378	0.0429	0.0550	0.0462	0.0668	0.0688
tmnt mike	0.0412	0.0378	0.0000	0.0420	0.0489	0.0449	0.0475	0.0638
tmnt don	0.0444	0.0429	0.0420	0.0000	0.0537	0.0504	0.0619	0.0537
real leo	0.0425	0.0550	0.0489	0.0537	0.0000	0.0394	0.0534	0.0553
real raph	0.0496	0.0462	0.0449	0.0504	0.0394	0.0000	0.0504	0.0544
real mike	0.0648	0.0668	0.0475	0.0619	0.0534	0.0504	0.0000	0.0580
real don	0.0671	0.0688	0.0638	0.0537	0.0553	0.0544	0.0580	0.0000



(a) Hierarchical Agglomerative Clustering with complete linkage



(b) Hierarchical Agglomerative Clustering with complete linkage

1d

The top 20 most common words and their word count are found in Table 3. They make up 24.3425% of the word count.

Table 3: Top 20 most common words

Word	the	and	his	was	leonardo	that	for	with	michelangelo	raphael
Count	2664	1127	636	453	342	297	282	279	277	212
Word	from	this	which	turtles	donatello	him	series	were	who	one
Count	209	189	129	127	117	116	115	111	110	103

1e

Our collection of 8 wikipedia articles appears to follow Zipf's law. When looking at Figure 2, you can see the data is generally linear with a slope = 1. This implies that the number of times words appear in our article generally follows Zipf's law.

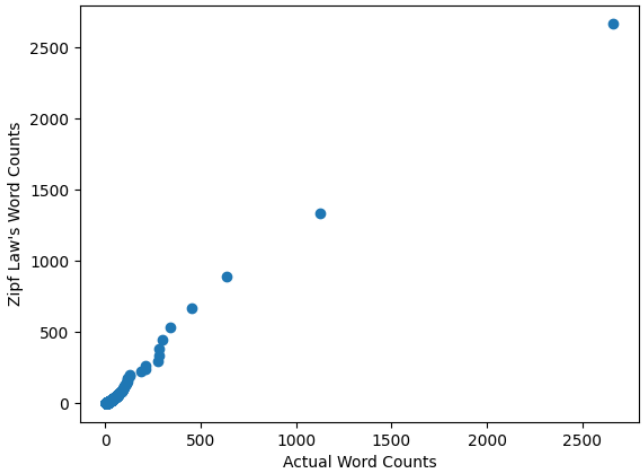


Figure 2: Word counts in our articles vs Zipf's Law theoretical word counts

Problem 2

See attached solutions.