How to Detect Plagiarism, or "Who *Really* Wrote Shakespeare's Plays?"
(Institution A)

Warmup
Discuss what it means for two objects to be "similar." What features are important, and what are non-essential? Then sketch the following vectors:

$$\mathbf{u} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} -7 \\ -4 \end{bmatrix}, \mathbf{t} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{s} = \begin{bmatrix} 4 \\ -7 \end{bmatrix}$$

Which of the vectors **v, w, t, s** is *most* similar to **u**? Which is least similar to **u**? Justify your answers.

Part One

1. Explain how you could use the dot product to determine whether two vectors are similar. Give specific details about how the dot product can be used to answer this question.
2. Suppose you have the vector **a**. Of all possible vectors, what vector is most similar to **a**? What will be the dot product of these two vectors?
3. Of all vectors, what vector is *least* similar to **a**? Justify your answer.
4. Are the vectors **a** and -**a** similar? Justify your answer *without* referring to the dot product; then justify the same answer by *referring* to the dot product.
5. Are the vectors **a** and 3**a** similar? Justify your answer *without* referring to the dot product; then justify the same answer by *referring* to the dot product.

Part Two
A common use of the dot product is to compare two groups using a *frequency vector*. For example, we might ask two groups of people what device they use most to access the Internet and obtain the following results:

	Group A	Group B
Desktop	25	40
Laptop	30	70
Phone	80	25
None	1	15

This would give us the frequency vector $\begin{bmatrix} 25 \\ 30 \\ 80 \\ 1 \end{bmatrix}$ for group A, and for $\begin{bmatrix} 40 \\ 70 \\ 25 \\ 15 \end{bmatrix}$ group B. The dot product of the *normalized* frequency vectors is one way to measure the similarity of the two groups.

1. Explain why the dot product must be done with the *normalized* frequency vectors.
2. Suppose two groups are very similar. What would you expect the dot product of their frequency vectors to be? Justify your answer.

Part Three
One of the more unusual uses of the dot product is to identify the author of a text, when the authorship is either unknown or disputed. For example, did Shakespeare actually write the plays attributed to him, or were they written by Francis Bacon, Queen Elizabeth, Walter Raleigh, or the Earl of Oxford, the last being the subject of the recent film Anonymous (2011)?

The problem is approached as follows. First, a group of common words are selected; studies show that the best words to select are "function" words that are almost meaningless by themselves: for example, "the",

"and", "an", "of", etc. Next, two frequency vectors are constructed: one from the works known to be by the author in question, and one for the work whose authorship is disputed. Finally, the dot product of the normalized vectors is found. If the dot product of the normalized vectors meets certain requirements, then the two works are judged to be by the same author; otherwise, the authorship remains in question.

On the last page are three texts (where "xxx" is a proper name that is irrelevant). Complete the table below to construct the frequency vectors, then use the frequency vectors to determine which of the two texts are probably by the same author.

Word	and	to	not	the	is	for	as	that
Text 1								
Text 2								
Text 3								

1. Using your frequency vectors, determine which of the two texts are most like each other, and so (probably) by the same author. Justify your answer mathematically.
2. Can you "beat the system" and produce a paragraph of text that a frequency analysis (like the above) will say is very similar to Text 1, even though by any reasonable assessment it could not be? If you can, provide such a text. If you can't, why not?

Bonus

1. Why is it more useful to use the words "and", "to", "not", etc., as opposed to words like "cowardice", "money", "provinces", etc.?
2. A friend in class suggests that frequency vectors are used by dating sites to measure compatibility. Discuss how this might work, or what problems might arise with this application.

References

P. D. Turney, P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37 (2010) 141-188. Available at http://www.jair.org/media/2934/live-2934-4846-jair.pdf

Texts for Analysis

Text One

And none need think it cowardice for a number of confederates to pause before they attack a single city. The xxx have allies as numerous as our own, and allies that pay tribute, and war is a matter not so much of arms as of money, which makes arms of use. And this is more than ever true in a struggle between a continental and a maritime power. First, then, let us provide money, and not allow ourselves to be carried away by the talk of our allies before we have done so: as we shall have the largest share of responsibility for the consequences be they good or bad, we have also a right to a tranquil inquiry respecting them.

Text Two

At first the xxx trusted the words of xxx, through their friendship for him; but when others arrived, all distinctly declaring that the work was going on and already attaining some elevation, they did not know how to disbelieve it. Aware of this, he told them that rumours are deceptive, and should not be trusted; they should send some reputable persons from xxx to inspect, whose report might be trusted. They dispatched them accordingly. Concerning these xxx secretly sent word to the xxx to detain them as far as possible without putting them under open constraint, and not to let them go until they had themselves returned.

Text Three

We are also told, that in the provinces he constantly maintained two tables, one for the officers of the army, and the gentry of the country, and the other for xxx of the highest rank, and provincials of the first distinction. He was so very exact in the management of his domestic affairs, both little and great, that he once threw a baker into prison, for serving him with a finer sort of bread than his guests; and put to death a freed-man, who was a particular favorite, for debauching the lady of a xxx knight, although no complaint had been made to him of the affair.