Name:

1. (**2 Pts.**) List two common applications of text mining:

Any two of:

- Predictive/Supervised or unsupervised models that include customer center notes, website forms, e-mails, and Tweets, or other social media text
- Spam Detection
- Document Categorization (Clustering)
- Topic Extraction
- Information Retrieval
- Anomaly Detection
- Processing large numbers of legal documents
- Hospital admission prediction models incorporating medical records notes as a new source of information
- Insurance fraud modeling using adjustor notes
- Sentiment categorization from customer comments
- Stylometry or forensic applications that identify the author of a writing sample

(Other reasonable examples considered)

2. **(6 Pts.)** Follow the directions from top to bottom to create a document by term matrix.

1. Remove all punctuation
2. Stem all nouns to singular form
3. Remove all stop words
4. Remove all terms that are less than or equal to 4 characters in length
5. Create a document by term matrix with terms as columns and documents as rows

Stoplist: a about at but for is it me than thing to was you

Document 1: to err is human, but to really foul things up you need a computer
Document 2: computer science is no more about computers than astronomy is about telescopes
Document 3: a computer once beat me at chess, but it was no match for me at kick boxing

	human	really	computer	science	astronomy	telescope	chess	match	boxing
doc 1	1	1	1	0	0	0	0	0	0
doc 2	0	0	2	1	1	1	0	0	0
doc 3	0	0	1	0	0	0	1	1	1

(Column and row order don't matter)

Name:

3. **(2 pts.)** Consider another unrelated term by document matrix ***A***. ***A*** has ***N*** rows which represent terms and ***p*** columns which represent documents. (We say ***A*** is an ***N*** x ***p*** matrix.) We use singular value decomposition (SVD) to extract ***k*** SVD features from ***A***. Given that SVD follows the well-known equation:

$$A = U\Sigma V^T$$

where ***U*** is an ***N*** x ***k*** matrix and $\boldsymbol{V^T}$ is an ***k*** x ***p*** matrix.

Is ***U*** or ***V*** more ideal to analyze the relationship between ideas in the documents and each term?

U

Is ***U*** or ***V*** more ideal to analyze the relationship between ideas in the documents and each document?

V