DNSC 6279 Name:
Quiz 08

1. **(6 Pts.)** Follow the directions from top to bottom to create a document by term matrix.

- Remove all URLs.
- Remove all @-mentions (users names preceded by '@').
- Remove all non-alphabetic characters (like numbers and punctuation).
- Remove all terms that are less than 10 characters in length.
- Create a document by term matrix with terms as columns and documents as rows. Do not consider the case of a token when counting tokens for the document by term matrix.

Document1: .@jehanramez also check out a new addition: https://github.com/sassoftware/dm-flow ⍰

Document 2: Good input @Petzoldt! CV is now in @SASSoftware PROC GLMSELECT & EM decision tree and least angle regression nodes. U want more?

Document 3: More @SASsoftware #machinelearning resources on @Github http://ow.ly/Tb3cx #datascience #SASUsers

Document 4: "Project Freedom" by @SebastianThrun on @LinkedIn https://www.linkedin.com/pulse/project-freedom-liberate-your-low-performers-sebastian-thrun ⍰

Document 5: Intuitive explanations for advanced #machinelearning and #deeplearning concepts: http://colah.github.io/ - Thanks @ch402!

Document1: ~~.@jehanramez also check out a new addition: https://github.com/sassoftware/dm-flow ⍰~~

Document 2: ~~Good input @Petzoldt~~! ~~CV is now in @SASSoftware PROC GLMSELECT & EM decision tree and least angle~~ regression ~~nodes. U want more?~~

Document 3: ~~More @SASsoftware~~ #machinelearning ~~resources on @Github http://ow.ly/Tb3cx~~ #datascience ~~#SASUsers~~

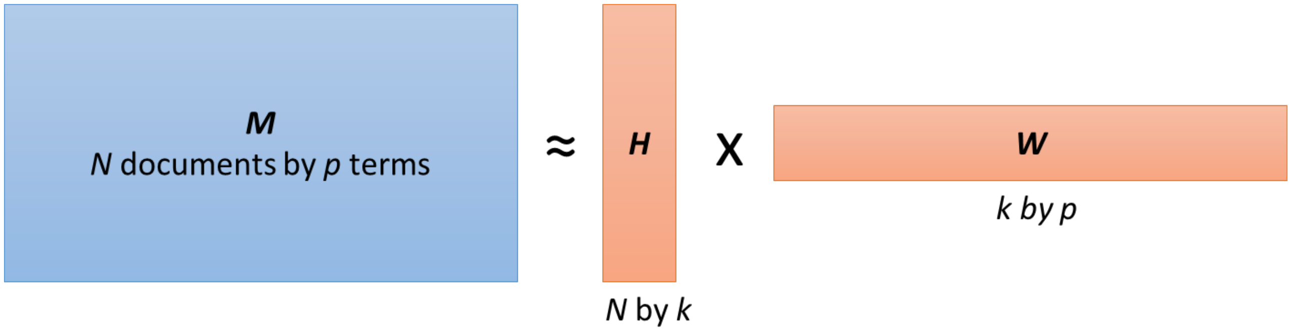
Document 4: ~~"Project Freedom" by @SebastianThrun on @LinkedIn https://www.linkedin.com/pulse/project-freedom-liberate-your-low-performers-sebastian-thrun ⍰~~

Document 5: ~~Intuitive~~ explanations ~~for advanced~~ #machinelearning ~~and~~ #deeplearning ~~concepts: http://colah.github.io/ - Thanks @ch402!~~

	regression	machinelearning	datascience	explanations	deeplearning
Document 1	0	0	0	0	0
Document 2	1	0	0	0	0
Document 3	0	1	1	0	0
Document 4	0	0	0	0	0
Document 5	0	1	0	1	1

2 pts. each for documents 2,3 and 5.

2. **(2 pts.)** Much like singular value decomposition (SVD), non-negative matrix factorization (NMF) is often used in text mining to factorize a wide, sparse document-by-term matrix, ***M***, into two smaller, dense matrices ***H*** and ***W***. The size of ***H*** and ***W*** is determined by ***M*** and the number of features desired, *k*.



Which matrix would be more suitable for creating clusters of documents?
H (1 pt.)

Which matrix would be more suitable for interpreting topics in the space of the terms?
W (1 pt.)

3. **(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)