

Quiz 8

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

- Stem all nouns to singular form
- Remove all stop words
- Remove all terms that occur in only one document
- 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

	computer	no
Document 1	1	0
Document 2	2	1
Document 3	1	1

2. (4 pts.) Consider another unrelated term by document matrix \mathbf{A} . \mathbf{A} has \mathbf{N} rows which represent terms and \mathbf{p} columns which represent documents. (We say \mathbf{A} is an $\mathbf{N} \times \mathbf{p}$ matrix.) We use singular value decomposition (SVD) to extract \mathbf{k} SVD features from \mathbf{A} . Given that SVD follows the well-known equation:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where \mathbf{U} is an $\mathbf{N} \times \mathbf{k}$ matrix and \mathbf{V}^T is an $\mathbf{k} \times \mathbf{p}$ matrix.

Is \mathbf{U} or \mathbf{V} more ideal to analyze the relationship between topics in the documents and each document?

V

Is \mathbf{U} or \mathbf{V} more ideal to analyze the relationship between topics in the documents and each term?

U