Preprocessing

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

↓ Stem nouns to singular

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

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

Document 1: ~~To~~ err ~~is~~ human, ~~but to~~ really foul ~~thing~~ up ~~you~~ need ~~a~~ computer
Document 2: Computer science ~~is~~ no more ~~about~~ computer ~~than~~ astronomy ~~is about~~ telescope
Document 3: ~~A~~ computer once beat ~~me at~~ chess, ~~but it was~~ no match ~~for me at~~ kick boxing

↓ Remove terms that occur once

Document 1: ~~To err is human~~, ~~but to really foul thing up you need a~~ computer
Document 2: Computer ~~science is~~ no ~~more about~~ computer ~~than astronomy is about telescope~~
Document 3: ~~A~~ computer ~~once beat me at chess~~, ~~but it was~~ no ~~match for me at kick boxing~~

Bag of Words: Document by Term Matrix

Document 1: ~~To~~ ~~err~~ ~~is~~ ~~human~~, ~~but to~~ ~~really foul thing~~ ~~up~~ ~~you~~ ~~need~~ ~~a~~ computer
Document 2: Computer ~~science~~ ~~is~~ no ~~more~~ ~~about~~ computer ~~than~~ ~~astronomy~~ ~~is about~~ ~~telescope~~
Document 3: ~~A~~ computer ~~once beat me at~~ chess, ~~but it was~~ no ~~match~~ ~~for me at kick boxing~~

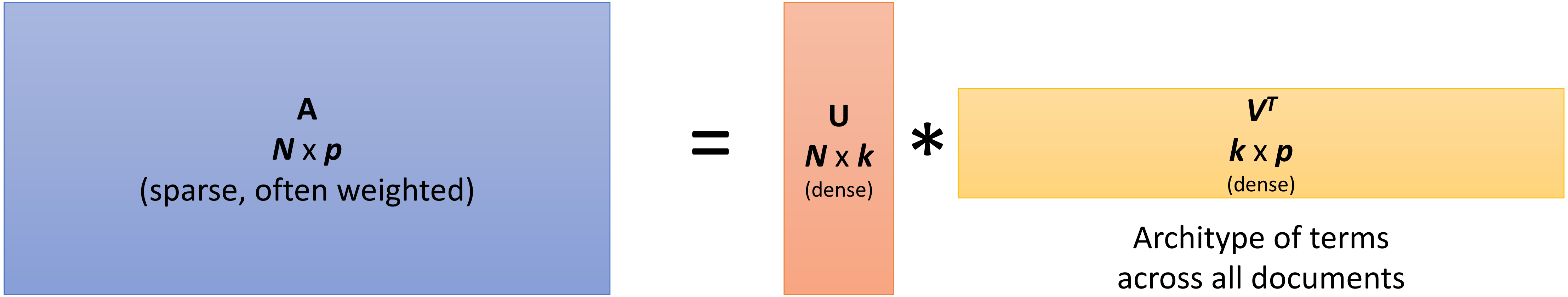
	computer	no
Document 1	1	0
Document 2	2	1
Document 3	1	1

Bag of Words: Matrix Factorization

Select ***k*** features to represent the corpus

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

Architypes of documents
across all terms



A
N x ***p***
(sparse, often weighted)

=

U
N x ***k***
(dense)

*

V^T
k x ***p***
(dense)

Architype of terms
across all documents

Term by Document Matrix
N terms
p documents

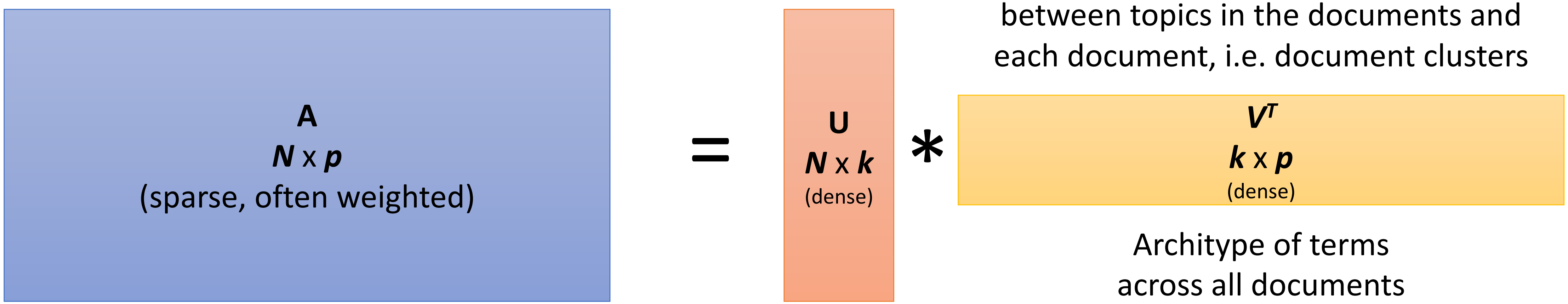
Bag of Words: Matrix Factorization

Select ***k*** features to represent the corpus

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

Architypes of documents across all terms

Best for analyzing the relationship between topics in the documents and each document, i.e. document clusters



A
N x ***p***
(sparse, often weighted)

=

U
N x ***k***
(dense)

*

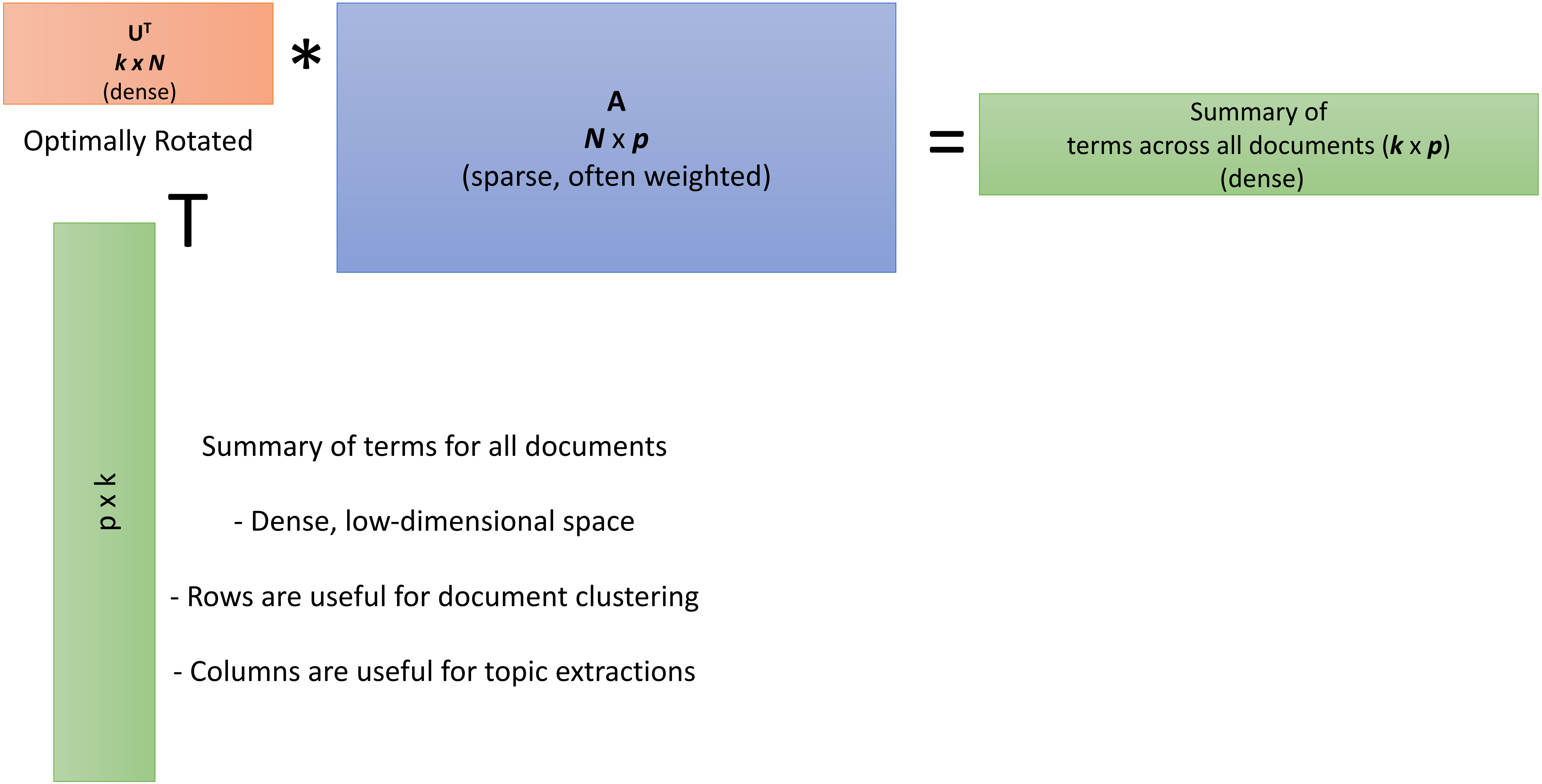
V^T
k x ***p***
(dense)

Architype of terms across all documents

Term by Document Matrix
N terms
p documents

Best for analyzing the relationship between topics in the documents and each term, i.e. topics composed of similar terms

Bag of Words: Matrix Factorization: document clustering and topic extraction
in SAS Text Miner



U^T
k* x *N
(dense)

Optimally Rotated

*

A
N x ***p***
(sparse, often weighted)

=

Summary of
terms across all documents (***k*** x ***p***)
(dense)

T

p x k

Summary of terms for all documents

- Dense, low-dimensional space

- Rows are useful for document clustering

- Columns are useful for topic extractions

Content-Sensitive: Matrix Factorization: Term Embedding: GloVe

	Term 1	Term 2	Term 3	Term 4	Term 5	…
Term1	90	2	0	1	0	…
Term 2	2	56	1	6	0	…
Term 3	0	1	78	0	1	…
Term 4	1	6	0	24	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Sparse, wide, fixed-length vectors that record term co-occurrence

Matrix Factorization

log bilinear
weighted least squares

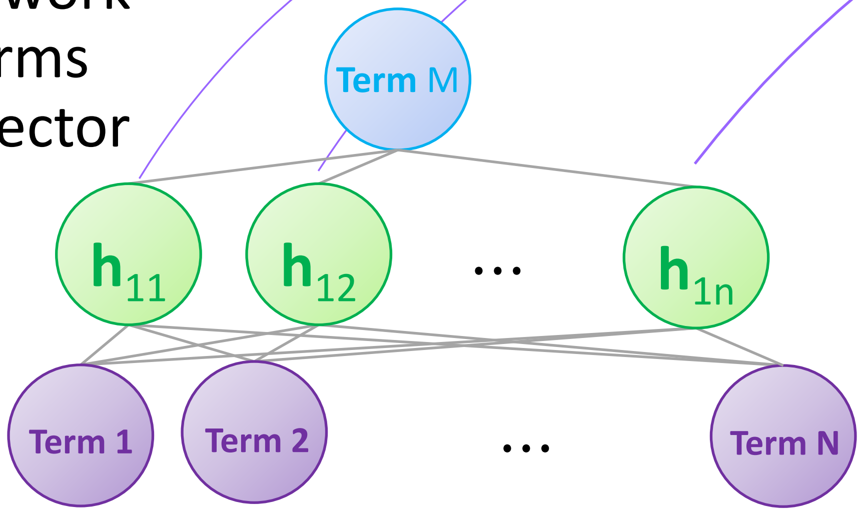
	Factor 1	Factor 2	Factor 3
Term 1	1.304	0.582	0.892
Term 2	0.897	0.843	0.885
Term 3	0.745	1.129	1.002
Term 4	0.921	0.962	0.714
⋮	⋮	⋮	⋮

Each row vector represents a term ("distributed representation")

Dense, fixed-length vectors for each term in the corpus

Content-Sensitive: Neural Networks: Term Embedding: Like Word2Vec

The output of a hidden layer of a neural network is used to embed terms into a fixed-length vector space from a simple encoding



	Term 1	Term 2	Term 3	Term 4	Term 5	…
Document 1	0	0	0	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

	Factor 1	Factor 2	…	Factor N
Term 1	1.304	0.582	…	0.892
Term 2	0.897	0.843	…	0.885
Term 3	0.745	1.129	…	1.002
Term 4	0.921	0.962	…	0.714
⋮	⋮	⋮	…	⋮

Each row vector represents a term ("distributed representation")

Dense, fixed-length vectors for each term in the corpus)