Introduction to NLP

Video 6: Methods of Text Preprocessing - Part 3

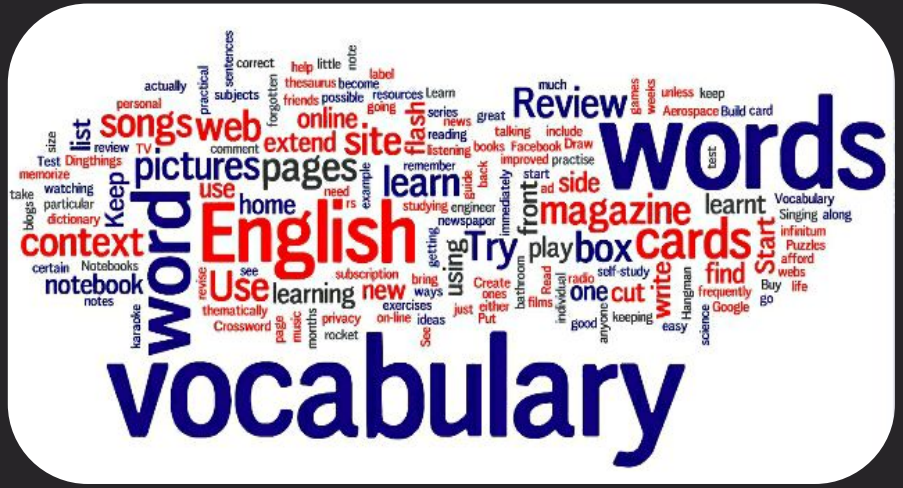
Analytics Vidhya

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters
- ✓ Stop Words Removal
- ✓ Stemming and Lemmatization
- Vocabulary Building
- Vectorization

Vocabulary Building



Select and define the set of unique words NLP model will analyse.

Tokens left after preprocessing steps : Base for the model

Analytics Vidhya

Jupyter Notebook

Analytics Vidhya

Vectorization

Vectorization

```
of': 0.0585545379819860 46
'member': 0.0539465764765739 44
'popular': 0.039600491970539 09
'is': 0.0388509404656887 05
'an': 0.036895200103282 93
'one': 0.0215294470487833 02
'love': 0.0121942529672384 26
'helps': -0.0059114586110119 82
'important': -0.0118997250276207 92
'spiderman': -0.035659030556440 35
```

Convert words into numerical representations called vectors.

Vectorization

- One Hot Encoding
- Count Vectorizer
- TF-IDF
- Word Embedding

Vectorization

- One Hot Encoding
- Count Vectorizer
- TF-IDF
- Word Embedding

Vectorization

- One Hot Encoding
- Count Vectorizer
- TF-IDF
- Word Embedding

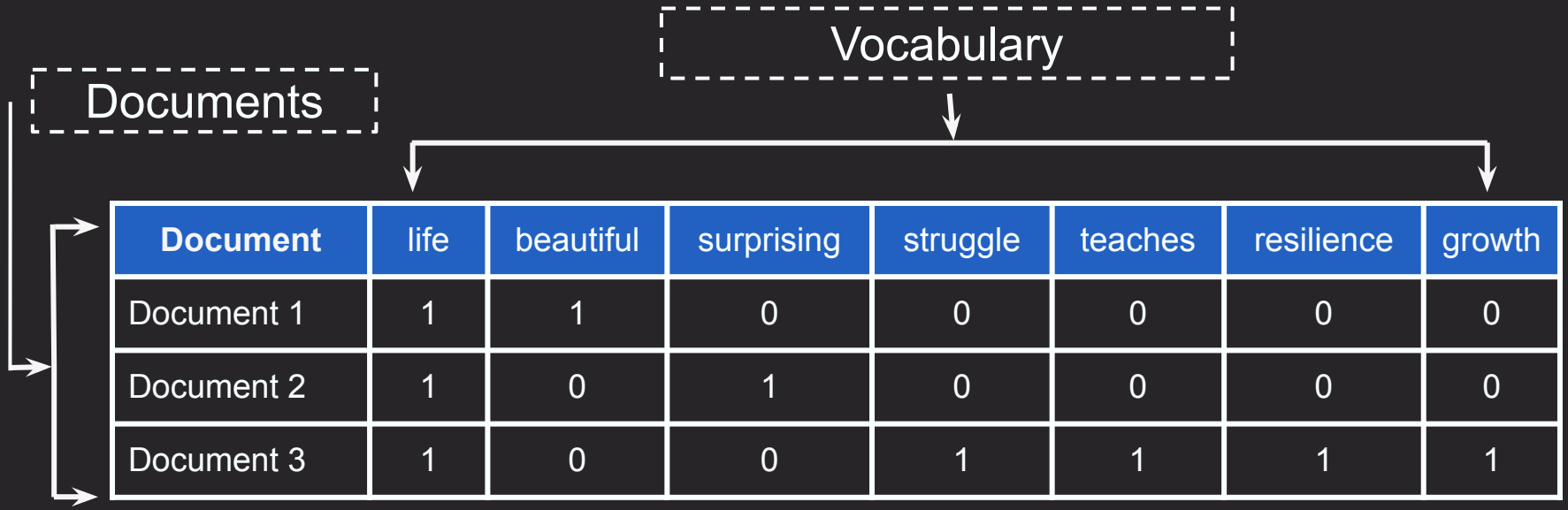
One-Hot Encoding

Binary vector to show whether a word appears in a document.

1 **Word is present**

0 **Word is absent**

One Hot Encoding



Document	life	beautiful	surprising	struggle	teaches	resilience	growth
Document 1	1	1	0	0	0	0	0
Document 2	1	0	1	0	0	0	0
Document 3	1	0	0	1	1	1	1

One Hot Encoding

Document 1 : “Life is beautiful.”

Document 2 : “ Life is surprising.”

Document 3 : “Struggle in life teaches resilience and growth.”

One Hot Encoding

Document 1 : “**Life** is **beautiful**.”

Document 2 : “ Life is surprising.”

Document 3 : “Struggle in life teaches resilience and growth.”

life | **beautiful**

One Hot Encoding

Document 1 : “**Life** is **beautiful**.”

Document 2 : “ Life is **surprising**.”

Document 3 : “Struggle in life teaches resilience and growth.”

life	beautiful	surprising

Analytics Vidhya

One Hot Encoding

Document 1 : “**Life** is **beautiful**.”

Document 2 : “ Life is **surprising**.”

Document 3 : “**Struggle** in life **teaches resilience** and **growth**.”

life	beautiful	surprising	struggle	teaches	resilience	growth

Analytics Vidhya

One Hot Encoding

Document	life	beautiful	surprising	struggle	teaches	resilience	growth
Document 1	1	1	0	0	0	0	0
Document 2	1	0	1	0	0	0	0
Document 3	1	0	0	1	1	1	1

life | beautiful | surprising | struggle | teaches | resilience | growth

Analytics Vidhya

Jupyter Notebook

Vectorization

One Hot Encoding

Count Vectorization

TF-IDF

Word Embedding

Analytics Vidhya

Count Vectorization

Represented as matrix of token counts.

Document	life	beautiful	surprising	struggle	turn	teaches	resilience	growth
Document 1	3	2	0	0	1	0	0	0
Document 2	2	0	1	0	0	0	0	0
Document 3	3	0	0	1	1	1	1	1

Analytics Vidhya

Count Vectorization

Document 1 : “Life is life, life is beautiful and full of beautiful turns.”

Document 2 : “ Life surprises us at every turn of life”

Document 3 : “Struggle in life teaches resilience, life, and growth in life.”

Analytics Vidhya

Count Vectorization

Document 1 : “**Life** is **life**, **life** is **beautiful** and full **beautiful turns**.”

Document 2 : “ **Life surprises** us at every **turn** of **life**”

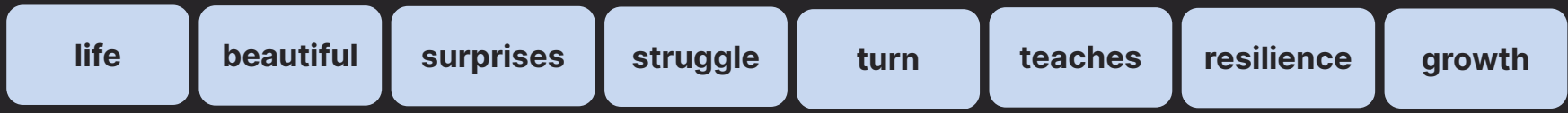
Document 3 : “**Struggle** in **life teaches resilience**, **life**, and **growth** in **life**.”

life	beautiful	surprises	struggle	turn	teaches	resilience	growth

Analytics Vidhya

Count Vectorization

Document	life	beautiful	surprising	struggle	turn	teaches	resilience	growth
Document 1	3	2	0	0	1	0	0	0
Document 2	2	0	1	0	0	0	0	0
Document 3	3	0	0	1	1	1	1	1



Analytics Vidhya

Jupyter Notebook