AASD 4004
Machine Learning - II

Applied AI Solutions Developer Program

Module 05
Topic Modeling

Vejey Gandyer

Agenda

Topic Modeling

LDA

Building a Topic Model

Topic Modeling

What is it?

Topic Modeling

Refers to the process of making sense of a collection of documents (corpus)

Splits / Groups a collection of documents into topics

Latent Dirichlet Allocation (LDA)

What is it?

LDA

Groups a collection of documents into "topics"

Input

D1: I like to eat broccoli and bananas.

D2: I ate a banana and salad for breakfast.

D3: Puppies and kittens are cute.

D4: My sister adopted a kitten yesterday.

D5: Look at this cute hamster munching on a piece of broccoli.

Output

Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching

Topic B: 20% puppies, 20% kittens, 20% cute, 15% hamster

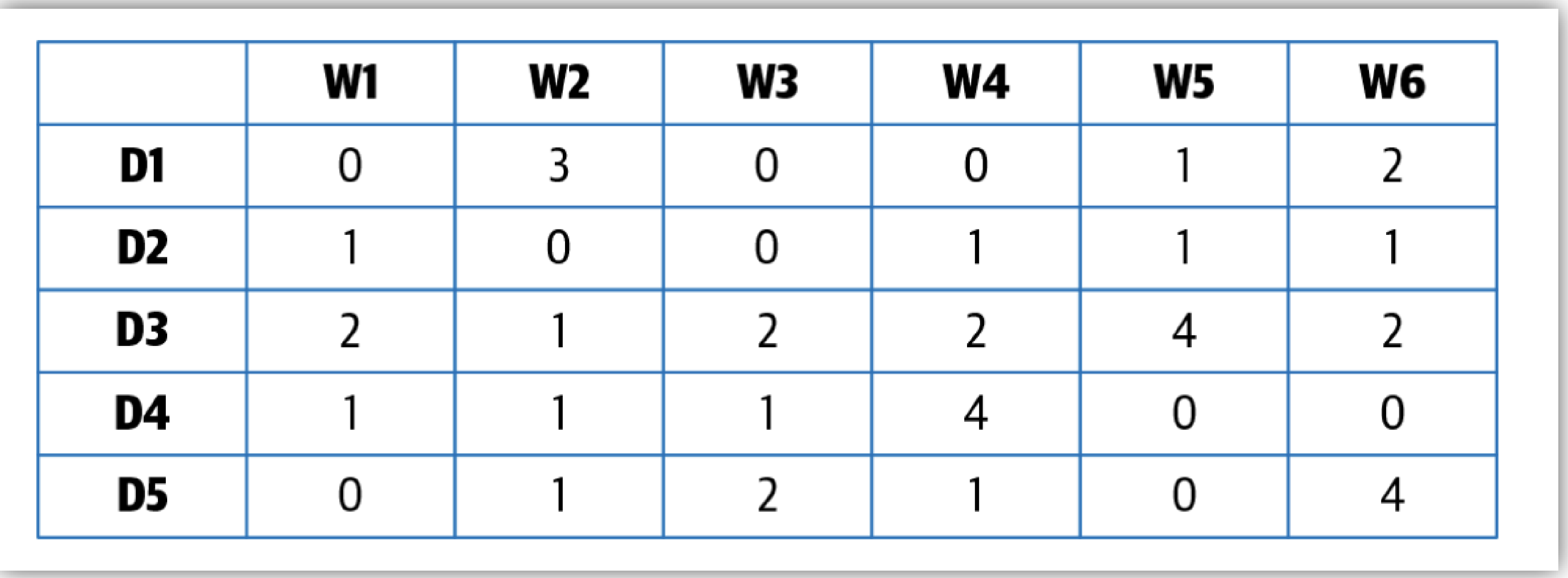
Document 1 and 2: 100% Topic A

Document 3 and 4: 100% Topic B

Document 5: 60% Topic A, 40% Topic B

LDA

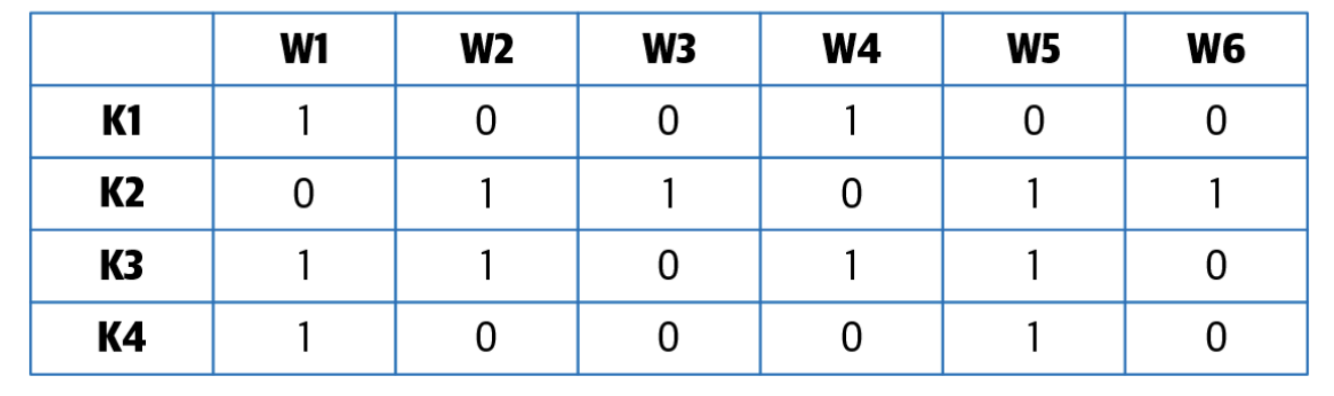
Document-Term Matrix (M)



	W1	W2	W3	W4	W5	W6
D1	0	3	0	0	1	2
D2	1	0	0	1	1	1
D3	2	1	2	2	4	2
D4	1	1	1	4	0	0
D5	0	1	2	1	0	4

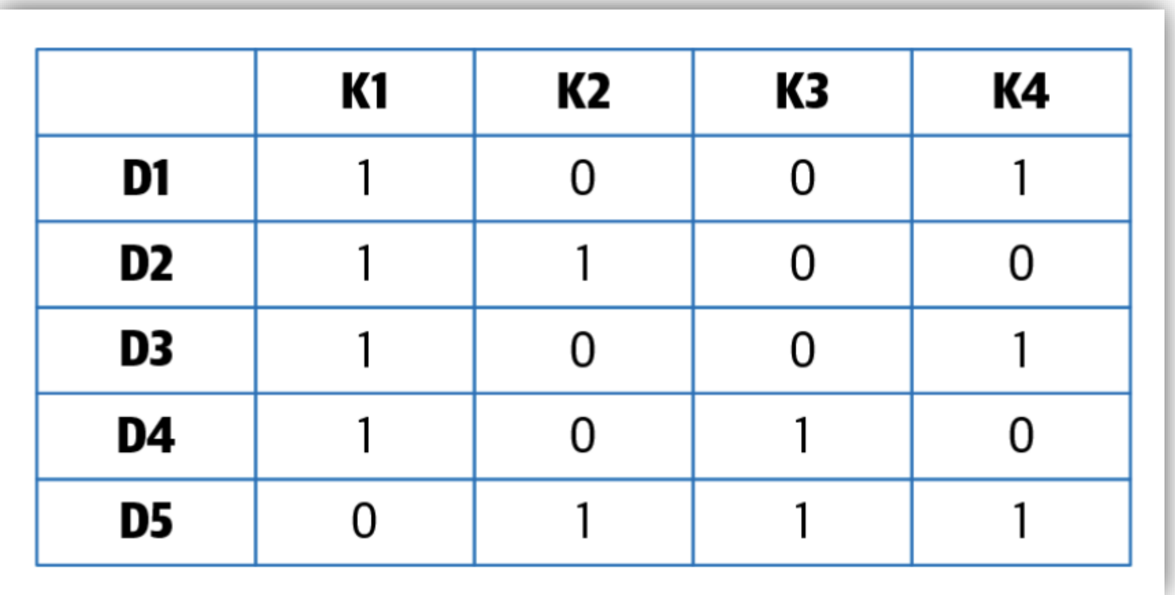
LDA - Factorization

Topics -Term Matrix (M2) K x N



	W1	W2	W3	W4	W5	W6
K1	1	0	0	1	0	0
K2	0	1	1	0	1	1
K3	1	1	0	1	1	0
K4	1	0	0	0	1	0

Document-Topics Matrix (M1) M x K



	K1	K2	K3	K4
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
D4	1	0	1	0
D5	0	1	1	1

LDA - Steps

Step 1: Import libraries

Step 2: Pre-process (Tokenize, remove stop-words, lowercase, …)

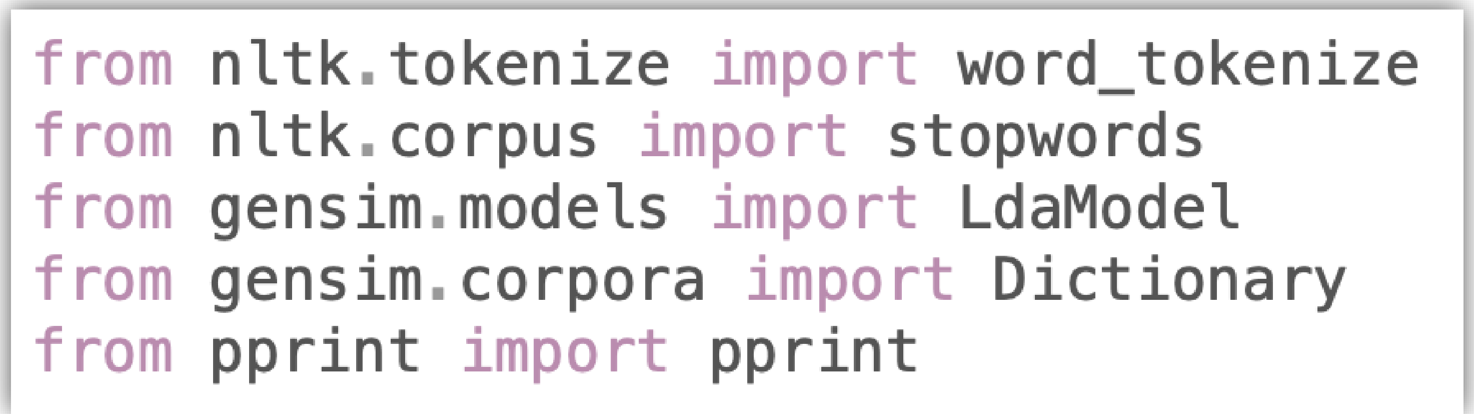
Step 3: Create a dictionary for the document

Step 4: Filter low frequency words

Step 5: Create a index to word dictionary

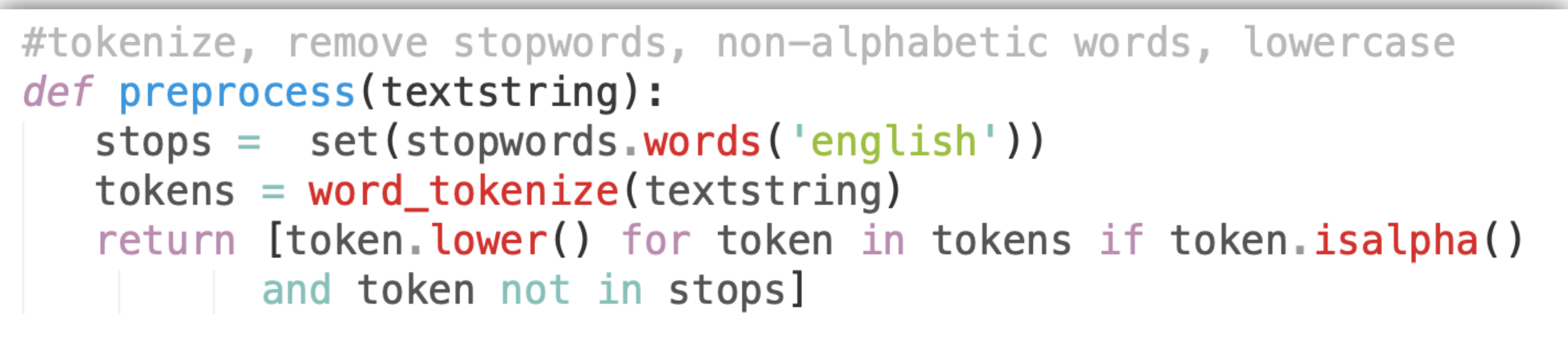
Step 6: Train the Topic Model

LDA - Importing libraries

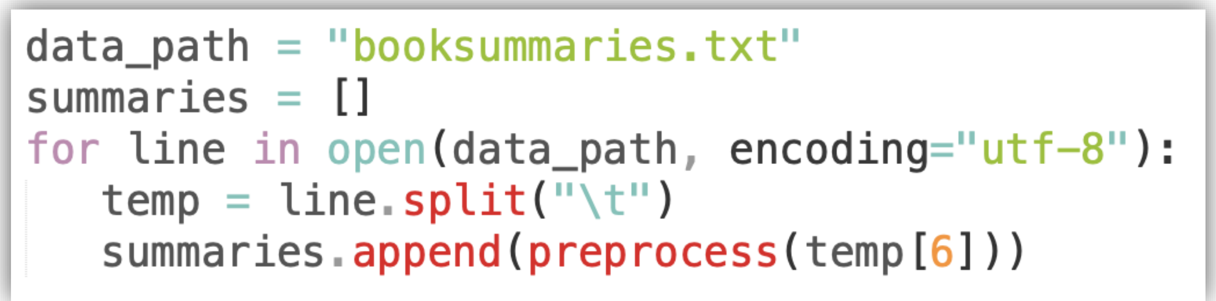


```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from gensim.models import LdaModel
from gensim.corpora import Dictionary
from pprint import pprint
```

LDA - Pre-process

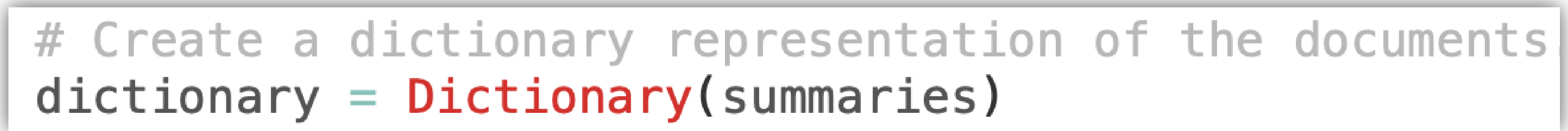


```python
#tokenize, remove stopwords, non-alphabetic words, lowercase
def preprocess(textstring):
   stops =  set(stopwords.words('english'))
   tokens = word_tokenize(textstring)
   return [token.lower() for token in tokens if token.isalpha()
             and token not in stops]
```



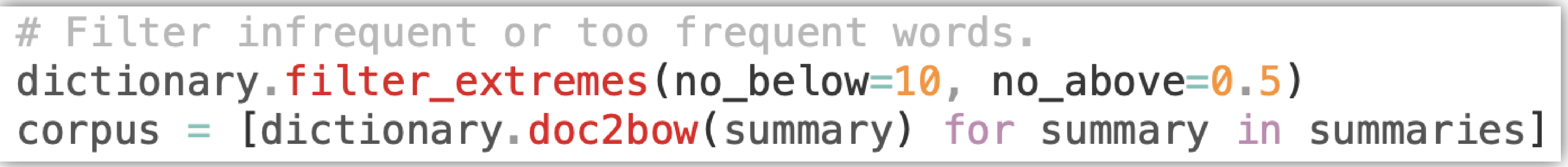
```python
data_path = "booksummaries.txt"
summaries = []
for line in open(data_path, encoding="utf-8"):
   temp = line.split("\t")
   summaries.append(preprocess(temp[6]))
```

LDA - Creating a dictionary



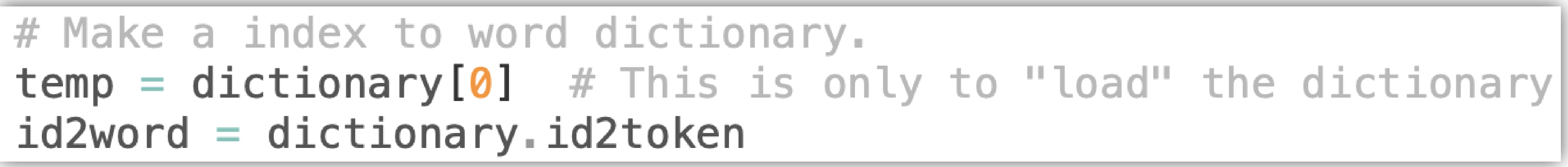
```
# Create a dictionary representation of the documents
dictionary = Dictionary(summaries)
```

LDA - Filter low-frequency words



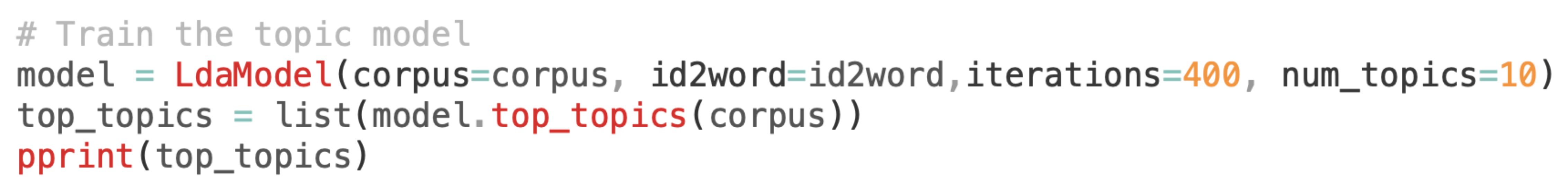
```
# Filter infrequent or too frequent words.
dictionary.filter_extremes(no_below=10, no_above=0.5)
corpus = [dictionary.doc2bow(summary) for summary in summaries]
```

LDA - Index to word dictionary



```
# Make a index to word dictionary.
temp = dictionary[0]  # This is only to "load" the dictionary
id2word = dictionary.id2token
```

LDA - Train the Topic Model



```
# Train the topic model
model = LdaModel(corpus=corpus, id2word=id2word,iterations=400, num_topics=10)
top_topics = list(model.top_topics(corpus))
pprint(top_topics)
```

Exercise: Building a Topic Model for the given dataset

https://raw.githubusercontent.com/subashgandyer/datasets/main/kaggledatasets.csv

Building a Topic Model

Create a Topic Model for a given dataset

- LDA

Further Reading

LDA

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation