Information Retrieval – Assignment 18

# Code

# IR21A.py CS5154/6054 cheng 2022  
# decompose the document-term matrix  
# Usage: python IR21A.py  
  
import numpy as np  
from sklearn.feature\_extraction.text import TfidfVectorizer  
from sklearn.decomposition import TruncatedSVD, LatentDirichletAllocation  
from matplotlib import pyplot as plt  
  
f = open("bible.txt", "r")  
docs = f.readlines()  
f.close()  
N = len(docs)  
  
vectorizer = TfidfVectorizer(max\_df=1000, min\_df=100)  
X = vectorizer.fit\_transform(docs)  
words = vectorizer.get\_feature\_names()  
  
truncated\_svd\_model = TruncatedSVD(n\_components=5)  
W = truncated\_svd\_model.fit\_transform(X).T  
H = truncated\_svd\_model.components\_  
  
from wordcloud import WordCloud  
  
for topic in range(5):  
 size = {}  
 largest = H[topic].argsort()[::-1]   
 for i in range(40):  
 size[words[largest[i]]] = abs(H[topic][largest[i]])  
 wc = WordCloud(background\_color="white", max\_words=100, width=960, height=540)  
 wc.generate\_from\_frequencies(size)  
 plt.imshow(wc, interpolation='bilinear')  
 plt.axis("off")  
 plt.show()  
  
lda\_model = LatentDirichletAllocation(n\_components=5)  
W = lda\_model.fit\_transform(X).T  
H = lda\_model.components\_  
  
for topic in range(5):  
 size = {}  
 largest = H[topic].argsort()[::-1]  
 for i in range(40):  
 size[words[largest[i]]] = abs(H[topic][largest[i]])  
 wc = WordCloud(background\_color="white", max\_words=100, width=960, height=540)  
 wc.generate\_from\_frequencies(size)  
 plt.imshow(wc, interpolation='bilinear')  
 plt.axis("off")  
 plt.show()

# Results

## Truncated SVD

### Iteration 1

Text

Description automatically generated

### Iteration 2

Text

Description automatically generated

### Iteration 3

Text

Description automatically generated

### Iteration 4

Text

Description automatically generated

### Iteration 5

Text

Description automatically generated\

## Latent Dirichlet Allocation

### Iteration 1

Text

Description automatically generated

### Iteration 2

Text

Description automatically generated

### Iteration 3

Text

Description automatically generated

### Iteration 4

A picture containing text

Description automatically generated

### Iteration 5

Text

Description automatically generated with medium confidence