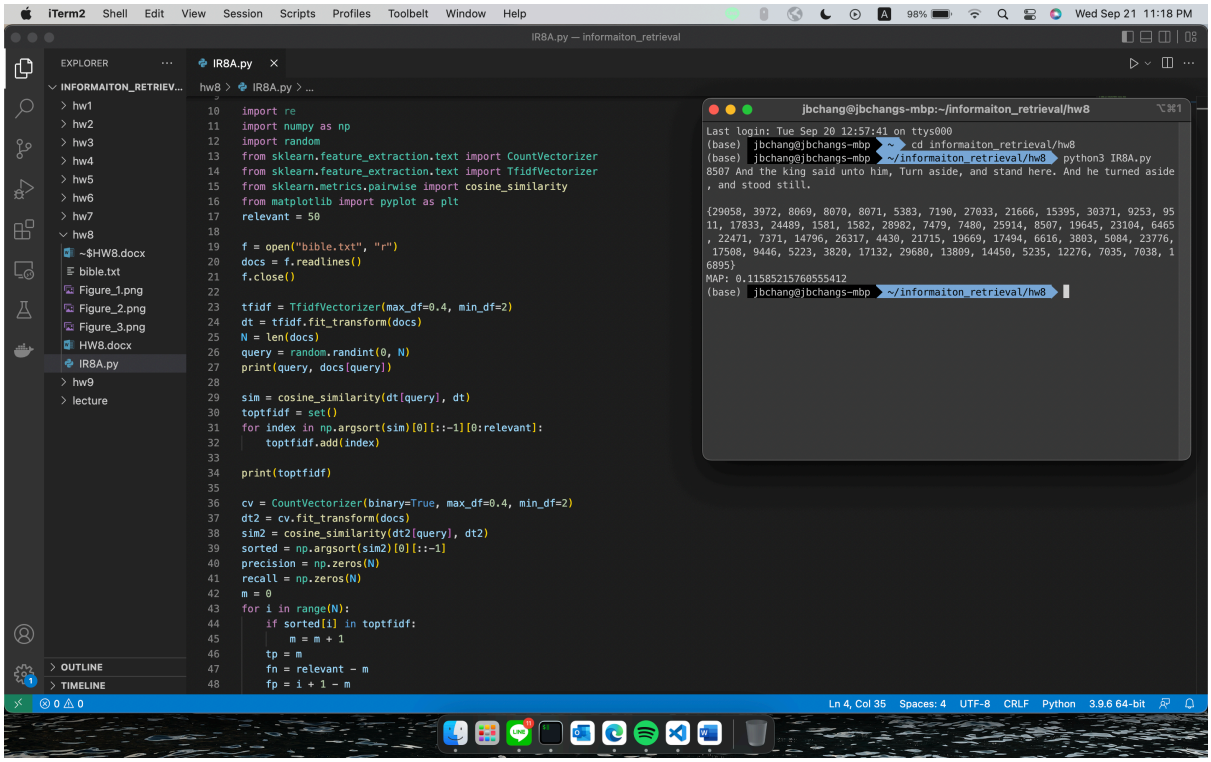
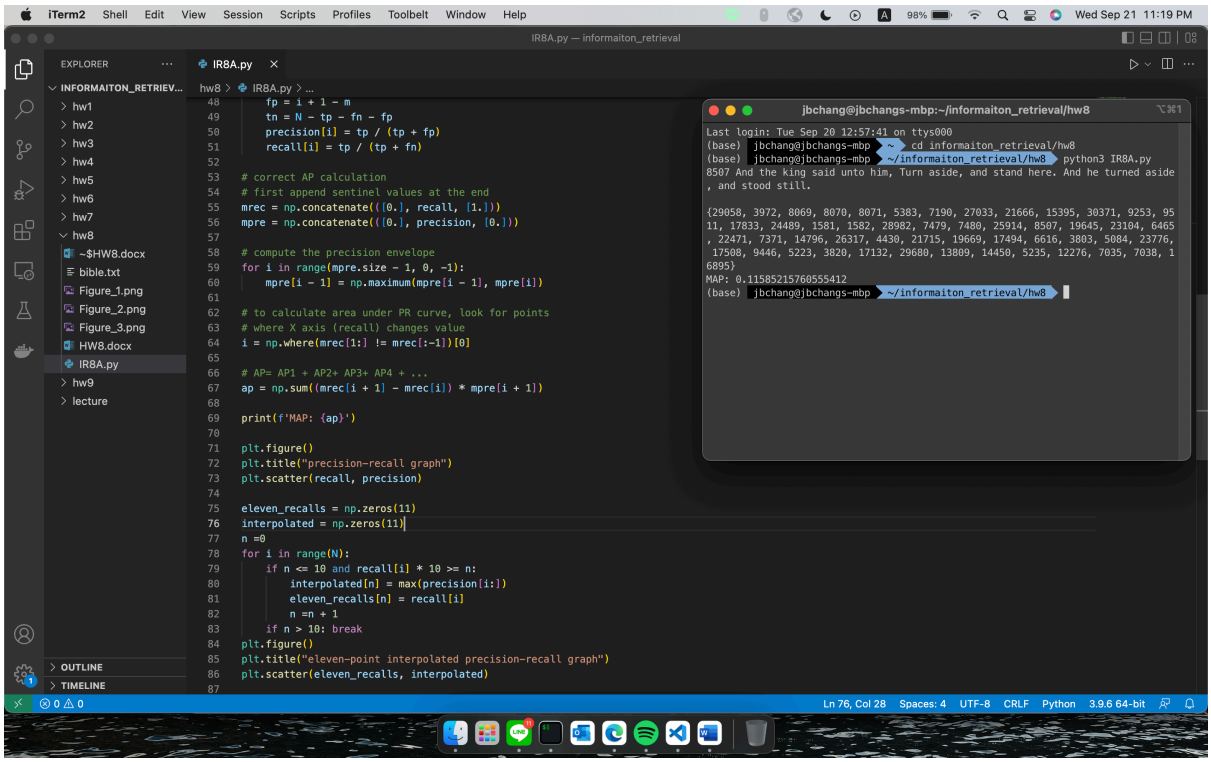
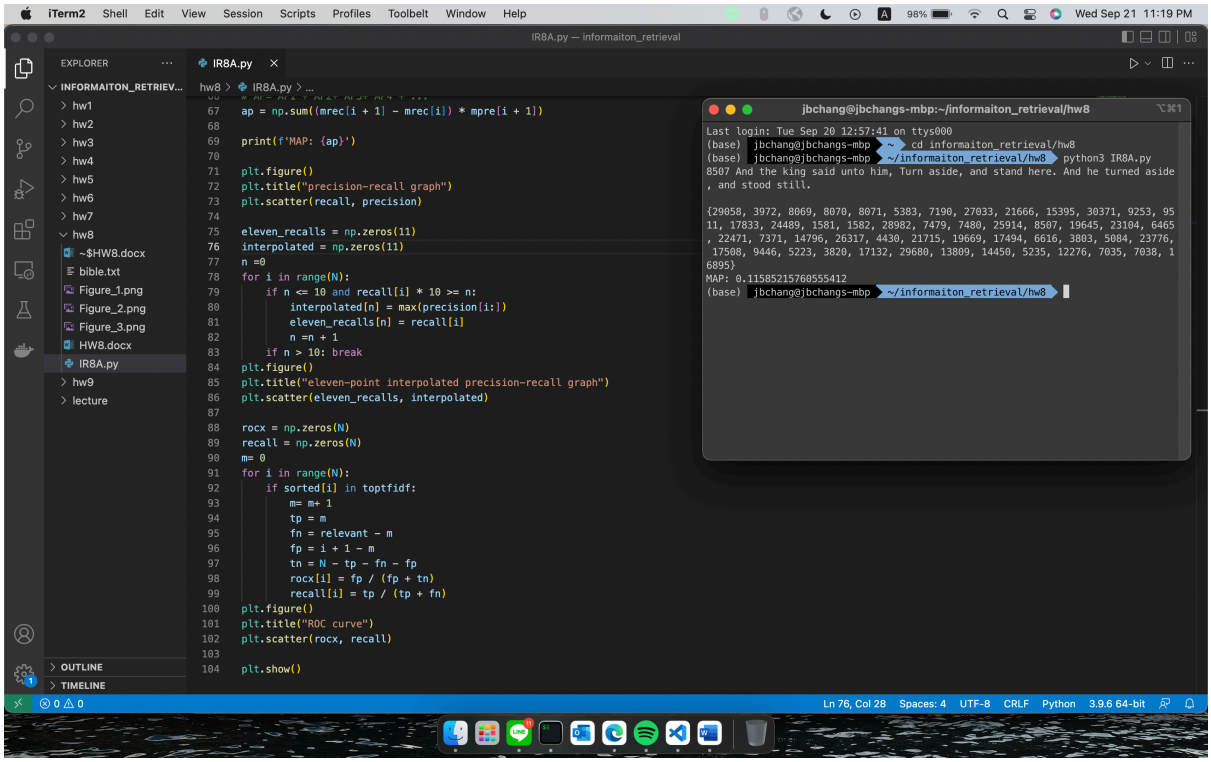
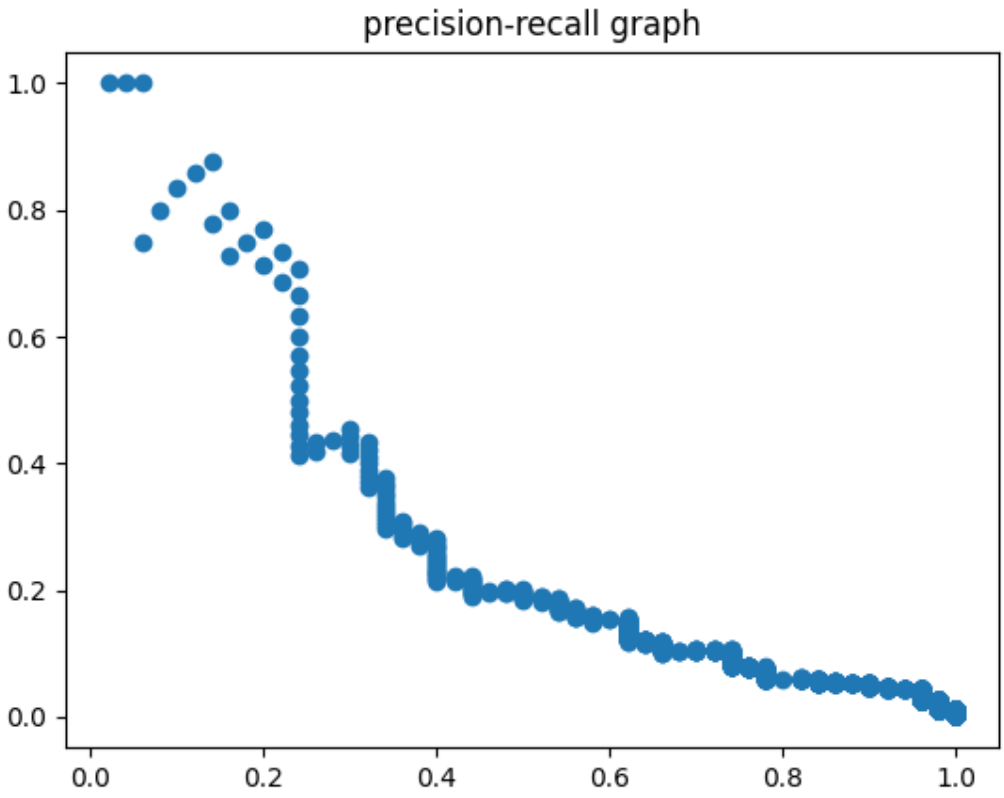
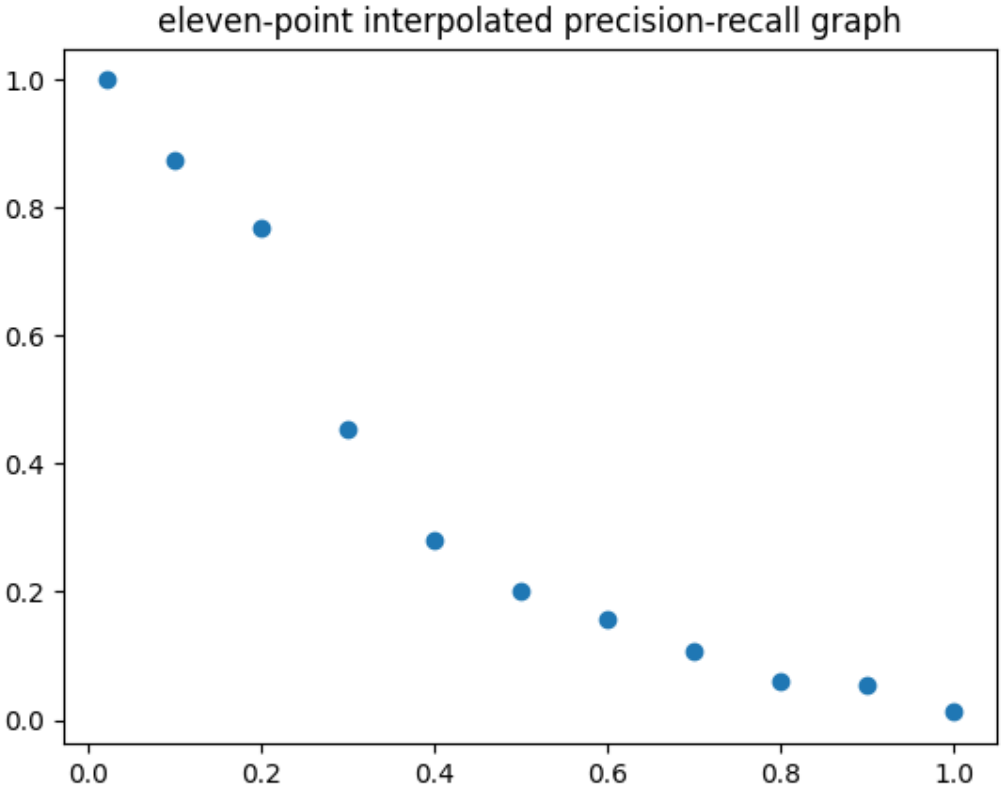
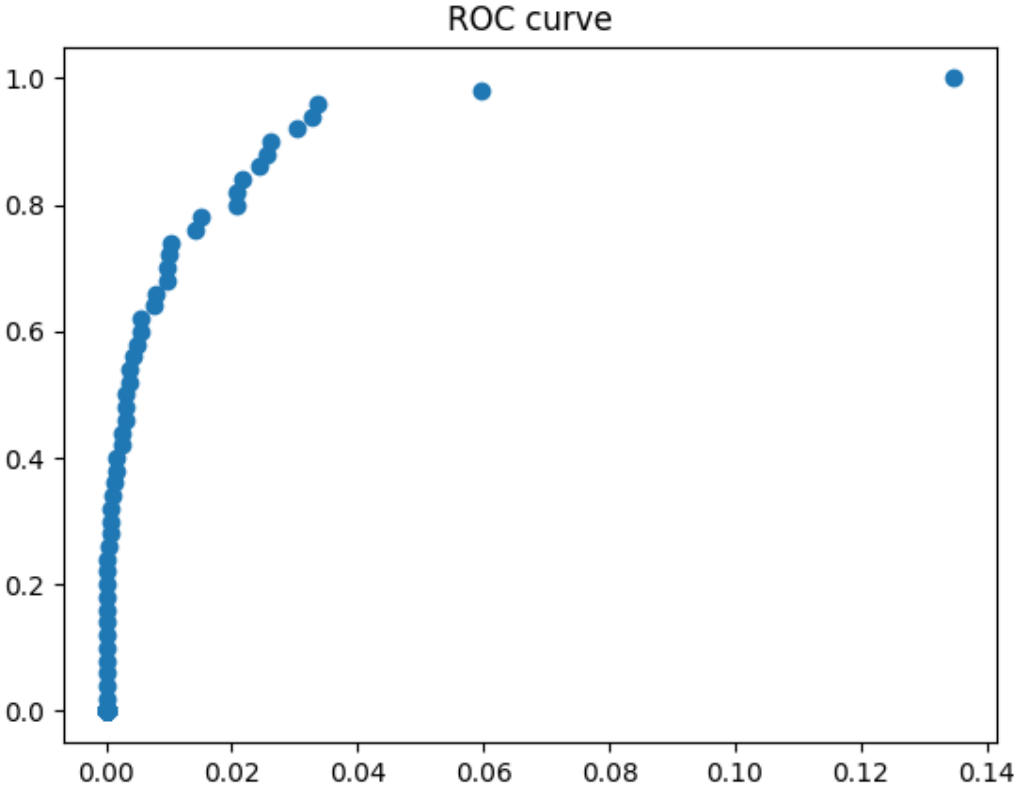

iTerm2 Shell Edit View Session Scripts Profiles Toolbelt Window Help
Wed Sep 21 11:18 PM
IR8A.py — informaiton_retrieval
EXPLORER
IR8A.py
INFORMAITON_RETRIEV...
hw1
hw2
hw3
hw4
hw5
hw6
hw7
hw8
~$HW8.docx
bible.txt
Figure_1.png
Figure_2.png
Figure_3.png
HW8.docx
IR8A.py
hw9
lecture
OUTLINE
TIMELINE
hw8 > IR8A.py > ...
10 import re
11 import numpy as np
12 import random
13 from sklearn.feature_extraction.text import CountVectorizer
14 from sklearn.feature_extraction.text import TfidfVectorizer
15 from sklearn.metrics.pairwise import cosine_similarity
16 from matplotlib import pyplot as plt
17 relevant = 50
18
19 f = open("bible.txt", "r")
20 docs = f.readlines()
21 f.close()
22
23 tfidf = TfidfVectorizer(max_df=0.4, min_df=2)
24 dt = tfidf.fit_transform(docs)
25 N = len(docs)
26 query = random.randint(0, N)
27 print(query, docs[query])
28
29 sim = cosine_similarity(dt[query], dt)
30 toptfidf = set()
31 for index in np.argsort(sim)[0][::-1][0:relevant]:
32 toptfidf.add(index)
33
34 print(toptfidf)
35
36 cv = CountVectorizer(binary=True, max_df=0.4, min_df=2)
37 dt2 = cv.fit_transform(docs)
38 sim2 = cosine_similarity(dt2[query], dt2)
39 sorted = np.argsort(sim2)[0][::-1]
40 precision = np.zeros(N)
41 recall = np.zeros(N)
42 m = 0
43 for i in range(N):
44 if sorted[i] in toptfidf:
45 m = m + 1
46 tp = m
47 fn = relevant - m
48 fp = i + 1 - m
jbchang@jbchangs-mbp:~/informaiton_retrieval/hw8
Last login: Tue Sep 20 12:57:41 on ttys000
(base) jbchang@jbchangs-mbp ~ cd informaiton_retrieval/hw8
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8 python3 IR8A.py
8507 And the king said unto him, Turn aside, and stand here. And he turned aside , and stood still.
{29058, 3972, 8069, 8070, 8071, 5383, 7190, 27033, 21666, 15395, 30371, 9253, 95 11, 17833, 24489, 1581, 1582, 28982, 7479, 7480, 25914, 8507, 19645, 23104, 6465 , 22471, 7371, 14796, 26317, 4430, 21715, 19669, 17494, 6616, 3803, 5084, 23776, 17508, 9446, 5223, 3820, 17132, 29680, 13809, 14450, 5235, 12276, 7035, 7038, 1 6895}
MAP: 0.11585215760555412
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8
Ln 4, Col 35 Spaces: 4 UTF-8 CRLF Python 3.9.6 64-bit


iTerm2 Shell Edit View Session Scripts Profiles Toolbelt Window Help
Wed Sep 21 11:19 PM
IR8A.py — informaiton_retrieval
EXPLORER
IR8A.py
INFORMAITON_RETRIEV...
hw1
hw2
hw3
hw4
hw5
hw6
hw7
hw8
~$HW8.docx
bible.txt
Figure_1.png
Figure_2.png
Figure_3.png
HW8.docx
IR8A.py
hw9
lecture
OUTLINE
TIMELINE
hw8 > IR8A.py > ...
48 fp = i + 1 - m
49 tn = N - tp - fn - fp
50 precision[i] = tp / (tp + fp)
51 recall[i] = tp / (tp + fn)
52
53 # correct AP calculation
54 # first append sentinel values at the end
55 mrec = np.concatenate(([0.], recall, [1.]))
56 mpre = np.concatenate(([0.], precision, [0.]))
57
58 # compute the precision envelope
59 for i in range(mpre.size - 1, 0, -1):
60 mpre[i - 1] = np.maximum(mpre[i - 1], mpre[i])
61
62 # to calculate area under PR curve, look for points
63 # where X axis (recall) changes value
64 i = np.where(mrec[1:] != mrec[:-1])[0]
65
66 # AP= AP1 + AP2+ AP3+ AP4 + ...
67 ap = np.sum((mrec[i + 1] - mrec[i]) * mpre[i + 1])
68
69 print(f'MAP: {ap}')
70
71 plt.figure()
72 plt.title("precision-recall graph")
73 plt.scatter(recall, precision)
74
75 eleven_recalls = np.zeros(11)
76 interpolated = np.zeros(11)
77 n =0
78 for i in range(N):
79 if n <= 10 and recall[i] * 10 >= n:
80 interpolated[n] = max(precision[i:])
81 eleven_recalls[n] = recall[i]
82 n =n + 1
83 if n > 10: break
84 plt.figure()
85 plt.title("eleven-point interpolated precision-recall graph")
86 plt.scatter(eleven_recalls, interpolated)
87
jbchang@jbchangs-mbp:~/informaiton_retrieval/hw8
Last login: Tue Sep 20 12:57:41 on ttys000
(base) jbchang@jbchangs-mbp ~ cd informaiton_retrieval/hw8
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8 python3 IR8A.py
8507 And the king said unto him, Turn aside, and stand here. And he turned aside , and stood still.
{29058, 3972, 8069, 8070, 8071, 5383, 7190, 27033, 21666, 15395, 30371, 9253, 95 11, 17833, 24489, 1581, 1582, 28982, 7479, 7480, 25914, 8507, 19645, 23104, 6465 , 22471, 7371, 14796, 26317, 4430, 21715, 19669, 17494, 6616, 3803, 5084, 23776, 17508, 9446, 5223, 3820, 17132, 29680, 13809, 14450, 5235, 12276, 7035, 7038, 1 6895}
MAP: 0.11585215760555412
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8
Ln 76, Col 28 Spaces: 4 UTF-8 CRLF Python 3.9.6 64-bit


iTerm2 Shell Edit View Session Scripts Profiles Toolbelt Window Help
Wed Sep 21 11:19 PM
IR8A.py — informaiton_retrieval
EXPLORER
INFORMAITON_RETRIEV...
hw1
hw2
hw3
hw4
hw5
hw6
hw7
hw8
~$HW8.docx
bible.txt
Figure_1.png
Figure_2.png
Figure_3.png
HW8.docx
IR8A.py
hw9
lecture
OUTLINE
TIMELINE
IR8A.py
hw8 > IR8A.py > ...
ap = np.sum((mrec[i + 1] - mrec[i]) * mpre[i + 1])
print(f'MAP: {ap}')
plt.figure()
plt.title("precision-recall graph")
plt.scatter(recall, precision)
eleven_recalls = np.zeros(11)
interpolated = np.zeros(11)
n =0
for i in range(N):
if n <= 10 and recall[i] * 10 >= n:
interpolated[n] = max(precision[i:])
eleven_recalls[n] = recall[i]
n =n + 1
if n > 10: break
plt.figure()
plt.title("eleven-point interpolated precision-recall graph")
plt.scatter(eleven_recalls, interpolated)
rocx = np.zeros(N)
recall = np.zeros(N)
m= 0
for i in range(N):
if sorted[i] in toptfidf:
m= m+ 1
tp = m
fn = relevant - m
fp = i + 1 - m
tn = N - tp - fn - fp
rocx[i] = fp / (fp + tn)
recall[i] = tp / (tp + fn)
plt.figure()
plt.title("ROC curve")
plt.scatter(rocx, recall)
plt.show()
jbchang@jbchangs-mbp:~/informaiton_retrieval/hw8
Last login: Tue Sep 20 12:57:41 on ttys000
(base) jbchang@jbchangs-mbp ~ cd informaiton_retrieval/hw8
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8 python3 IR8A.py
8507 And the king said unto him, Turn aside, and stand here. And he turned aside
, and stood still.
{29058, 3972, 8069, 8070, 8071, 5383, 7190, 27033, 21666, 15395, 30371, 9253, 95
11, 17833, 24489, 1581, 1582, 28982, 7479, 7480, 25914, 8507, 19645, 23104, 6465
, 22471, 7371, 14796, 26317, 4430, 21715, 19669, 17494, 6616, 3803, 5084, 23776,
17508, 9446, 5223, 3820, 17132, 29680, 13809, 14450, 5235, 12276, 7035, 7038, 1
6895}
MAP: 0.11585215760555412
(base) jbchang@jbchangs-mbp ~/informaiton_retrieval/hw8
Ln 76, Col 28 Spaces: 4 UTF-8 CRLF Python 3.9.6 64-bit


precision-recall graph
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


eleven-point interpolated precision-recall graph
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14