Week 5 Homework

This is the function modified to my personal refactoring.

```
from typing import List
from collections import defaultdict
import re
import math
import numpy as np

# used for unseen words in training vocabularies
UNK = None
# sentence start and end
SENTENCE_START = "<s>"
SENTENCE_END = "</s>"

def read_sentences_from_file(file_path: str) -> List[List[str]]:
    '''
        read the files.
    '''
    with open(file_path, "r") as f:
        return [re.split("\s+", line.rstrip('\n')) for line in f]

class UnigramLanguageModel:
    def __init__(self, sentences, global_vocabs, mode="collection",
smoothing=False):

        '''
            sentences: sentences of the dataset
            mode: whether this language model is for the whole
corpus/collection or just a single document
            smoothing: add-one smoothing
        '''

        self.mode = mode
        self.word_freq = defaultdict(int)
        self.smoothing = smoothing
        self.global_vocabs = global_vocabs

        for sentence in sentences:
            for word in sentence:
                if word != SENTENCE_START and word != SENTENCE_END:
                    self.word_freq[word] += 1

        if self.mode == "collection":
            self.word_freq[UNK] = 1

        self.total = sum(self.word_freq.values())
    def calculate_unigram_probability(self, word):
        '''
            calculate unigram probability of a word
        '''
        if not self.smoothing:
            return self.word_freq[word] / self.total
        else:
            if self.mode == "document":
                # print("NUMERATOR", word, self.mode, (self.vocab[word] + 1))
                return (self.word_freq[word] + 1) / (self.total + len(self.global_vocabs))
            if word in self.word_freq:
                count = self.word_freq[word]
            else:
                count = self.word_freq[UNK]
                # print(count)
            # print("NUMERATOR", word, self.mode, (count + 1))
            return ((count + 1) / (self.total + len(self.global_vocabs)))



    def calculate_sentence_probability(self, sentence, normalize_probability=True):
        '''
            calculate score/probability of a sentence or query using the unigram language model.
            sentence: input sentence or query
            normalize_probability: If true then log of probability is not computed. Otherwise take log2 of the probability score.
        '''
        prob = 1
        for word in sentence:
            prob *= self.calculate_unigram_probability(word)

        return prob if normalize_probability else math.log(prob, 2)



def calculate_interpolated_sentence_probability(sentence, doc, collection, alpha=0.75, normalize_probability=True):
    '''
        calculate interpolated sentence/query probability using both sentence and collection unigram models.
        sentence: input sentence/query
        doc: unigram language model a doc. HINT: this can be an instance of the UnigramLanguageModel class
        collection: unigram language model a collection. HINT: this can be an instance of the UnigramLanguageModel class
        alpha: the hyperparameter to combine the two probability scores coming from the document and collection language models.
        normalize_probability: If true then log of probability is not computed. Otherwise take log2 of the probability score.
    '''
    prob = 1
    for word in sentence:
        if word == SENTENCE_START or word == SENTENCE_END:
            continue
        prob *= (alpha * doc.calculate_unigram_probability(word)) + ((1-alpha) * collection.calculate_unigram_probability(word))
    return prob if normalize_probability else math.log(prob, 2)

if __name__ == '__main__':
    global_vocabs = read_sentences_from_file('./train_vocab.txt')
    global_vocabs.append([UNK])
    # print(global_vocabs)



    actual_dataset = read_sentences_from_file("./train.txt")
    doc1_dataset = read_sentences_from_file("./doc1.txt")
    doc2_dataset = read_sentences_from_file("./doc2.txt")
    doc3_dataset = read_sentences_from_file("./doc3.txt")
    actual_dataset_test = read_sentences_from_file("./test.txt")

    doc1 = UnigramLanguageModel(doc1_dataset, global_vocabs, mode="document", smoothing=True)
    doc2 = UnigramLanguageModel(doc2_dataset, global_vocabs, mode="document", smoothing=True)
    doc3 = UnigramLanguageModel(doc3_dataset, global_vocabs, mode="document", smoothing=True)
    collection = UnigramLanguageModel(actual_dataset, global_vocabs, mode="collection", smoothing=True)

    num_doc1 = 0
    num_doc2 = 0
    num_doc3 = 0


    for sentence in actual_dataset_test:
        print(f"query: {' '.join(sentence)}")

        prob_doc1 = calculate_interpolated_sentence_probability(sentence, doc1, collection)
        prob_doc2 = calculate_interpolated_sentence_probability(sentence, doc2, collection)
        prob_doc3 = calculate_interpolated_sentence_probability(sentence, doc3, collection)

        max_prob = max(prob_doc1, prob_doc2, prob_doc3)
        print("Interpolated Prob\ndoc1:", prob_doc1, "\ndoc2:", prob_doc2, "\ndoc3:", prob_doc3, "\n")

        choose_doc = ""

        if max_prob == prob_doc1:
            num_doc1 += 1
            choose_doc = "doc1"
        elif max_prob == prob_doc2:
            num_doc2 += 1
            choose_doc = "doc2"
        elif max_prob == prob_doc3:
            num_doc3 += 1
            choose_doc = "doc3"

        print(f"Most probable document: {choose_doc}")

    print("\n")
    print(f"Number of doc1 {num_doc1}")
    print(f"Number of doc2 {num_doc2}")
    print(f"Number of doc3 {num_doc3}")

    '''
        Question: for each of the test queries given in test.txt, find out best matching document/doc
        according to their interpolated sentence probability.
        Optional: Extend the model to bigram language modeling.
    '''
```

This is the result for this code.

```

query: <s> the website and monthly newsletter is run by a sub-committee that is independent to the parish council and is financed through selling advertisement space to local businesses </s>
Interpolated Prob
doc1: 2.06867347917177e-84
doc2: 4.832057773568354e-84
doc3: 5.841100583891947e-84

Most probable document: doc3
query: <s> uk was designed and built by chris chambers the site is designed in a way that when content is added to the site no previous content needs to be edited therefore creating a archive over time meaning that every single article that is added is stored and never deleted allowing users to search and read articles going far back as the website launch date </s>
Interpolated Prob
doc1: 4.881811503463049e-205
doc2: 5.8887576296024185e-205
doc3: 5.7738753795445575e-205

Most probable document: doc2
query: <s> chelmondiston is a small village and civil parish in suffolk england on the south bank of the river orwell and num miles south-east of ipswich </s>
Interpolated Prob
doc1: 2.8913969221756948e-77
doc2: 2.9740210951922733e-77
doc3: 4.767083580657596e-77
```

```
Most probable document: doc3
query: <s> formerly known as chelmington and was in the old hundreds of
suffolk of babergh </s>
Interpolated Prob
doc1: 2.5616731669652173e-42
doc2: 3.003325040465502e-42
doc3: 5.54224750358778e-42

Most probable document: doc3
query: <s> the former church site has not been a very lucky one </s>
Interpolated Prob
doc1: 1.3561872121008167e-35
doc2: 1.82917556228002e-35
doc3: 2.0892193354627122e-35

Most probable document: doc3
query: <s> one night in late num a v2 rocket hit hakewill church build
work and the little church was almost completely destroyed </s>
Interpolated Prob
doc1: 2.1605396902097275e-68
doc2: 1.6206420123871048e-68
doc3: 5.1737998443413365e-68

Most probable document: doc3
query: <s> it was not until num that basil hatcher was given the
commission to provide a replacement st andrew church </s>
Interpolated Prob
doc1: 4.999736942182e-60
doc2: 4.20012873828841e-60
doc3: 1.1067350959837566e-59

Most probable document: doc3
query: <s> the church includes a fine set of stained glass windows the num
work of francis skeat </s>
Interpolated Prob
doc1: 1.8692249698929316e-49
doc2: 1.6333792048145694e-49
doc3: 3.5523682336737106e-49

Most probable document: doc3
query: <s> there is a newer methodist church on the main road and a
baptist on pin mill road </s>
Interpolated Prob
doc1: 3.8068247006498425e-54
doc2: 5.066814258505366e-54
doc3: 1.0666128693269939e-53

Most probable document: doc3
query: <s> pubs in the area include the butt and oyster and the red lion
</s>
Interpolated Prob
doc1: 8.334917791248071e-39
doc2: 9.155199981407728e-39
doc3: 1.3011487284208682e-38
```

```
Most probable document: doc3
query: <s> dinan gallo dinan is a walled breton town and a commune in the
tes-d armor department in north-western france </s>
Interpolated Prob
doc1: 3.0347641262941492e-61
doc2: 3.423942861980613e-61
doc3: 4.3486718490936325e-61

Most probable document: doc3
query: <s> its geographical setting is exceptional </s>
Interpolated Prob
doc1: 1.4131042400358672e-17
doc2: 1.8934473664687355e-17
doc3: 1.8863105601319568e-17

Most probable document: doc2
query: <s> instead of nestling on the valley floor like morlaix most urban
development has been on the hillside overlooking the river rance </s>
Interpolated Prob
doc1: 4.179401939468469e-68
doc2: 7.952305313294264e-68
doc3: 5.680663069337795e-68

Most probable document: doc2
query: <s> the area alongside the river rance is known as the port of
dinan and is connected to the town by the steep streets rue jerzual and
its continuation outside the walls the rue de petit fort this river has
moderate turbidity and its brownish water is somewhat low in velocity due
to the slight gradient of the watercourse ph levels have been measured at
within the city of dinan and electrical conductivity of the waters have
tested at num micro-siemens per centimetre </s>
Interpolated Prob
doc1: 8.405855551990266e-256
doc2: 3.8500833454064486e-255
doc3: 2.729813630018384e-254

Most probable document: doc3
query: <s> in the center of dinan the rance summer flows are typically in
the range of num cubic feet per second </s>
Interpolated Prob
doc1: 1.203645830227692e-60
doc2: 6.529541042946083e-61
doc3: 1.852041548044045e-60

Most probable document: doc3
query: <s> for many years the bridge over the river rance at dinan was the
most northerly crossing point on the river but the tidal power station at
the mouth of the estuary constructed in the num downstream from dinan
incorporates a num meter long tidal barrage which doubles as a crossing
point nearer to the sea </s>
Interpolated Prob
doc1: 1.340343104034258e-169
doc2: 3.0966050601110305e-169
```

```
doc3: 4.673339984727321e-169

Most probable document: doc3
query: <s> the medieval town on the hilltop has many fine old buildings
some as early as num century </s>
Interpolated Prob
doc1: 8.591638551694525e-54
doc2: 1.507371865586215e-53
doc3: 1.145553362282478e-53

Most probable document: doc2
query: <s> the town retains a large section of the city walls part of
which can be walked around </s>
Interpolated Prob
doc1: 4.72632199302111e-53
doc2: 5.493489914042508e-53
doc3: 8.366994434197404e-53

Most probable document: doc3
query: <s> major historical attractions include the jacobins theatre
dating from num the flamboyant gothic st malo church the romanesque st
saviour basilica duchess anne tower and the cents teau de dinan </s>
Interpolated Prob
doc1: 4.5146770295459775e-101
doc2: 5.3772457146721465e-101
doc3: 5.272373053473593e-100

Most probable document: doc3
query: <s> a major highlight in the calendar is dinan te des remparts </s>
Interpolated Prob
doc1: 4.262636179516314e-35
doc2: 1.568899253168657e-35
doc3: 1.6976156725459216e-35

Most probable document: doc1
query: <s> the town is transformed with decoration and many locals dress
up in medieval garb for this two-day festival </s>
Interpolated Prob
doc1: 2.1607217985849332e-58
doc2: 2.8105850058576446e-58
doc3: 3.308449477390122e-58

Most probable document: doc3
query: <s> it occurs only every other year </s>
Interpolated Prob
doc1: 1.6529169393084515e-20
doc2: 6.946189593443052e-21
doc3: 8.599257299699822e-21

Most probable document: doc1
query: <s> inhabitants of dinan are called dinannais </s>
Interpolated Prob
doc1: 2.079041947563658e-20
doc2: 2.5660837138177766e-20
```

```
doc3: 3.032278494997937e-20

Most probable document: doc3
query: <s> in num of the children attended bilingual schools in primary
education </s>
Interpolated Prob
doc1: 9.512019812136916e-34
doc2: 8.364331407476082e-34
doc3: 9.895908154863033e-34

Most probable document: doc3
query: <s> dinan was also a favorite place for artists to visit in search
of picturesque views </s>
Interpolated Prob
doc1: 1.819605271225023e-47
doc2: 1.9018556577270946e-47
doc3: 3.1501542052891213e-47

Most probable document: doc3
query: <s> the british artist john everett millais lived there as a child
</s>
Interpolated Prob
doc1: 1.7805105040626853e-36
doc2: 1.887023801504471e-36
doc3: 2.1165483241056905e-36

Most probable document: doc3
query: <s> it was also painted by edward ward and horace tuck among others
</s>
Interpolated Prob
doc1: 4.528654044032606e-40
doc2: 3.511320336453354e-40
doc3: 2.996824226535092e-40

Most probable document: doc1
query: <s> guingamp is a commune in the tes-d armor department in brittany
in north-western france </s>
Interpolated Prob
doc1: 5.204622980902006e-44
doc2: 6.197163397739529e-44
doc3: 7.045403991418316e-44

Most probable document: doc3
query: <s> inhabitants of guingamp are called guingampais </s>
Interpolated Prob
doc1: 2.079041947563658e-20
doc2: 2.5660837138177766e-20
doc3: 3.032278494997937e-20

Most probable document: doc3
query: <s> the municipality launched a linguistic plan through ya
brezhoneg on num july num </s>
Interpolated Prob
doc1: 5.484899650032761e-42
```

```
doc2: 4.2991434144884727e-42
doc3: 4.909838343154321e-42

Most probable document: doc1
query: <s> in num of the children attended the bilingual schools in
primary education </s>
Interpolated Prob
doc1: 1.551175910623506e-35
doc2: 1.4386130077228085e-35
doc3: 1.7621840830828082e-35

Most probable document: doc3
query: <s> the breton dance festival of saint-loup is held every year in
mid-august </s>
Interpolated Prob
doc1: 1.954353585838299e-38
doc2: 1.925237702900165e-38
doc3: 1.5669246537462867e-38

Most probable document: doc1
query: <s> then there is the annual which brings pilgrims to pay homage to
the black in the basilica of notre dame de bon secours </s>
Interpolated Prob
doc1: 3.1362933017900096e-72
doc2: 5.161195133308263e-72
doc3: 7.604082664078553e-72

Most probable document: doc3
query: <s> guingamp is home to the num coupe de france holders en avant de
guingamp a football team in ligue num the second-highest league in french
football </s>
Interpolated Prob
doc1: 7.4006188053682195e-84
doc2: 6.676910824765447e-84
doc3: 1.3570027570605087e-83

Most probable document: doc3
query: <s> the town has like many others in the region a rich and
interesting history </s>
Interpolated Prob
doc1: 3.657357570405796e-43
doc2: 4.8653604008079636e-43
doc3: 7.927780701246691e-43

Most probable document: doc3
query: <s> this is exemplified in the remains of s three castles razed to
ground level by the order of richelieu and now reduced to three towers
</s>
Interpolated Prob
doc1: 3.9971250472570055e-77
doc2: 3.8195371062827755e-77
doc3: 9.858065086949427e-77

Most probable document: doc3
```

```
query: <s> vincent de bourbon great grandson of louis xiv was count of
guingamp from num till his death in num </s>
Interpolated Prob
doc1: 2.0656331327670926e-61
doc2: 2.2520188577111437e-61
doc3: 2.8850431111649694e-61

Most probable document: doc3
query: <s> lannion is a commune in the tes-d armor department of brittany
in north-western france </s>
Interpolated Prob
doc1: 8.349831819860936e-44
doc2: 9.678866037149346e-44
doc3: 1.1259443174952766e-43

Most probable document: doc3
query: <s> it is a sous-pr fecture of tes-d armor the capital of gor and
the center of an urban area of almost inhabitants </s>
Interpolated Prob
doc1: 4.913059929355598e-66
doc2: 5.011811744816403e-66
doc3: 6.766184884453187e-66

Most probable document: doc3
query: <s> inhabitants of lannion are called lannionnais </s>
Interpolated Prob
doc1: 2.079041947563658e-20
doc2: 2.5660837138177766e-20
doc3: 3.032278494997937e-20

Most probable document: doc3
query: <s> lannion takes its name from lann huon in breton or land of huon
in english </s>
Interpolated Prob
doc1: 1.269098444879289e-49
doc2: 9.03060823123984e-49
doc3: 2.8288612415081055e-49

Most probable document: doc2
query: <s> most of the area indeed use to belong to lord huon </s>
Interpolated Prob
doc1: 1.405432684193616e-34
doc2: 2.281271376453339e-34
doc3: 2.0478839135304133e-34

Most probable document: doc2
query: <s> the old quarter of lannion attracts many tourists to the city
</s>
Interpolated Prob
doc1: 2.3067800704529225e-34
doc2: 2.6382868478842495e-34
doc3: 3.104387358766442e-34

Most probable document: doc3
```

```
query: <s> the old quarter contains old squares a church called venez
half-timbered houses chapels and frescoes </s>
Interpolated Prob
doc1: 1.2835158561234358e-51
doc2: 1.7569419803800887e-51
doc3: 3.595706086000904e-51

Most probable document: doc3
query: <s> on num october num the municipality launched a linguistic plan
to promote breton language through the ya brezhoneg yes to breton charter
</s>
Interpolated Prob
doc1: 7.798761025912808e-71
doc2: 7.442150820162508e-71
doc3: 9.689789375041292e-71

Most probable document: doc3
query: <s> in num of the children attended bilingual schools in primary
education </s>
Interpolated Prob
doc1: 9.512019812136916e-34
doc2: 8.364331407476082e-34
doc3: 9.895908154863033e-34

Most probable document: doc3
query: <s> lannion is a large telecommunications research center in france
with several firms such as alcatel-lucent orange france telecom and sagem
operating there </s>
Interpolated Prob
doc1: 3.181900220658696e-74
doc2: 3.3393245051670758e-74
doc3: 4.124064669254868e-74

Most probable document: doc3
query: <s> regular concerts are held in the town square during the summer
months known as les tardives </s>
Interpolated Prob
doc1: 3.7056665868134834e-52
doc2: 3.2000611086927653e-52
doc3: 2.1681554303667573e-52

Most probable document: doc1
query: <s> lannion is also home to the magique a well known theater
company in the area </s>
Interpolated Prob
doc1: 3.251563687710432e-46
doc2: 3.646547759327327e-46
doc3: 6.813898501334283e-46

Most probable document: doc3
query: <s> lannion is served by extensive transport links </s>
Interpolated Prob
doc1: 8.605293392386266e-25
doc2: 9.107801991113934e-25
```

```
doc3: 9.053833620103095e-25

Most probable document: doc2
query: <s> the nearby airport lannion te de granit airport was recently
expanded to accommodate larger flights arriving from paris and other
french destinations </s>
Interpolated Prob
doc1: 1.4210025297333835e-75
doc2: 2.2728230120881583e-75
doc3: 4.526474178631939e-75

Most probable document: doc3
query: <s> the station provides tgv services to brest brieuc rennes and
paris as well as ter links to local stations </s>
Interpolated Prob
doc1: 3.647277871308314e-63
doc2: 4.6098866560187813e-63
doc3: 1.2518807744287826e-62

Most probable document: doc3
query: <s> saint-brieuc breton sant-brieg gallo saent-berioec is a commune
in the tes-d armor department in brittany in north-western france </s>
Interpolated Prob
doc1: 4.69241157991566e-59
doc2: 5.291139795400612e-59
doc3: 5.989104419017699e-59

Most probable document: doc3
query: <s> it has a cathedral </s>
Interpolated Prob
doc1: 1.1887087415535818e-12
doc2: 1.3614998747902798e-12
doc3: 1.027277010869512e-12

Most probable document: doc2
query: <s> saint-brieuc is named after a welsh monk brioc who evangelized
the region in the num century and established an oratory there </s>
Interpolated Prob
doc1: 1.413846972232076e-65
doc2: 3.977469439371831e-65
doc3: 2.4526630180835535e-65

Most probable document: doc2
query: <s> bro de saint-brieuc one of the nine traditional bishoprics of
brittany which were used as administrative areas before the french
revolution was named after saint-brieuc </s>
Interpolated Prob
doc1: 1.435800948064984e-80
doc2: 4.559306417419135e-80
doc3: 1.3155355030731692e-80

Most probable document: doc2
query: <s> saint-brieuc is one of the towns in europe that hosts the iu
honors program </s>
```

```
Interpolated Prob
doc1: 5.673246000916324e-43
doc2: 7.541541445620983e-43
doc3: 6.922671224308266e-43

Most probable document: doc2
query: <s> the cemetery of saint michel contains graves of several notable
bretons and sculptures by paul le goff and jean boucher </s>
Interpolated Prob
doc1: 3.6994038249735896e-66
doc2: 3.2187138751964634e-66
doc3: 4.547505430306867e-66

Most probable document: doc3
query: <s> outside the wall is armel beaufils statue of anatole le braz
</s>
Interpolated Prob
doc1: 6.203663071430771e-37
doc2: 6.689477124404356e-37
doc3: 7.207494631140045e-37

Most probable document: doc3
query: <s> le goff who was killed with his two brothers in world war i is
also commemorated in a street and with his major sculptural work la forme
se gageant de la re in the central gardens which also includes a memorial
to him by jules-charles le bozec and work by francis renaud </s>
Interpolated Prob
doc1: 1.3295978941573395e-168
doc2: 1.2219539928436755e-168
doc3: 2.3454964495329637e-168

Most probable document: doc3
query: <s> the town of brieux in saskatchewan canada is named after saint-
brieuc of brittany </s>
Interpolated Prob
doc1: 1.6756016882765874e-40
doc2: 5.312376487499578e-40
doc3: 2.1800976256825654e-40

Most probable document: doc2
query: <s> it was founded by immigrants from this region in brittany </s>
Interpolated Prob
doc1: 5.222930244612062e-32
doc2: 8.913464020136339e-32
doc3: 4.636995214521987e-32

Most probable document: doc2
query: <s> it was settled in the early num </s>
Interpolated Prob
doc1: 1.290822101808736e-19
doc2: 1.1982146534079151e-19
doc3: 1.0141671749167324e-19

Most probable document: doc1
```

```
query: <s> the town is located by the english channel in the bay of saint-
brieuc </s>
Interpolated Prob
doc1: 5.481337259939806e-38
doc2: 6.909241336802038e-38
doc3: 8.201656666087094e-38

Most probable document: doc3
query: <s> two rivers flow through saint-brieuc the t and the dic </s>
Interpolated Prob
doc1: 4.6793214225265477e-32
doc2: 4.8020400004882084e-32
doc3: 5.735136278259441e-32

Most probable document: doc3
query: <s> other towns of notable size in the partement of tes are dinan
and all sous-pr fectures </s>
Interpolated Prob
doc1: 1.7208359555738055e-50
doc2: 1.4517749472249194e-50
doc3: 2.132901634708337e-50

Most probable document: doc3
query: <s> in num large amounts of sea lettuce a type of algae washed up
on many beaches of brittany and when it rotted it emitted dangerous levels
of hydrogen sulphide </s>
Interpolated Prob
doc1: 1.7757440596733716e-92
doc2: 9.418008220665403e-93
doc3: 1.148290547787817e-92

Most probable document: doc1
query: <s> a horse and some dogs died and a council worker driving a
truckload of it fell unconscious at the wheel and died </s>
Interpolated Prob
doc1: 3.711654755435629e-68
doc2: 4.394532984510055e-68
doc3: 5.668899567902865e-68

Most probable document: doc3
query: <s> the beach at saint-brieuc suffered bad damage and had to be
shut </s>
Interpolated Prob
doc1: 1.0598711202385091e-38
doc2: 1.141393039519684e-38
doc3: 1.7843039531598955e-38

Most probable document: doc3
query: <s> langueux la augon rin ploufragan gueux and muson </s>
Interpolated Prob
doc1: 2.9545658881273156e-29
doc2: 2.6829975648801457e-29
doc3: 2.985513810214616e-29
```

```
Most probable document: doc3
query: <s> inhabitants of saint-brieuc are called briochins or briochains
</s>
Interpolated Prob
doc1: 6.331336484578037e-27
doc2: 1.0381955899333389e-26
doc3: 1.066808468346599e-26

Most probable document: doc3
query: <s> in num of the children attended the bilingual schools in
primary education </s>
Interpolated Prob
doc1: 1.551175910623506e-35
doc2: 1.4386130077228085e-35
doc3: 1.7621840830828082e-35

Most probable document: doc3
query: <s> the gare de saint-brieuc railway station is connected by tgv
atlantique to paris montparnasse station </s>
Interpolated Prob
doc1: 3.075834454454937e-51
doc2: 3.5562680417301707e-51
doc3: 7.255278367832452e-51

Most probable document: doc3
query: <s> an air service from saint-brieuc armor airport to newquay in
cornwall is operated by isles of scilly skybus four days per week </s>
Interpolated Prob
doc1: 9.499078756003669e-74
doc2: 4.460437115427016e-74
doc3: 5.2432188040663774e-74

Most probable document: doc1
query: <s> bourseul gallo rsoeut is a commune in the tes-d armor
department in bretagne in north-western france </s>
Interpolated Prob
doc1: 1.56276143875768e-51
doc2: 1.810802525743204e-51
doc3: 2.0541582261054865e-51

Most probable document: doc3
query: <s> inhabitants of bourseul are called bourseulais </s>
Interpolated Prob
doc1: 2.079041947563658e-20
doc2: 2.5660837138177766e-20
doc3: 3.032278494997937e-20

Most probable document: doc3
query: <s> stoke by nayland in the english county of suffolk lies close to
the border with essex in what is sometimes referred to as constable
country </s>
Interpolated Prob
doc1: 7.132317672107545e-78
doc2: 1.0577831481149523e-77
```

```
doc3: 1.4538528470167791e-77

Most probable document: doc3
query: <s> it contains a church st mary part of the deanery of hadleigh in
the diocese of chelmsford </s>
Interpolated Prob
doc1: 4.3587082987943045e-51
doc2: 4.1546561810433474e-51
doc3: 2.169246535137942e-50

Most probable document: doc3
query: <s> the incumbent is the revd </s>
Interpolated Prob
doc1: 2.662918139394243e-14
doc2: 3.275114249718185e-14
doc3: 3.502151408646949e-14

Most probable document: doc3
query: <s> v armstrong </s>
Interpolated Prob
doc1: 3.0026410068358886e-08
doc2: 2.9219860919008703e-08
doc3: 2.9156003383305802e-08

Most probable document: doc1
query: <s> the village located within babergh district contains many
cottages and timber framed houses and all surround a large recreation
field which makes up the center of the village </s>
Interpolated Prob
doc1: 1.2090688578887644e-91
doc2: 1.3027628843142716e-91
doc3: 3.530611574790511e-91

Most probable document: doc3
query: <s> the population of the stoke-by-nayland civil parish at the num
census was num comprising num males and num females </s>
Interpolated Prob
doc1: 2.623767676718617e-55
doc2: 1.6562077212705663e-55
doc3: 2.4263401679516104e-55

Most probable document: doc1
query: <s> two schools a primary and a middle school are in the village as
are two public houses the angel inn and the crown </s>
Interpolated Prob
doc1: 2.95268340933078e-69
doc2: 2.712809935820428e-69
doc3: 5.497150202826608e-69

Most probable document: doc3
query: <s> stoke by nayland is about miles from nayland and stands on a
ridge overlooking the stour and box valleys </s>
Interpolated Prob
doc1: 5.4580113863119514e-61
```

```
doc2: 5.482706310687335e-61
doc3: 4.80911324638082e-61

Most probable document: doc2
query: <s> immediately to the north of the village lies the hamlet of
scotland street </s>
Interpolated Prob
doc1: 1.0388417432632586e-38
doc2: 1.2366457611508498e-38
doc3: 1.5798523026978213e-38

Most probable document: doc3
query: <s> st mary church was rebuilt in the fifteenth century and
renovated in num </s>
Interpolated Prob
doc1: 5.218922662987555e-40
doc2: 5.080561056401103e-40
doc3: 2.38673681319946e-39

Most probable document: doc3
query: <s> the church is on the site of a num century minster </s>
Interpolated Prob
doc1: 7.974481928007722e-31
doc2: 9.22724484531569e-31
doc3: 1.9210226739520473e-30

Most probable document: doc3
query: <s> a saxon monastery was founded here during the time of king
edmund by earl alfgar who died in num </s>
Interpolated Prob
doc1: 4.893682342613935e-60
doc2: 1.4445290271197034e-60
doc3: 1.9160960602452367e-60

Most probable document: doc1
query: <s> the num national gazetteer of great britain describes the
village such </s>
Interpolated Prob
doc1: 2.1325814002312858e-34
doc2: 1.870892864183069e-34
doc3: 2.1769814065748977e-34

Most probable document: doc3
query: <s> stoke-by-nayland a parish in the hundred of babergh county
suffolk num mile of nayland and num miles of bures railway station </s>
Interpolated Prob
doc1: 3.7934350002117285e-65
doc2: 2.609321419150703e-65
doc3: 3.944551466627415e-65

Most probable document: doc3
query: <s> colchester is its post town </s>
Interpolated Prob
doc1: 1.2832448202307842e-17
```

```
doc2: 1.7173408391452845e-17
doc3: 1.7106983497908842e-17

Most probable document: doc2
query: <s> the village which was formerly a market town is situated near
the river stour </s>
Interpolated Prob
doc1: 1.257880049818965e-44
doc2: 1.8735217887697266e-44
doc3: 1.9377116230660464e-44

Most probable document: doc3
query: <s> the parish contains the chapelry of leavenheath and had a
monastery endowed by the saxon earl of algar traces of which are still
existing </s>
Interpolated Prob
doc1: 1.9761013587831016e-73
doc2: 2.0249375165984936e-73
doc3: 5.702472543703445e-73

Most probable document: doc3
query: <s> the living is a vicarage in the diocese of ely value num </s>
Interpolated Prob
doc1: 4.12281191573753e-34
doc2: 4.414904482515214e-34
doc3: 5.4164834480801046e-34

Most probable document: doc3
query: <s> the church dedicated to mary is an ancient structure with a
tower and six bells </s>
Interpolated Prob
doc1: 4.6388041226369145e-47
doc2: 5.737901874116101e-47
doc3: 1.2407513214754869e-46

Most probable document: doc3
query: <s> there is also a district church at leavenheath the living of
which is a perpetual curacy value num </s>
Interpolated Prob
doc1: 1.3379496801839427e-54
doc2: 1.4230507228104398e-54
doc3: 3.7829720404264065e-54

Most probable document: doc3
query: <s> the parochial charities produce about num per annum exclusive
of some almshouses </s>
Interpolated Prob
doc1: 1.4288340909698129e-39
doc2: 8.924433062615427e-40
doc3: 7.5467802934012e-40

Most probable document: doc1
query: <s> num go towards lady windsor hospital </s>
Interpolated Prob
```

```
doc1: 8.670472632225368e-22
doc2: 6.960711873131749e-22
doc3: 7.208003865825148e-22

Most probable document: doc1
query: <s> there is a national school for both sexes </s>
Interpolated Prob
doc1: 7.132918041123936e-26
doc2: 8.424106592848712e-26
doc3: 8.611811033911283e-26

Most probable document: doc3
query: <s> tendring hall is the principal residence </s>
Interpolated Prob
doc1: 4.9031245115084665e-20
doc2: 5.564060558277e-20
doc3: 5.734126791394035e-20

Most probable document: doc3
query: <s> the birthstone of april is the diamond and the birth flower is
typically listed as either the daisy or the sweet pea </s>
Interpolated Prob
doc1: 7.492556358098532e-65
doc2: 1.56909532723612e-64
doc3: 1.193798895948853e-61

Most probable document: doc3



Number of doc1 14
Number of doc2 15
Number of doc3 72
```