Praktikum 5 - Model Probabilistik Bahasa

October 25, 2021

1 Model Probabilistik Bahasa

Corpus

[134]:
```
corpus = """Can you tell me any good cantonese restaurants close by?
Mid priced Thai Food is what i'm looking for.
Tell me about chez panisse!
Can you give me a listing of the kinds of food that are available?
I'm looking for a good place to eat breakfast.
When is caffe venezia open during the day.
"""

#corpus = """I am Sam.
#Sam I am.
#I do not like green eggs and ham.
#"""

print(corpus)
```

```
Can you tell me any good cantonese restaurants close by?
Mid priced Thai Food is what i'm looking for.
Tell me about chez panisse!
Can you give me a listing of the kinds of food that are available?
I'm looking for a good place to eat breakfast.
When is caffe venezia open during the day.
```

Tokenize Per Kalimat

[135]:
```
import nltk

list_kalimat = nltk.sent_tokenize(corpus)

print(list_kalimat)
```

```
['Can you tell me any good cantonese restaurants close by?', "Mid priced Thai
Food is what i'm looking for.", 'Tell me about chez panisse!', 'Can you give me
a listing of the kinds of food that are available?', "I'm looking for a good
place to eat breakfast.", 'When is caffe venezia open during the day.']
```

Lowercase Semua Huruf

[136]:
```
for indeks, kalimat in enumerate(list_kalimat):
    list_kalimat[indeks] = kalimat.lower()

print(list_kalimat)
```

```
['can you tell me any good cantonese restaurants close by?', "mid priced thai
food is what i'm looking for.", 'tell me about chez panisse!', 'can you give me
a listing of the kinds of food that are available?', "i'm looking for a good
place to eat breakfast.", 'when is caffe venezia open during the day.']
```

Punctuation Removal

[137]:
```
import re

pola_simbol = r'[^\w\s]'

for indeks, kalimat in enumerate (list_kalimat):
    list_kalimat[indeks] = re.sub(pola_simbol,'',kalimat)

print(list_kalimat)
```

```
['can you tell me any good cantonese restaurants close by', 'mid priced thai
food is what im looking for', 'tell me about chez panisse', 'can you give me a
listing of the kinds of food that are available', 'im looking for a good place
to eat breakfast', 'when is caffe venezia open during the day']
```

Penambahan Pseudo String

[138]:
```
for indeks, kalimat in enumerate(list_kalimat):
    list_kalimat[indeks] = "^ " + kalimat + " $"

for kalimat in list_kalimat:
    print(kalimat)
```

```
^ can you tell me any good cantonese restaurants close by $
^ mid priced thai food is what im looking for $
^ tell me about chez panisse $
^ can you give me a listing of the kinds of food that are available $
^ im looking for a good place to eat breakfast $
^ when is caffe venezia open during the day $
```

Percobaan Bigram

[139]:
```
from nltk.util import ngrams

kalimat = list_kalimat[0]
list_kata = nltk.word_tokenize(kalimat)
```

```
n_gram = 2

bigrams = ngrams(list_kata,n_gram)

print("List Kata")
print(list_kata)
print()

print("Bigram")
list_gram = []
for gram in bigrams:
    list_gram.append(list(gram))

print(list_gram)
```

```
List Kata
['^', 'can', 'you', 'tell', 'me', 'any', 'good', 'cantonese', 'restaurants',
'close', 'by', '$']

Bigram
[['^', 'can'], ['can', 'you'], ['you', 'tell'], ['tell', 'me'], ['me', 'any'],
['any', 'good'], ['good', 'cantonese'], ['cantonese', 'restaurants'],
['restaurants', 'close'], ['close', 'by'], ['by', '$']]
```

Listing All Gram

[140]:
```
from nltk.util import ngrams

n_gram = 2

print("List Kalimat")
for kalimat in list_kalimat:
    print(kalimat)

print()
print("List n-gram")
list_ngram = []
for kalimat in list_kalimat:
    list_kata = nltk.word_tokenize(kalimat)
    bigrams = ngrams(list_kata,n_gram)
    for gram in bigrams:
        list_ngram.append(list(gram))

for item in list_ngram:
    print(item)
```

```
List Kalimat
```

```
^ can you tell me any good cantonese restaurants close by $
^ mid priced thai food is what im looking for $
^ tell me about chez panisse $
^ can you give me a listing of the kinds of food that are available $
^ im looking for a good place to eat breakfast $
^ when is caffe venezia open during the day $

List n-gram
['^', 'can']
['can', 'you']
['you', 'tell']
['tell', 'me']
['me', 'any']
['any', 'good']
['good', 'cantonese']
['cantonese', 'restaurants']
['restaurants', 'close']
['close', 'by']
['by', '$']
['^', 'mid']
['mid', 'priced']
['priced', 'thai']
['thai', 'food']
['food', 'is']
['is', 'what']
['what', 'im']
['im', 'looking']
['looking', 'for']
['for', '$']
['^', 'tell']
['tell', 'me']
['me', 'about']
['about', 'chez']
['chez', 'panisse']
['panisse', '$']
['^', 'can']
['can', 'you']
['you', 'give']
['give', 'me']
['me', 'a']
['a', 'listing']
['listing', 'of']
['of', 'the']
['the', 'kinds']
['kinds', 'of']
['of', 'food']
['food', 'that']
['that', 'are']
```

```
['are', 'available']
['available', '$']
['^', 'im']
['im', 'looking']
['looking', 'for']
['for', 'a']
['a', 'good']
['good', 'place']
['place', 'to']
['to', 'eat']
['eat', 'breakfast']
['breakfast', '$']
['^', 'when']
['when', 'is']
['is', 'caffe']
['caffe', 'venezia']
['venezia', 'open']
['open', 'during']
['during', 'the']
['the', 'day']
['day', '$']
```

Pembuatan Bag of Words

[141]:
```
list_semua_kata = []

for kalimat in list_kalimat:
    list_kata = nltk.word_tokenize(kalimat)
    list_semua_kata.extend(list_kata)

himpunan_kata = set(list_semua_kata)

bow = {}

for kata in himpunan_kata:
    bow[kata] = list_semua_kata.count(kata)

print(bow)
```

```
{'any': 1, 'venezia': 1, 'of': 2, '$': 6, 'place': 1, 'thai': 1, 'eat': 1,
'you': 2, 'cantonese': 1, 'during': 1, 'is': 2, 'available': 1, 'close': 1,
'breakfast': 1, 'kinds': 1, 'looking': 2, 'about': 1, 'caffe': 1, 'by': 1, 'a':
2, 'panisse': 1, 'are': 1, 'im': 2, 'tell': 2, 'mid': 1, 'when': 1, 'priced': 1,
'the': 2, 'day': 1, 'chez': 1, 'food': 2, 'listing': 1, 'for': 2, '^': 6,
'restaurants': 1, 'that': 1, 'me': 3, 'open': 1, 'can': 2, 'give': 1, 'what': 1,
'to': 1, 'good': 2}
```

Menghitung Probabilitas Kata Berikutnya

[148]:
```
#contoh string_sejarah = "^", string_prediksi = "i"
#maka akan melakukan perhitungan berapa kemungkinan kemunculan string "^ i"
#p(i|^)

string_sejarah = "place"
string_prediksi = "good"

string_prediksi_lengkap = [string_sejarah,string_prediksi]

n_spl = list_ngram.count(string_prediksi_lengkap) #jumlah ditemukannya string␣
 ↪prediksi lengkap
n_sj =  bow[string_sejarah] #jumlah ditemukannya string sejarah

probabilitas = n_spl / n_sj

print("P(%s|%s) = %d/%d = %f" %␣
 ↪(string_prediksi,string_sejarah,n_spl,n_sj,probabilitas))
```

```
P(good|place) = 0/1 = 0.000000
```

[ ]:
```

```