Pengenalan Natural Language Toolkit (NLTK) Bagian 2 (Reuters Corpus)

Yunita Sari

yunita.sari@ugm.ac.id

Departemen Ilmu Komputer dan Elektronika, UGM

Desember 2019

Pada bagian ke-2 dari tutorial NLTK kali ini, kita akan meng-eskplore beberapa fitur dan resource lain. NLTK menyediakan beberapa *Text Corpora and Lexical Resources* dalam berbagai bentuk. Diantaranya: Buku, Gutenberg Corpus, Web & chat Text, Brown dan Reuters Corpus. Corpora ini bisa kita gunakan untuk meng-eksplore berbagai macam fitur yang tersedia. Tutorial kali ini akan mengeksplore beberapa NLTK corpus.

Corpus pada NLTK biasanya belum terinstal secara langsung ketika instalasi awal. Untuk menginstal corpus, kita bisa men-spesifikan package apa yang akan diinstall. Pada contoh ini, corpus Reuters telah terinstall.

```
In [11]: import nltk
         nltk.download("reuters")
```

```
[nltk_data] Downloading package reuters to /home/yunita/nltk_data...
[nltk_data]   Package reuters is already up-to-date!
```

```
Out[11]: True
```

Reuters corpus dikembangkan dari kumpulan berita dari tahun 1987. Pada tahun 1990, dokumen pada Reuters corpus dipublikasikan secara luas akan tetapi hanya untuk keperluan riset saja. Reuters Corpus yang asli terdiri dari 22173 dokumen sehingga disebut "Reuters-22173" corpus. Versi yang asli ini kemudian dimodifikasi dengan membuang dokumen-dokumen yang sama, menghasilkan Reuters-21578. Reuters corpus pada NLTK terdiri dari 10788 dokumen dan mengandung toal 1.3 juta kata. Dokumen pada Reuters corpus telah diklasifikasikan ke 90 topik/kategori dan dibagi menjadi 2 set sebagai training dan test.

```
In [17]: from nltk.corpus import reuters
         categories = reuters.categories()
         print("Jumlah Topik/Kategori:",len(categories))
         print(categories[0:9])
```

```
Jumlah Topik/Kategori: 90
['acq', 'alum', 'barley', 'bop', 'carcass', 'castor-oil', 'cocoa', 'coconut', 'coconut-oil']
```

Beberapa topik/kategori dalam Reuters corpus diantara-nya adalah *acq, alum, barley, bop, carcass, castor-oil* Kita juga bisa melihat apakah sebuah dokumen/file di Reuters corpus masuk ke dalam test atau training set. Sebagai contoh test/14826 berarti file dengan id: 14826 masuk dalam test set

```
In [ ]: reuters.fileids()
```

Pada Reuters corpus, sebuah dokumen bisa masuk ke dalam beberapa kategori sekaligus. Sebagai contoh, dokumen dengan file id: 9865 masuk ke training data dengan kategori *barley, corn, grain* dan *wheat*

```
In [19]: reuters.categories('training/9865')

Out[19]: ['barley', 'corn', 'grain', 'wheat']

In [20]: reuters.categories(['training/9865', 'training/9880'])

Out[20]: ['barley', 'corn', 'grain', 'money-fx', 'wheat']

In [ ]: reuters.fileids(['barley', 'corn'])
```

Selanjutnya, kita bisa mengeksplorasi Reuters Corpus, misalkan dengan menghitung total kata keseluruhan, maupun total kata dalam sebuah file maupun kategori. Selain itu, kita bisa juga melihat kata apa saja yang muncul dalam sebuah file.

```
In [23]: total_words = reuters.words()
         print("number of words", len(total_words) )

number of words 1720901


In [25]: individual_file_words = reuters.words('training/9865')
         print("number of words", len(individual_file_words))

number of words 114


In [28]: tradeWords = reuters.words(categories = 'trade')
         print("number of words", len(tradeWords))

number of words 142723


In [26]: reuters.words('training/9865')[:14]

Out[26]: ['FRENCH',
          'FREE',
          'MARKET',
          'CEREAL',
          'EXPORT',
          'BIDS',
```

```
'DETAILED',
'French',
'operators',
'have',
'requested',
'licences',
'to',
'export']
```

Dengan Reuters corpus ini, kita bisa mengaplikasikan fitur-fitur NLTK yang telah dijelaskan pada bagian 1 dari tutorial ini. Selain, itu kita bisa juga mulai meng-eksplor NLP task seperti *topic classification*