Praktikum 13 - Part of Speech Tagging

January 2, 2022

1 Part of Speech Tagging

Pada praktikum kali ini, kita akan belajar salah satu contoh penerapan Part of Speech Tagging. Seperti yang kita ketahui, Part of Speeh tagging pada prinsipnya adalah memberikan penanda terhdap setiap token berdasarkan posisi/peran token tersebut pada sebuah dokumen. Terdapat cukup banyak *tag set* yang dapat dilakukan untuk melakukan *tagging*. Namun pada praktikum kali ini kita akan menggunakan *tagging* bawaan dari NLTK, yaitu Penn Treebank.

Skenarionya, kita akan menggunakan sebuah dokumen untuk dilakukan penandaan. Nantinya dokumen tersebut akan dibuat menjadi dua versi. Dokumen versi pertama akan dilakukan penandaan tanpa dilakukan proses pra-pengolahan teks terlebih dahulu. Dokumen versi kedua akan dilakukan penandaan setelah dilakukan proses pra pengolahan teks. Nantinya akan kita bandingkan hasil dari dua versi dokumen tersebut.

1.1 Mendefinisikan Teks/Dokumen

Pada tahap ini kita definisikan terlebih dahulu teks yang akan dilakukan penandaan. Pada contoh kali ini saya ambil dari narasi intro film Star Wars.

[31]:

```
dokumen = """It is a period of civil war.
Rebel spaceships, striking from a hidden base, have won their first victory
against the evil Galactic Empire. During the battle,
Rebel spies managed to steal secret plans to the
Empire’s ultimate weapon, the DEATH STAR, an armored
space station with enough power to destroy an entire
planet. Pursued by the Empire’s sinister agents,
Princess Leia races home aboard her starship, custodian of
the stolen plans that can save her people and restore freedom to the galaxy ...
"""

print(dokumen)
```

```
It is a period of civil war.
Rebel spaceships, striking from a hidden base, have won their first victory
against the evil Galactic Empire. During the battle,
Rebel spies managed to steal secret plans to the
Empire’s ultimate weapon, the DEATH STAR, an armored
space station with enough power to destroy an entire
```

```
planet. Pursued by the Empire’s sinister agents,
Princess Leia races home aboard her starship, custodian of
the stolen plans that can save her people and restore freedom to the galaxy …
```

1.2 Membuat Cheat Sheet Penn Treebank Tagset

Pada bagian ini, kita akan mendefinisikan cheat sheet yang nantinya akan memudahkan kita untuk menerjemahkan tag dari masing-masin token.

[5]:

```
tagset_helper = {
    'CC' : 'coordinating conjunction',
    'CD' : 'cardinal number',
    'DT' : 'determiner',
    'EX' : 'existensial "there"',
    'FW' : 'foreign word',
    'IN' : 'prepotition/subordin-conj',
    'JJ' : 'adjective',
    'JJR' : 'comparative adjective',
    'JJS' : 'superlative adjective"',
    'LS' : 'list item marker',
    'MD' : 'modal',
    'NN' : 'sing or mass noun',
    'NNS' : 'noun, plural',
    'NNP' : 'proper noun, sing.',
    'NNPS' : 'proper noun, plu.',
    'PDT' : 'predeterminer',
    'POS' : 'possesive ending',
    'PRP' : 'personal pronoun',
    'PRP$' : 'possess, pronoun',
    'RB' : 'adverb',
    'RBR' : 'comparative adverb',
    'RBS' : 'superlatv, adverb',
    'RP' : 'particle',
    'SYM' : 'symbol',
    'TO' : '"to"',
    'UH' : 'interjection',
    'VB' : 'verb base form',
    'VBD' : 'verb past tense',
    'VBG' : 'verb gerund',
    'VBN' : 'verb past part',
    'VBP' : 'verb non-3sg present',
    'VBZ' : 'verb 3sg pres',
    'WDT' : 'wh-determ.',
    'WP' : 'wh-pronoun',
    'WP$' : 'wh-possess.',
    'WRB' : 'wh-adverb',
    '$' : 'dollar sign',
    '#' : 'pound sign',
    '".' : 'left quote',
    '."' : 'right quote',
    '(' : 'left paren',
    ')' : 'right paren',
    ',' : 'comma',
    '.' : 'sent-end punc',
    ':' : 'sent-mid punc',
}

print(tagset_helper)
```

```
{'CC': 'coordinating conjunction', 'CD': 'cardinal number', 'DT': 'determiner',
'EX': 'existensial "there"', 'FW': 'foreign word', 'IN': 'prepotition/subordin-
conj', 'JJ': 'adjective', 'JJR': 'comparative adjective', 'JJS': 'superlative
adjective"', 'LS': 'list item marker', 'MD': 'modal', 'NN': 'sing or mass noun',
'NNS': 'noun, plural', 'NNP': 'proper noun, sing.', 'NNPS': 'proper noun, plu.',
'PDT': 'predeterminer', 'POS': 'possesive ending', 'PRP': 'personal pronoun',
'PRP$': 'possess, pronoun', 'RB': 'adverb', 'RBR': 'comparative adverb', 'RBS':
'superlatv, adverb', 'RP': 'particle', 'SYM': 'symbol', 'TO': '"to"', 'UH':
'interjection', 'VB': 'verb base form', 'VBD': 'verb past tense', 'VBG': 'verb
gerund', 'VBN': 'verb past part', 'VBP': 'verb non-3sg present', 'VBZ': 'verb
3sg pres', 'WDT': 'wh-determ.', 'WP': 'wh-pronoun', 'WP$': 'wh-possess.', 'WRB':
'wh-adverb', '$': 'dollar sign', '#': 'pound sign', '".': 'left quote', '."':
'right quote', '(': 'left paren', ')': 'right paren', ',': 'comma', '.': 'sent-
end punc', ':': 'sent-mid punc'}
```

1.3 Skenario Dokumen Dengan Pra Pengolahan Teks

Pada skenario ini dokumen akan dilakukan *tagging* setelah melalui pra pengolahan teks.

1.3.1 Pra Pengolahan Teks Terhadap Dokumen

[11]:
```
import re
import nltk

#polaSimbol menyimbolkan regex daftar simbol
polaSimbol = r'[^\w\s]'

# dokumen_prerocessed menyimpan dokumen yang telah melalui pra pengolahan teks
dokumen_prerocessed = dokumen.casefold() #casefolding
dokumen_prerocessed = re.sub(polaSimbol,'',dokumen_prerocessed) #punctuation removal
dokumen_prerocessed = dokumen_prerocessed.strip() #whitespace removal
dokumen_prerocessed = nltk.word_tokenize(dokumen_prerocessed) #word tokenization

print(dokumen_prerocessed)
```

```
['it', 'is', 'a', 'period', 'of', 'civil', 'war', 'rebel', 'spaceships',
'striking', 'from', 'a', 'hidden', 'base', 'have', 'won', 'their', 'first',
'victory', 'against', 'the', 'evil', 'galactic', 'empire', 'during', 'the',
'battle', 'rebel', 'spies', 'managed', 'to', 'steal', 'secret', 'plans', 'to',
'the', 'empires', 'ultimate', 'weapon', 'the', 'death', 'star', 'an', 'armored',
'space', 'station', 'with', 'enough', 'power', 'to', 'destroy', 'an', 'entire',
'planet', 'pursued', 'by', 'the', 'empires', 'sinister', 'agents', 'princess',
'leia', 'races', 'home', 'aboard', 'her', 'starship', 'custodian', 'of', 'the',
'stolen', 'plans', 'that', 'can', 'save', 'her', 'people', 'and', 'restore',
'freedom', 'to', 'the', 'galaxy']
```

1.3.2 Tagging

[13]:
```
#Menggunduh library yang dibutuhkan (biasanya cukup dilakukan 1x)
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\user\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger.zip.
```

[13]:
```
True
```

[29]:
```
#library dan fungsi untuk melakukan tagging
from nltk import pos_tag

hasil_tagging = pos_tag(dokumen_prerocessed)

print("----------------------")
print("| No     | token -> tag")
print("----------------------")
for index, item in enumerate(hasil_tagging):
    print("| {}\t| {} -> {} ({})".format(index+1,item[0],item[1],tagset_helper.
 ↪get(item[1]),"-"))
```

```
----------------------
| No     | token -> tag
----------------------
| 1      | it -> PRP (personal pronoun)
| 2      | is -> VBZ (verb 3sg pres)
| 3      | a -> DT (determiner)
| 4      | period -> NN (sing or mass noun)
| 5      | of -> IN (prepotition/subordin-conj)
| 6      | civil -> JJ (adjective)
| 7      | war -> NN (sing or mass noun)
| 8      | rebel -> NN (sing or mass noun)
| 9      | spaceships -> NNS (noun, plural)
| 10     | striking -> VBG (verb gerund)
| 11     | from -> IN (prepotition/subordin-conj)
| 12     | a -> DT (determiner)
| 13     | hidden -> JJ (adjective)
| 14     | base -> NN (sing or mass noun)
| 15     | have -> VBP (verb non-3sg present)
| 16     | won -> VBN (verb past part)
| 17     | their -> PRP$ (possess, pronoun)
| 18     | first -> JJ (adjective)
| 19     | victory -> NN (sing or mass noun)
| 20     | against -> IN (prepotition/subordin-conj)
| 21     | the -> DT (determiner)
| 22     | evil -> JJ (adjective)
| 23     | galactic -> JJ (adjective)
| 24     | empire -> NN (sing or mass noun)
| 25     | during -> IN (prepotition/subordin-conj)
| 26     | the -> DT (determiner)
| 27     | battle -> NN (sing or mass noun)
| 28     | rebel -> NN (sing or mass noun)
| 29     | spies -> NNS (noun, plural)
| 30     | managed -> VBD (verb past tense)
| 31     | to -> TO ("to")
| 32     | steal -> VB (verb base form)
| 33     | secret -> JJ (adjective)
| 34     | plans -> NNS (noun, plural)
| 35     | to -> TO ("to")
| 36     | the -> DT (determiner)
| 37     | empires -> NNS (noun, plural)
| 38     | ultimate -> JJ (adjective)
| 39     | weapon -> IN (prepotition/subordin-conj)
| 40     | the -> DT (determiner)
| 41     | death -> NN (sing or mass noun)
| 42     | star -> NN (sing or mass noun)
| 43     | an -> DT (determiner)
| 44     | armored -> JJ (adjective)
| 45     | space -> NN (sing or mass noun)
| 46     | station -> NN (sing or mass noun)
| 47     | with -> IN (prepotition/subordin-conj)
| 48     | enough -> JJ (adjective)
| 49     | power -> NN (sing or mass noun)
| 50     | to -> TO ("to")
| 51     | destroy -> VB (verb base form)
| 52     | an -> DT (determiner)
| 53     | entire -> JJ (adjective)
| 54     | planet -> NN (sing or mass noun)
| 55    | pursued -> VBN (verb past part)
| 56    | by -> IN (prepotition/subordin-conj)
| 57    | the -> DT (determiner)
| 58    | empires -> NNS (noun, plural)
| 59    | sinister -> VBP (verb non-3sg present)
| 60    | agents -> NNS (noun, plural)
| 61    | princess -> JJ (adjective)
| 62    | leia -> JJ (adjective)
| 63    | races -> NNS (noun, plural)
| 64    | home -> NN (sing or mass noun)
| 65    | aboard -> IN (prepotition/subordin-conj)
| 66    | her -> PRP$ (possess, pronoun)
| 67    | starship -> JJ (adjective)
| 68    | custodian -> NN (sing or mass noun)
| 69    | of -> IN (prepotition/subordin-conj)
| 70    | the -> DT (determiner)
| 71    | stolen -> VBN (verb past part)
| 72    | plans -> NNS (noun, plural)
| 73    | that -> WDT (wh-determ.)
| 74    | can -> MD (modal)
| 75    | save -> VB (verb base form)
| 76    | her -> PRP$ (possess, pronoun)
| 77    | people -> NNS (noun, plural)
| 78    | and -> CC (coordinating conjunction)
| 79    | restore -> NN (sing or mass noun)
| 80    | freedom -> NN (sing or mass noun)
| 81    | to -> TO ("to")
| 82    | the -> DT (determiner)
| 83    | galaxy -> NN (sing or mass noun)
```

1.4 Skenario Dokumen Tanpa Pra Pengolahan Teks

Pada skenario ini dokumen akan dilakukan *tagging* tanpa melalui pra pengolahan teks, kecuali tokenisasi.

1.4.1 Tokenisasi Dokumen

[32]:

```python
#Variabel dokumen_tanpa_prerocessed menyimpan hasil tokenisasi terhadap dokumen
dokumen_tanpa_prerocessed = nltk.word_tokenize(dokumen)
print(dokumen_tanpa_prerocessed)
```

```
['It', 'is', 'a', 'period', 'of', 'civil', 'war', '.', 'Rebel', 'spaceships',
',', 'striking', 'from', 'a', 'hidden', 'base', ',', 'have', 'won', 'their',
'first', 'victory', 'against', 'the', 'evil', 'Galactic', 'Empire', '.',
'During', 'the', 'battle', ',', 'Rebel', 'spies', 'managed', 'to', 'steal',
'secret', 'plans', 'to', 'the', 'Empire', '’', 's', 'ultimate', 'weapon', ',',
'the', 'DEATH', 'STAR', ',', 'an', 'armored', 'space', 'station', 'with',
'enough', 'power', 'to', 'destroy', 'an', 'entire', 'planet', '.', 'Pursued',
'by', 'the', 'Empire', '’', 's', 'sinister', 'agents', ',', 'Princess', 'Leia',
'races', 'home', 'aboard', 'her', 'starship', ',', 'custodian', 'of', 'the',
'stolen', 'plans', 'that', 'can', 'save', 'her', 'people', 'and', 'restore',
'freedom', 'to', 'the', 'galaxy', '…']
```

1.4.2 Tagging

[33]:

```
hasil_tagging = pos_tag(dokumen_tanpa_prerocessed)

print("----------------------")
print("| No     | token -> tag")
print("----------------------")
for index, item in enumerate(hasil_tagging):
    print("| {}\t| {} -> {} ({})".format(index+1,item[0],item[1],tagset_helper.
 ↪get(item[1]),"-"))
```

```
----------------------
| No     | token -> tag
----------------------
| 1      | It -> PRP (personal pronoun)
| 2      | is -> VBZ (verb 3sg pres)
| 3      | a -> DT (determiner)
| 4      | period -> NN (sing or mass noun)
| 5      | of -> IN (prepotition/subordin-conj)
| 6      | civil -> JJ (adjective)
| 7      | war -> NN (sing or mass noun)
| 8      | . -> . (sent-end punc)
| 9      | Rebel -> NNP (proper noun, sing.)
| 10     | spaceships -> NNS (noun, plural)
| 11     | , -> , (comma)
| 12     | striking -> VBG (verb gerund)
| 13     | from -> IN (prepotition/subordin-conj)
| 14     | a -> DT (determiner)
| 15     | hidden -> JJ (adjective)
| 16     | base -> NN (sing or mass noun)
| 17     | , -> , (comma)
| 18     | have -> VBP (verb non-3sg present)
| 19     | won -> VBN (verb past part)
| 20     | their -> PRP$ (possess, pronoun)
| 21     | first -> JJ (adjective)
| 22     | victory -> NN (sing or mass noun)
| 23     | against -> IN (prepotition/subordin-conj)
| 24     | the -> DT (determiner)
| 25     | evil -> JJ (adjective)
| 26     | Galactic -> NNP (proper noun, sing.)
| 27    | Empire -> NNP (proper noun, sing.)
| 28    | . -> . (sent-end punc)
| 29    | During -> IN (prepotition/subordin-conj)
| 30    | the -> DT (determiner)
| 31    | battle -> NN (sing or mass noun)
| 32    | , -> , (comma)
| 33    | Rebel -> NNP (proper noun, sing.)
| 34    | spies -> VBZ (verb 3sg pres)
| 35    | managed -> VBN (verb past part)
| 36    | to -> TO ("to")
| 37    | steal -> VB (verb base form)
| 38    | secret -> JJ (adjective)
| 39    | plans -> NNS (noun, plural)
| 40    | to -> TO ("to")
| 41    | the -> DT (determiner)
| 42    | Empire -> NNP (proper noun, sing.)
| 43    | ’ -> NNP (proper noun, sing.)
| 44    | s -> NN (sing or mass noun)
| 45    | ultimate -> JJ (adjective)
| 46    | weapon -> NN (sing or mass noun)
| 47    | , -> , (comma)
| 48    | the -> DT (determiner)
| 49    | DEATH -> NNP (proper noun, sing.)
| 50    | STAR -> NNP (proper noun, sing.)
| 51    | , -> , (comma)
| 52    | an -> DT (determiner)
| 53    | armored -> JJ (adjective)
| 54    | space -> NN (sing or mass noun)
| 55    | station -> NN (sing or mass noun)
| 56    | with -> IN (prepotition/subordin-conj)
| 57    | enough -> JJ (adjective)
| 58    | power -> NN (sing or mass noun)
| 59    | to -> TO ("to")
| 60    | destroy -> VB (verb base form)
| 61    | an -> DT (determiner)
| 62    | entire -> JJ (adjective)
| 63    | planet -> NN (sing or mass noun)
| 64    | . -> . (sent-end punc)
| 65    | Pursued -> VBN (verb past part)
| 66    | by -> IN (prepotition/subordin-conj)
| 67    | the -> DT (determiner)
| 68    | Empire -> NNP (proper noun, sing.)
| 69    | ’ -> NNP (proper noun, sing.)
| 70    | s -> VBP (verb non-3sg present)
| 71    | sinister -> NN (sing or mass noun)
| 72    | agents -> NNS (noun, plural)
| 73    | , -> , (comma)
| 74    | Princess -> NNP (proper noun, sing.)
| 75    | Leia -> NNP (proper noun, sing.)
| 76    | races -> VBZ (verb 3sg pres)
| 77    | home -> NN (sing or mass noun)
| 78    | aboard -> IN (prepotition/subordin-conj)
| 79    | her -> PRP$ (possess, pronoun)
| 80    | starship -> NN (sing or mass noun)
| 81    | , -> , (comma)
| 82    | custodian -> NN (sing or mass noun)
| 83    | of -> IN (prepotition/subordin-conj)
| 84    | the -> DT (determiner)
| 85    | stolen -> VBN (verb past part)
| 86    | plans -> NNS (noun, plural)
| 87    | that -> WDT (wh-determ.)
| 88    | can -> MD (modal)
| 89    | save -> VB (verb base form)
| 90    | her -> PRP$ (possess, pronoun)
| 91    | people -> NNS (noun, plural)
| 92    | and -> CC (coordinating conjunction)
| 93    | restore -> NN (sing or mass noun)
| 94    | freedom -> NN (sing or mass noun)
| 95    | to -> TO ("to")
| 96    | the -> DT (determiner)
| 97    | galaxy -> NN (sing or mass noun)
| 98    | … -> : (sent-mid punc)
```

[ ]: