pos-tagging

October 4, 2024

Nama : Rika Ajeng Finatih

NIM : 121450036

Kelas : Pemrosesan Bahasa Alami

1 POS Tagging

POS Tagging atau Part of Speech Tagging merupakan prosesmengidentifikasian dan penandaan kategori tata bahasa (part of speech) dari setiap kata dalam suatu kalimat. Hal ini adalah langkah penting dalam pengolahan bahasa alami yang memungkinkan komputer untuk memahami makna dan struktur kalimat dengan lebih baik.

1.1 Import Library

Langkah pertama dalam proses POS tagging adalah menginstal paket yang diperlukan untuk pemrosesan bahasa alami. Berikut adalah paket-paket yang akan diinstal.

[121]:
```
# Install Packages
!pip install python-crfsuite
!pip install nltk
!pip install spacy
```

```
Requirement already satisfied: python-crfsuite in
/usr/local/lib/python3.10/dist-packages (0.9.11)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages
(3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages
(from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages
(from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.10/dist-packages (from nltk) (2024.9.11)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages
(from nltk) (4.66.5)
Requirement already satisfied: spacy in /usr/local/lib/python3.10/dist-packages
(3.7.5)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.0.12)
```

```
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.3.0,>=8.2.2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (8.2.5)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (0.12.5)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (4.66.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.32.3)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.9.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from spacy) (3.1.4)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-
packages (from spacy) (71.0.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (24.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.4.1)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.10/dist-
packages (from spacy) (1.26.4)
Requirement already satisfied: language-data>=1.2 in
/usr/local/lib/python3.10/dist-packages (from langcodes<4.0.0,>=3.2.0->spacy)
(1.2.0)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.10/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.7.0)
Requirement already satisfied: pydantic-core==2.23.4 in
/usr/local/lib/python3.10/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (2.23.4)
Requirement already satisfied: typing-extensions>=4.6.1 in
/usr/local/lib/python3.10/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (4.12.2)
```

```
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests<3.0.0,>=2.13.0->spacy) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(2024.8.30)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2->spacy)
(0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2->spacy)
(0.1.5)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.10/dist-
packages (from typer<1.0.0,>=0.3.0->spacy) (8.1.7)
Requirement already satisfied: shellingham>=1.3.0 in
/usr/local/lib/python3.10/dist-packages (from typer<1.0.0,>=0.3.0->spacy)
(1.5.4)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.10/dist-
packages (from typer<1.0.0,>=0.3.0->spacy) (13.8.1)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
/usr/local/lib/python3.10/dist-packages (from weasel<0.5.0,>=0.1.0->spacy)
(0.19.0)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
/usr/local/lib/python3.10/dist-packages (from weasel<0.5.0,>=0.1.0->spacy)
(7.0.4)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->spacy) (2.1.5)
Requirement already satisfied: marisa-trie>=0.7.7 in
/usr/local/lib/python3.10/dist-packages (from language-
data>=1.2->langcodes<4.0.0,>=3.2.0->spacy) (1.2.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.10/dist-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (2.18.0)
Requirement already satisfied: wrapt in /usr/local/lib/python3.10/dist-packages
(from smart-open<8.0.0,>=5.2.1->weasel<0.5.0,>=0.1.0->spacy) (1.16.0)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-
packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy)
(0.1.2)
```

Penjelasan Packages:

- **python-crfsuite**: mengimplementasikan Conditional Random Fields (CRF), yang merupakan model statistik yang sering digunakan untuk tugas-tugas pengenalan pola, termasuk POS tagging.
- **nltk** (Natural Language Toolkit): pustaka Python yang menyediakan alat dan sumber daya untuk pemrosesan bahasa alami. Pustaka ini mencakup berbagai fungsi, seperti tokenisasi, pengenalan entitas, dan POS tagging. NLTK juga menyediakan berbagai dataset dan korpus bahasa yang berguna untuk pelatihan dan pengujian model NLP.
- **spacy**: pustaka NLP modern yang cepat dan efisien, dirancang untuk penggunaan dalam aplikasi dunia nyata. SpaCy mendukung berbagai tugas NLP, termasuk tokenisasi, pengenalan entitas, dan POS tagging, serta memiliki dukungan untuk beberapa bahasa. SpaCy juga mudah digunakan dan terintegrasi dengan model yang sudah dilatih sebelumnya.

Setelah packgaes berhasil di install, langkah seanjutnya dengan cara mengimport library yang akan digunakan dalam proses POS Tagging.

[122]:
```
# Import Library
import pandas as pd
from nltk.tokenize import wordpunct_tokenize
from nltk.tag import CRFTagger
import nltk
from nltk import pos_tag

# Download NLTK
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

[122]:
```
True
```

1.2 Load Data

Import data Harry Potter 2 yang akan dilakukan POS Tagging.

[123]:
```
# Memuat model bahasa Inggris dari SpaCy
nlp = spacy.load("en_core_web_sm")

# Import Data
df = pd.read_csv('Harry Potter 2.csv', delimiter=';')
df.head()
```

[123]:
```
  Character                                          Sentence
0    HARRY                       I can’t let you out, Hedwig.
1    HARRY   I’m not allowed to use magic outside of school.
2    HARRY                            Besides, if Uncle Vernon…
3    VERNON                                      Harry Potter!
4     HARRY                                Now you’ve done it.
```

1.3 Preprocessing

[124]:
```
def preprocessing(text):
    case_folded = text.lower()  # Mengubah ke huruf kecil
    tokenized = word_tokenize(case_folded)  # Tokenisasi
    return " ".join(tokenized)  # Mengembalikan teks yang ditokenisasi dalam␣
 ↪bentuk string
```

[125]:
```
# Menerapkan preprocessing pada kolom 'Sentence'
df['preprocessed_Sentence'] = df['Sentence'].apply(preprocessing)
print(df['preprocessed_Sentence'])
```

```
0                         i can ’ t let you out , hedwig .
1       i ’ m not allowed to use magic outside of scho…
2                              besides , if uncle vernon…
3                                        harry potter !
4                                  now you ’ ve done it .
                              …
1695                                     sorry i 'm late .
1696    the owl that delivered my release papers got a…
1697                         some ruddy bird called errol .
1698    and i 'd just like to say that if it had n't b…
1699          there 's no hogwarts without you , hagrid .
Name: preprocessed_Sentence, Length: 1700, dtype: object
```

Setelah berhasil melakukan preocessing pada teks dengan mengubahnya menjadi huruf kecil dan melakukan tokenisasi, langkah selanjutnya masuk kedalam proses tagging bagian of speech (POS) menggunakan pustakan seperti NLTK atau spaCy.

1.4 POS Tagging

Langkah akhir setelah melakukan preprocessing data, kita dapat menggunakan model SpaCy untuk melakukan POS tagging pada kolom yang berisi kalimat hasil dari proses preprocessing.

[126]:
```
# Fungsi untuk melakukan POS Tagging
def pos_tagging(text):
    doc = nlp(text)  # Menggunakan model SpaCy untuk analisis teks
    words = [token.text for token in doc]  # Mengambil kata
    pos_tags = [token.pos_ for token in doc]  # Mengambil POS
    tags = [token.tag_ for token in doc]  # Mengambil tag
    deps = [token.dep_ for token in doc]  # Mengambil dependency
```

```
    return words, pos_tags, tags, deps  # Mengembalikan keempat komponen

# Membuat list untuk menyimpan hasil POS tagging dalam bentuk tabel
data = []

# Menerapkan POS tagging pada kalimat yang telah diproses
for index, row in df.iterrows():
    character = row['Character']  # Nama karakter
    sentence = row['preprocessed_Sentence']  # Kalimat yang telah diproses
    words, pos_tags, tags, deps = pos_tagging(sentence)  # Mendapatkan hasil␣
 ↪POS tagging

    # Menyimpan hasil ke dalam list 'data'
    data.append([character, sentence, words, pos_tags, tags, deps])

# Membuat DataFrame dari hasil POS tagging
pos_df = pd.DataFrame(data, columns=['Character', 'Sentence', 'Word', 'POS',␣
 ↪'Tag', 'Dependency'])

# Menampilkan hasil dalam bentuk tabel
pos_df
```

[126]:

	Character	Sentence	\
0	HARRY	i can ’ t let you out , hedwig .	
1	HARRY	i ’ m not allowed to use magic outside of scho…	
2	HARRY	besides , if uncle vernon…	
3	VERNON	harry potter !	
4	HARRY	now you ’ ve done it .	
…	…	…	
1695	HAGRID	sorry i 'm late .	
1696	HAGRID	the owl that delivered my release papers got a…	
1697	HAGRID	some ruddy bird called errol .	
1698	HAGRID	and i 'd just like to say that if it had n't b…	
1699	HARRY	there 's no hogwarts without you , hagrid .	

	Word	\
0	[i, can, ’, t, let, you, out, ,, hedwig, .]	
1	[i, ’, m, not, allowed, to, use, magic, outsid…	
2	[besides, ,, if, uncle, vernon, …]	
3	[harry, potter, !]	
4	[now, you, ’, ve, done, it, .]	
…	…	
1695	[sorry, i, ', m, late, .]	
1696	[the, owl, that, delivered, my, release, paper…	
1697	[some, ruddy, bird, called, errol, .]	
1698	[and, i, 'd, just, like, to, say, that, if, it…	
1699	[there, 's, no, hogwarts, without, you, ,, hag…	

```
                                                    POS  \
0     [PRON, AUX, PUNCT, PROPN, VERB, PRON, ADP, PUN…
1     [PRON, PUNCT, AUX, PART, VERB, PART, VERB, NOU…
2               [ADV, PUNCT, SCONJ, PROPN, NOUN, PUNCT]
3                                  [PROPN, PROPN, PUNCT]
4             [ADV, PRON, PUNCT, AUX, VERB, PRON, PUNCT]
…                                                     …
1695               [INTJ, PRON, VERB, VERB, ADJ, PUNCT]
1696  [DET, NOUN, PRON, VERB, PRON, NOUN, NOUN, AUX,…
1697               [DET, NOUN, NOUN, VERB, NOUN, PUNCT]
1698  [CCONJ, PRON, AUX, ADV, VERB, PART, VERB, SCON…
1699  [PRON, VERB, DET, NOUN, ADP, PRON, PUNCT, VERB…

                                                    Tag  \
0             [PRP, MD, '', NNP, VB, PRP, RP, ,, NNP, .]
1     [PRP, '', VBP, RB, VBN, TO, VB, NN, RB, IN, NN…
2                              [RB, ,, IN, NNP, NN, NFP]
3                                          [NNP, NNP, .]
4                       [RB, PRP, '', VBP, VBN, PRP, .]
…                                                     …
1695                          [UH, PRP, VBP, VBP, JJ, .]
1696  [DT, NN, WDT, VBD, PRP$, NN, NNS, VBD, DT, VBN…
1697                            [DT, NN, NN, VBN, NN, .]
1698  [CC, PRP, MD, RB, VB, TO, VB, IN, IN, PRP, VBD…
1699               [EX, VBZ, DT, NNS, IN, PRP, ,, VB, .]

                                                Dependency
0     [nsubj, aux, punct, nsubj, ROOT, nsubj, ccomp,…
1     [nsubjpass, punct, auxpass, neg, ROOT, aux, xc…
2             [ROOT, punct, mark, compound, meta, punct]
3                                [compound, ROOT, punct]
4        [advmod, nsubj, punct, aux, ROOT, dobj, punct]
…                                                     …
1695           [intj, ROOT, appos, appos, acomp, punct]
1696  [det, nsubjpass, nsubj, relcl, poss, compound,…
1697          [det, compound, nsubj, ROOT, oprd, punct]
1698  [cc, nsubj, aux, advmod, ROOT, aux, xcomp, mar…
1699  [expl, ROOT, det, attr, prep, pobj, punct, dep…

[1700 rows x 6 columns]
```

Hasil yang diperoleh diatas menujukkan bahwa kita telah berhasil menerapkan POS tagging pada kalimat dari karakter dalam dataset “Harry Poter 2”. Kalimat telah berhasil di kelompokkan dalan setiap POS, tag, dan dependensy. Penjelasan lebih detailnya sebagai berikut.

- **POS:** Mencantumkan kategori POS untuk setiap token. Misalnya, untuk kata “I”, POS-nya adalah PRON (kata ganti), untuk “can” adalah AUX (kata kerja bantu), dan untuk “let”

adalah VERB. Ini memberikan informasi tentang jenis kata dan fungsinya dalam kalimat.

- **Tag:** Bentuk yang lebih detail dari POS. Misalnya, untuk token "I", tag-nya adalah PRP (personal pronoun). Ini memberikan detail tambahan yang berguna untuk analisis lebih lanjut.
- **Dependency:** hubungan sintaksis antara kata dalam kalimat. Misalnya, nsubj menunjukkan bahwa kata tersebut adalah subjek nominal dari kata kerja. Ini memberikan struktur gramatikal yang lebih mendalam mengenai bagaimana kata-kata saling berhubungan dalam kalimat.

POS tagging merupakan langkah penting dalam proses analisis bahasa alami dan membuka jalan bagi analisis yang lebih mendalam serta aplikasi yang lebih kompleks.

1.5 Simpan Hasil

[127]:
```
# Simpan hasil POS tagging ke dalam file CSV
pos_df.to_csv('hasil_pos_tagging.csv', index=False)
```