**NLP Lab 4**

Dixitha Kasturi

[dkasturi@syr.edu](mailto:dkasturi@syr.edu)

**Topic: Parts of Speech Tagging**

File chosen : Leaves by Whitman, file 18, index 17

The number of tokens = 711215

After splitting it into lines = 3827

List of tagged tokens : POS, Stanford

1. POS tagger :

[[('[', 'NN'), ('Leaves', 'NN'), ('of', 'IN'), ('Grass', 'NN'), ('by', 'IN'), ('Walt', 'NNP'), ('Whitman', 'NN'), ('1855', 'NN'), (']', 'NN'), ('Come', 'NN'), (',', ','), ('said', 'VBD'), ('my', 'PRP$'), ('soul', 'NN'), (',', ','), ('Such', 'JJ'), ('verses', 'NN'), ('for', 'IN'), ('my', 'PRP$'), ('Body', 'NN'), ('let', 'VBD'), ('us', 'PRP'), ('write', 'VB'), (',', ','), ('(', 'NN'), ('for', 'IN'), ('we', 'PRP'), ('are', 'VBP'), ('one', 'CD'), (',', ','), (')', 'NN'), ('That', 'DT'), ('should', 'MD'), ('I', 'PRP'), ('after', 'IN'), ('return', 'NN'), (',', ','), ('Or', 'CC'), (',', ','), ('long', 'JJ'), (',', ','), ('long', 'JJ'), ('hence', 'NN'), (',', ','), ('in', 'IN'), ('other', 'JJ'), ('spheres', 'NN'), (',', ','), ('There', 'EX'), ('to', 'TO'), ('some', 'DT'), ('group', 'NN'), ('of', 'IN'), ('mates', 'NN'), ('the', 'DT'), ('chants', 'NN'), ('resuming', 'NN'), (',', ','), ('(', 'NN'), ('Tallying', 'NN'), ('Earth', 'NN'), ("'s", 'POS'), ('soil', 'NN'), (',', ','), ('trees', 'NNS'), (',', ','), ('winds', 'NN'), (',', ','), ('tumultuous', 'JJ'), ('waves', 'NN'), (',', ','), (')', 'NN'), ('Ever', 'NN'), ('with', 'IN'), ('pleas', 'NN'), ("'d", 'MD'), ('smile', 'NN'), ('I', 'PRP'), ('may', 'MD'), ('keep', 'VB'), ('on', 'IN'), (',', ','), ('Ever', 'NN'), ('and', 'CC'), ('ever', 'RB'), ('yet', 'RB'), ('the', 'DT'), ('verses', 'NN'), ('owning', 'VBG'), ('--', ':'), ('as', 'IN'), (',', ','), ('first', 'JJ'), (',', ','), ('I', 'PRP'), ('here', 'RB'), ('and', 'CC'), ('now', 'RB'), ('Signing', 'NN'), ('for', 'IN'), ('Soul', 'NN'), ('and', 'CC'), ('Body', 'NN'), (',', ','), ('set', 'VBN'), ('to', 'TO'), ('them', 'PRP'), ('my', 'PRP$'), ('name', 'NN'), (',', ','), ('Walt', 'NNP'), ('Whitman', 'NN'), ('[', 'NN'), ('BOOK', 'NN'), ('I.', 'NNP'), ('INSCRIPTIONS', 'NN'), (']', 'NN'), ('}', 'NN'), ("One's-Self", 'NN'), ('I', 'PRP'), ('Sing', 'NN'), ("One's-self", 'NN'), ('I', 'PRP'), ('sing', 'NN'), (',', ','), ('a', 'DT'), ('simple', 'JJ'), ('separate', 'JJ'), ('person', 'NN'), (',', ','), ('Yet', 'CC'), ('utter', 'NN'), ('the', 'DT'), ('word', 'NN'), ('Democratic', 'JJ'), (',', ','), ('the', 'DT'), ('word', 'NN'), ('En-Masse', 'NN'), ('.', '.')], [('Of', 'IN'), ('physiology', 'NN'), ('from', 'IN'), ('top', 'JJ'), ('to', 'TO'), ('toe', 'NN'), ('I', 'PRP'), ('sing', 'NN'), (',', ','), ('Not', 'RB'), ('physiognomy', 'NN'), ('alone', 'RB'), ('nor', 'CC'), ('brain', 'NN'), ('alone', 'RB'), ('is', 'VBZ'), ('worthy', 'JJ'), ('for', 'IN'), ('the', 'DT'), ('Muse', 'NN'), (',', ','), ('I', 'PRP'), ('say', 'VBP'), ('the', 'DT'), ('Form', 'NN'), ('complete', 'VB'), ('is', 'VBZ'), ('worthier', 'NN'), ('far', 'RB'), (',', ','), ('The', 'NNP'), ('Female', 'NN'), ('equally', 'RB'), ('with', 'IN'), ('the', 'DT'), ('Male', 'NN'), ('I', 'PRP'), ('sing', 'NN'), ('.', '.')]]

1. Stanford tagger result:
2. [[('[', 'JJ'), ('Leaves', 'NNS'), ('of', 'IN'), ('Grass', 'NNP'), ('by', 'IN'), ('Walt', 'NNP'), ('Whitman', 'NNP'), ('1855', 'CD'), (']', 'NNP'), ('Come', 'NNP'), (',', ','), ('said', 'VBD'), ('my', 'PRP$'), ('soul', 'NN'), (',', ','), ('Such', 'JJ'), ('verses', 'NNS'), ('for', 'IN'), ('my', 'PRP$'), ('Body', 'NNP'), ('let', 'VB'), ('us', 'PRP'), ('write', 'VB'), (',', ','), ('(', '('), ('for', 'IN'), ('we', 'PRP'), ('are', 'VBP'), ('one', 'CD'), (',', ','), (')', ')'), ('That', 'WDT'), ('should', 'MD'), ('I', 'PRP'), ('after', 'IN'), ('return', 'NN'), (',', ','), ('Or', 'CC'), (',', ','), ('long', 'RB'), (',', ','), ('long', 'JJ'), ('hence', 'NN'), (',', ','), ('in', 'IN'), ('other', 'JJ'), ('spheres', 'NNS'), (',', ','), ('There', 'EX'), ('to', 'TO'), ('some', 'DT'), ('group', 'NN'), ('of', 'IN'), ('mates', 'VBZ'), ('the', 'DT'), ('chants', 'NNS'), ('resuming', 'NN'), (',', ','), ('(', '('), ('Tallying', 'VBG'), ('Earth', 'NNP'), ("'s", 'POS'), ('soil', 'NN'), (',', ','), ('trees', 'NNS'), (',', ','), ('winds', 'NNS'), (',', ','), ('tumultuous', 'JJ'), ('waves', 'NNS'), (',', ','), (')', ')'), ('Ever', 'RB'), ('with', 'IN'), ('pleas', 'NNS'), ("'d", 'MD'), ('smile', 'VB'), ('I', 'PRP'), ('may', 'MD'), ('keep', 'VB'), ('on', 'IN'), (',', ','), ('Ever', 'NNP'), ('and', 'CC'), ('ever', 'RB'), ('yet', 'RB'), ('the', 'DT'), ('verses', 'NNS'), ('owning', 'VBG'), ('--', ':'), ('as', 'IN'), (',', ','), ('first', 'RB'), (',', ','), ('I', 'PRP'), ('here', 'RB'), ('and', 'CC'), ('now', 'RB'), ('Signing', 'VBG'), ('for', 'IN'), ('Soul', 'NNP'), ('and', 'CC'), ('Body', 'NNP'), (',', ','), ('set', 'VBN'), ('to', 'TO'), ('them', 'PRP'), ('my', 'PRP$'), ('name', 'NN'), (',', ','), ('Walt', 'NNP'), ('Whitman', 'NNP'), ('[', 'NNP'), ('BOOK', 'NNP'), ('I.', 'NNP'), ('INSCRIPTIONS', 'NNP'), (']', 'NNP'), ('}', ')'), ("One's-Self", 'NNP'), ('I', 'PRP'), ('Sing', 'VBG'), ("One's-self", 'PRP'), ('I', 'PRP'), ('sing', 'VBP'), (',', ','), ('a', 'DT'), ('simple', 'JJ'), ('separate', 'JJ'), ('person', 'NN'), (',', ','), ('Yet', 'CC'), ('utter', 'JJ'), ('the', 'DT'), ('word', 'NN'), ('Democratic', 'NNP'), (',', ','), ('the', 'DT'), ('word', 'NN'), ('En-Masse', 'NNP'), ('.', '.')], [('Of', 'IN'), ('physiology', 'NN'), ('from', 'IN'), ('top', 'JJ'), ('to', 'TO'), ('toe', 'VB'), ('I', 'PRP'), ('sing', 'VBG'), (',', ','), ('Not', 'RB'), ('physiognomy', 'VB'), ('alone', 'RB'), ('nor', 'CC'), ('brain', 'NN'), ('alone', 'RB'), ('is', 'VBZ'), ('worthy', 'JJ'), ('for', 'IN'), ('the', 'DT'), ('Muse', 'NNP'), (',', ','), ('I', 'PRP'), ('say', 'VBP'), ('the', 'DT'), ('Form', 'NNP'), ('complete', 'NN'), ('is', 'VBZ'), ('worthier', 'JJR'), ('far', 'RB'), (',', ','), ('The', 'DT'), ('Female', 'NNP'), ('equally', 'RB'), ('with', 'IN'), ('the', 'DT'), ('Male', 'NNP'), ('I', 'PRP'), ('sing', 'VBP'), ('.', '.')]]

Frequencies :

1. Frequencies of parts of speech tags are

NN 43648

, 17936

IN 14385

DT 13612

PRP 8423

CC 7152

JJ 5491

RB 4797

NNS 3585

. 3559

VB 3317

VBP 2818

PRP$ 2604

MD 2554

NNP 2302

TO 2151

VBZ 2054

: 1359

VBG 1109

VBD 1084

VBN 886

WP 684

CD 636

WRB 619

POS 590

WDT 449

RP 299

JJR 237

RBR 196

EX 171

JJS 145

LS 114

NNPS 94

'' 72

RBS 33

WP$ 22

`` 9

PDT 3

-NONE- 1

FW 1

b)Treebank:

N 49629

, 17936

I 14385

D 13612

P 11620

V 11268

C 7788

J 5873

R 5325

. 3559

M 2554

T 2151

W 1774

: 1359

E 171

L 114

' 72

` 9

- 1

F 1

Reporting :

The POS tags for the text in Whitman-leaves.txt is lesser when compared to the treebank. This tells us that the parts of speech tagged in ‘leaves’ is less compared to treebank text. POS used in leaves despite being diverse are not as rich as the ones in treebank.