K. J. SOMAIYA COLLEGE OF ENGINEERING

(Autonomous College Affiliated to University of Mumbai)

Candidate Roll No. _____ (In figures)

Test Exam.

Name : _____

Date : _____ 20

Examination : _____ Branch/Semester _____

Subject : _____

Junior Supervisor's full Signature with Date

Question No.	1	2	3	4	5	6	7	8	9	10	11	12	Total
Marks Obtained													

Stemming with Porter Stemmer Algorithm

→ Suffix Stripping.
↳ It is operation by which suffixes are removed from terms.

→ A document is represented by vector of words, or terms.

→ Terms with a common stem:-
- Connect
- connected
- Connecting
- Connection
- connections.

⟹ Connect (Root word)

→ Stem is nothing but root-

* Stemming
→ In linguistic morphology and information retrieval, stemming is process of reducing terms to their "roots" before indexing.

→ Stemming Programs are commonly referred to as stemming algorithms or stemmers.

→ "Stemming" suggest- crude affix chopping
↳ it is language dependent-

E.g. automate(s) automatic, automation all reduced to automat

i.e automat is actual word, from which all other words are generated.

∴ For Stemming one very Popular algorithm is Porter Stemmer

→ Common algorithm for Stemming English
→ The algorithm consists of 5 steps of rules applied in order i.e Sequentially.
→ Each Phase consists of Set of Commands.
→ Sample convention :- of the rules in a Compound command, Select the one that applies to longest Suffix

The Porter Stemmer :- definitions.

Definitions :-
→ CONSONANT :- a letter other than A, E, I, O, U and Y Preceded by consonant.
→ VOWEL :- any other letter. (i.e A, E, I, O and U)

∴ with this definition in Porter Stemming algorithm, all the words are of the form.

$$(C)(VC)^m(V)$$

C = String of one or more consonants. (cont).
V = String of one or more vowels.
m = measure of word or word Part, when represented in this form (VC).

imp

* Example for m (Measure of Word)

For eg: 1) M=0 TREE, BY, TR.
i.e Vowel consonant combination.

TREE ∴ no combination of Vc.
∴ M=0.

2) M=1 ; TROUBLE OATS TREES, IVY.

∴ one combination ∴ M=1.

3) M=2, TROUBLES PRIVATE, OATEN, ORRERY.

∴ two combination in one word ∴ M=2

* Rule for removing a suffix :-
The rules are of the form:-

(condition) S1 → S2 → Porter Stemmer applies rules in this form.

i.e condition followed by suffix S1 maps to suffix S2.

where S1 and S2 are suffixes

This means that if a word ends with a suffix S1, and the stem before S1 satisfies the given condition then S1 is replaced by S2.

→ E.g if rule is (m>1) EMENT →

→ In this s_1 is EMENT and s_2 is NULL.

→ So this would map REPLACEMENT to REPLAC

M=4 i.e M>1 satisfied.

∴ EMENT is replaced by Null value

* The condition part may also contain one of following :-

m	The measure of the stem
*s	The stem ends with s
v	The stem contains a vowel
*d	The stem ends with double consonant (TT, SS)
*o	The stem ends in cvc (second c should not be w, x or y) e.g. WIL, HOP

→ The condition part may also contain expressions with and, or and not -

→ e.g. (m>1 and (*s or *t)) : tests for a stem with m>1 ending in s or t.

1/18/20,000

K. J. SOMAIYA COLLEGE OF ENGINEERING

(Autonomous College Affiliated to University of Mumbai)

(6)

Candidate Roll No. _____ (In figures)

Test Exam.

Name : _____

Date : _____ 20

Examination : _____ Branch/Semester _____

Subject : _____

Junior Supervisor's full Signature with Date

Question No.	1	2	3	4	5	6	7	8	9	10	11	12	Total
Marks Obtained													

→ The Porter Stemmer : Step 1a (Step 1, Part-a).

i) SSES → SS

i.e. after applying Porter Stemmer.

Caresses → caress

ii) IES → I

↳ Ponies → Poni

↳ ties → ti

iii) SS → SS

↳ caress → caress

iv) S → ε

↳ null.

↳ cats → cat.

→ Step 1b { with rules condition }

i) (m>0) EED → EE.

- Condition Verified : agreed → agree.

i.e. agreed

↓ ↓

v.c. except.

i.e only for Stem we have m=1 i.e m>0.

∴ EED is replaced by EE.

• Condition not Verified : ~~bb~~ feed → feed.
No combin" of Vc.
∴ no change.

ii) (*V*) ED → ε
↳ NULL

• Condition Verified : plastered → plastery
V V.
∴ ED applied by Null.

• Condition not verified bled → bled
No vowel.

iii) (*V*)ING → ε

• Condition Verified :- motoring → motor.
• Condition not Verified : sing → sing.

Next Step (cleanup)
These rules are applied if second or third rule in 1b apply).

i) AT → ATE

↳ conflat(ed) → conflate.
ii) BL → BLE ↳ convert into proper word.
↳ Troubl(ing) → trouble.

iii) (*d & !(*L or *S or *Z)) → Single letter
i.e stem with double consonant and that consonant is ~~ents~~ not L, S or Z. then converted into single letter.

- Condition verified : hopp(ing) → hop
  i.e double consonant - then double pp replaced by single P.
  tann(ed) → tan.
- Condition Not verified fall(ing) → fall.

iv) (m=1 & *o)
↳ Measure =1 and ∝o i.e cvc form.
- Condition verified : fil(ing) → file.
  c v v c
- Condition not verified : fail → fail.
  v v c
  ∴ condition not satisfied.

~~4.~~ The Porter Stemmer : Step 1c and 2

i) Step 1c : Y Elimination (*V*) Y → I.
- Condition Verified : happy → happi
  if we chop y, then it contains a vowel.
  ∴ y is replaced by i.
- Condition not verified : sky → sky.

~~iii)~~ Step 2 Derivational Morphology I.
i) . (m>0) ATIONAL → ATE
↳ Relational → relate.
ii) (m>0) IZATION → IZE
↳ generalization → generalize.
iii) (m>0) BILITI → BLE
↳ sensibiliti → sensible.

Step 3 Derivational Morphology, II.

i) (m>0) ICATE → IC
↳ triplicate → triplic.

ii) (m>0) FUL → ε
↳ hopeful → hope.

iii) (m>0) NESS → ε
↳ goodness → good.

Step 4 Derivational Morphology, III.

i) (m>0) ANCE → ε
↳ allowance → allow.

ii) (m>0) ENT → ε
↳ dependent → depend

iii) (m>0) IVE → ε
↳ effective → effect.

Step 5 (cleanUP).

Step 5a

i) (m>1) E → ε
↳ Probate → Probat.

ii) (m=1 & !*o) NESS → ε
↳ goodness → good.

Step 5b

(m>1 & *d & *L) → single letter;
↳ Condition Verified ∴ control → control
↳ Condition Not Verified roll → roll.

∴ This is how Porter Stemmer works on tokens and gives the roots words.