**Etichetarea părţii de vorbire**

**Bărbulescu Adrian, ISM 244/1**

Se adauga mai tarziu aici ..

Font – Times New Roman

Chapter Heading – 24 + bold

Subchapter heading – 18 + bold

Sub titles – 16 + bold

Main Body Text – 12

Bibliography text - 12

Links - 12

Captions – 10

C# code - 10

**III. Consideratii practice**

1. **Descriere generala proiect – Modelul de predictie**

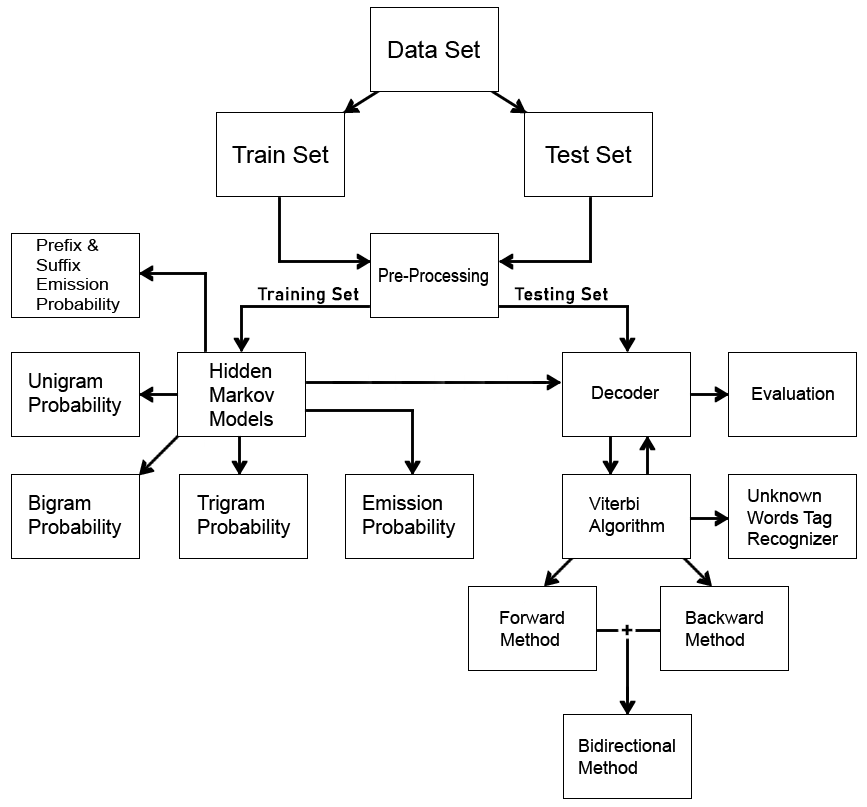


Figura 1 - Diagrama bloc funcţional

**2. Descriere separata bloc**

* 1. **Data Set**

Pentru setul de date se va folosi **Brown Corpus**, o colectie de propozitii si fraze in limba engleza colectate si organizate de W. Nelson Francis & Henry Kucera din departamentul lingvistic de la Universitatea Brown. Colectia aceasta are peste 1 milion de cuvinte (~1,014,312) in total si exact 500 de documente. Cele 500 de documente sunt impartite in 2 mari categorii, prima categorie fiind proza informativa cu urmatoarele sub-categorii:

A.Presă: Reportaje – 44 documente

B.Presă: Editorial – 27 documente

C.Presă: Recenzii (teatru, carti, muzica, dans) – 17 documente

D.Religie – 17 documente

E.Skill-uri si hobby-uri – 36 documente

F.Folclor popular – 48 documente

G.Scrisori, bibliografii, biografii – 75 documente

H.Diverse – 30 documente

J.Articole stiintifice– 80 documente

TOTAL – 374 documente

iar a doua categorie fiind proza imaginativa cu urmatoarele sub-categorii:

K.Fictiune generala – 29 documente

L.Mister și ficțiune detectiva – 24 documente

M.Opere științifico-fantastice – 6 documente

N.Aventura și ficțiunea western – 29 documente

P.Povesti romantice și de dragoste – 29 documente

R.Umor – 9 documente

TOTAL – 126 documente

Fiecare document are peste 2000 de cuvinte iar pentru fiecare cuvant se adauga si eticheta cu partea de vorbire aferenta acestuia, sub forma “cuvant/tag”. Acest set de date este folositor pentru un algoritm de *invatare supervizata pe o clasa* sau *one class classification*, deoarece fiecare cuvant are un singur tag indicat de un supraveghetor, astfel acest proces fiind unul supervizat.

ex. *Mr./np Remarque's/np$ conception/nn of/in this/dt novel/nn was/bedz sound/jj and/cc perhaps/rb even/rb noble/jj ./.*

Exista peste 100 de taguri (~103) individuale in Brown Corpus, multe fiind derivate de la forma de baza a partii de vorbire, de exemplu tagul *np$* este substantiv propriu-zis posesiv la singular, fiind derivat din substantiv. Unele taguri pot aparea combinate cu delimitatorul ‘+’:

ex. … Y'all/ppss ***wanna/vb+to*** walk/vb …

, deoarece *want/vb* + *to/to 🡺 wanna/vb+to.*

Alte taguri pot aparea formate din tagul propriu-zis si un tag prefix de indicatie a unei informatii suplimentare folosind delimitatorul ‘-‘, ca fie este un cuvant preluat din alta limba ‘FW’ (foreign word), fie ca acel cuvant apare in titlu ‘TL’ sau alte informatii de forma aceasta care ar putea fi importante la clasificare.

ex. … yesterday/nr ***en/fw-in route/fw-nn*** to/in his/pp$ …

*, unde /fw-in* este tagul pentru prepozitie intr-o limba straina si */fw-nn* estesubstantiv la singular intr-o limba straina.

In unele tag-uri pot aparea si simboluri precum ‘$’, ‘\*’, etc., mai multe exemple si informatii despre tag-uri pot fi gasite in manualul pentru Brown Corpus.

* 1. **Train & Test Set**

Inainte ca modelul de predictie sa fie folosit pe date reale acesta mai intai este evaluat. Pentru a realiza evaluarea modelului, trebuie sa impartim setul de date intr-un set de antrenament (train set) si un set de testare (test set). Exista 2 metode de impartire a setului de date din proiect:

**2.2.1. Impartire 70% - setul de antrenament, 30% - setul de testare**

Aceasta metoda este oarecum evidenta, pentru fiecare sub-categorie din Brown Corpus, se aleg din documentele aferente acesteia, 70% documente pentru antrenare si 30% documente pentru testare.

ex. Pentru sub-categoria *J. Articole stiintifice,* avem in total 80 de documente, primele 56 documente (70%) le vom folosi pentru antrenare iar ultimele 24 documente (30%) le vom folosi pentru testare.

**2.2.2. Cross-Validation**

Numita si “rotation estimation”, este o tehnica de validare a modelului, pentru a vedea rezultatul generalizat al modelului pentru un set de date independent. Acest mod de validare este folositor pentru a vedea abilitatea modelului la predictia pentru date noi, inlaturand probleme precum “overfitting” sau “selection-bias”, probleme care presupun impartirea setului de date astfel incat pentru o portiune aleasa din setul de antrenament ar da predictii foarte mici daca ar fi considerata set de testare sau de validare.

**2.2.3. K folds cross-validation**

Acesta presupune mai intai alegerea unui numar K (deobicei K=4 sau K=10) si apoi impartirea setului de date in K folds. Dupa aceasta impartire, se va itera setul de date astfel incat fiecare fold individual sa fie set de testare si restul set de antrenare. Dupa fiecare evaluare a setului de testare se va salva acuratetea modelului pt. acel fold si se va pastra pana cand vom evalua toate fold-urile.



Figura 2 – exemplu cross-validation k = 4, sursa: <https://blog.quantinsti.com/cross-validation-machine-learning-trading-models/>

La sfarsit dupa ce am evaluat fiecare fold in parte, se va calcula media aritmetica pentru fiecare fold pentru a obtine o acuratete generala pe model.

k - numarul de fold-uri alese

– acuratetea pe fold-ul i

**2.2.4. Implementare K folds cross-validation**

Functia principala pentru cross-validation are 3 parametrii de intrare: *filePath* pentru path-ul la folder-ul cu documentele Brown Corpus, *fold* pentru numarul de fold-uri folosite pentru impartire a fisierelor (parametru default = 4) si shuffle pentru optiunea booleana daca a amesteca documentele sau a le lasa cum apar in director (parametru default = False)

List**<**string**>** files **=** FileReader**.**GetAllTextFromDirectoryAsList**(**filePath**);**

int filesPerFold **=** files**.**Count **/** fold**;**

**this.**TestFile **=** **new** string**[**fold**];**

**this.**TrainFile **=** **new** string**[**fold**];**

**if** **(**shuffle**)**

files **=** **this.**Shuffle**(**files**);**

**for** **(**int crossIndex **=** 0**;** crossIndex **<** fold**;** crossIndex**++)**

**{**

var IndividualTrainFiles **=** **new** List**<**string**>();**

var IndividualTestFiles **=** **new** List**<**string**>();**

**for** **(**int i **=** 0**;** i **<** files**.**Count**;** i**++)**

**{**

**if** **(**i **>=** **(**filesPerFold **\*** crossIndex**)** **&&** i **<** **(**filesPerFold **\*** **(**crossIndex **+** 1**)))**

**{**

IndividualTestFiles**.**Add**(**files**[**i**]);**

**}**

**else**

**{**

IndividualTrainFiles**.**Add**(**files**[**i**]);**

**}**

**}**

string trainf **=** String**.**Join**(**" "**,** IndividualTrainFiles**);**

string testf **=** String**.**Join**(**" "**,** IndividualTestFiles**);**

**this.**TrainFile**[**crossIndex**]** **=** trainf**;**

**this.**TestFile**[**crossIndex**]** **=** testf**;**

**}**

In functia aceasta, se vor citi toate documentele din folderul cu fisierele Brown si se vor returna textul acestora ca lista pentru fiecare document separat (files.count = 500).

In continuare se vor initia vectorii proprietati ai clasei pentru fisierele de antrenare si de testare, se va calcula cate fisiere sunt per fold (de exemplu pentru fold = 4, vor exista 125 documente per fold) si dupa aceea se va verifica daca optiunea de shuffle a fost “activata”, daca da atunci lista de fisiere cu documente Brown va fi amestecata.

In for-ul din functia principala, se va itera de la 0 pana la fold number ales, se va calcula lista individuala de fisiere pt. antrenare si testare si dupa care din aceste liste se va concatena tot continutul acestora prin simbolul char de spatiu si rezultatul se va salva la crossIndex-ul fold-ului in vectorul propietate de antrenare si de testare.

Pentru a intelege mai bine conditia colorata, luam ca exemplu filesPerFold = 125, crossIndex = 3 (adica ultimul fold), atunci aceasta conditie s-ar traduce => i(index-ul pentru document) mai mare sau egal ca 375 si i mai mic strict ca 500, ceea ce reprezinta 25% pentru setul de testare salvat in lista *IndividualTestFiles* iar celelalte fisiere de la 0 la 374 adica restul de 75% se vor salva in lista de antrenare *IndividualTrainFiles*.

Metoda statica GetAllTextFromDirectoryAsList(path) din clasa statica FileReader este implementata in felul urmator:

List**<**string**>** outputFile **=** **new** List**<**string**>();**

var files **=** Directory**.**GetFiles**(**inputDir**);**

**foreach** **(**string file **in** files**)**

**{**

string elem **=** GetTextFromFileAsString**(**file**);**

outputFile**.**Add**(**elem**);**

**}**

**return** outputFile**;**

GetTextFromFileAsString este tot o metoda statica care returneaza tot textul dintr-un fisier ca string (**File.ReadAllText(inputFile);**).

Functia de amestecare implementeaza algoritmul *Fisher–Yates Shuffling,* acesta alege un numar random intre index-ul de la primu element din lista si ultimul element din lista si incepe sa numere pana la acesta, ignorand elementele din lista deja folosite in amestecare (scratched elements), odata ajuns la index-ul respectiv, va prelua elementul de acolo si il va adauga la finalul listei nou formate, punand astfel un flag de folosit pt. indexul respectiv si decrementand numarul total de numere.

ex.Fisher–Yates: lista cu numere: [1, 2, 3, 4, 5], n = 5

alegem un numar random de la 1 la n (1 fiind index-ul la primul element din lista, n fiind index-ul la ultimul element din lista)

Random = 2 🡺 noua\_lista = [2], lista\_veche = [1, **2**, 3, 4, 5], n =4

Random = 3 🡺 noua\_lista = [2, 4], lista\_veche = [1, **2**, 3, **~~4~~**, 5], n =3

Random = 1 🡺 noua\_lista = [2, 4, 1], lista\_veche = [**1**, **2**, 3, **4**, 5], n =2

Random = 2 🡺 noua\_lista = [2, 4, 1, 5], lista\_veche = [**1**, **2**, 3, **4**, **5**], n =1

noua\_lista\_finala = [2, 4, 1, 5, 3]

Implementarea acestuia este destul de ineficienta si complicata, pentru algoritm ar trebuii sa folosim 2 liste si sa avem si un flag pentru indexul elementelor pe care le-am folosit deja, dar exista totusi o metoda mai usoara de implementare, anume *Durstenfeld's version.* Acesta presupune interschimbarea elementului la index-ul n cu elementul la index-ul ales random si decrementarea lui n dupa fiecare interschimbare.

ex. Durstenfeld’s version: lista: [1, 2, 3, 4, 5], n = 5

Random = 2 🡺 dfeld\_list = [1, **5**, 3, 4 | **2**], n = 4

Random = 3 🡺 dfeld\_list = [1, **5**, **4** | **3**, **2**], n = 3

Random = 3 🡺 dfeld\_list = [1, **5** | **4**, **3**, **2**], n = 2

Random = 1 🡺 dfeld\_list = [**5** | **1**, **4**, **3**, **2**], n = 1

dfeld\_list\_final = [5, 1, 4, 3, 2]

Functia pentru implementarea versiunii lui Durstenfeld, cu un numar random generat de metoda Next din clasa fundamentala Random din c#:

Random rng **=** **new** Random**();**

int n **=** list**.**Count**;**

**while** **(**n **>** 1**)**

**{**

n**--;**

int k **=** rng**.**Next**(**n **+** 1**);**

string value **=** list**[**k**];**

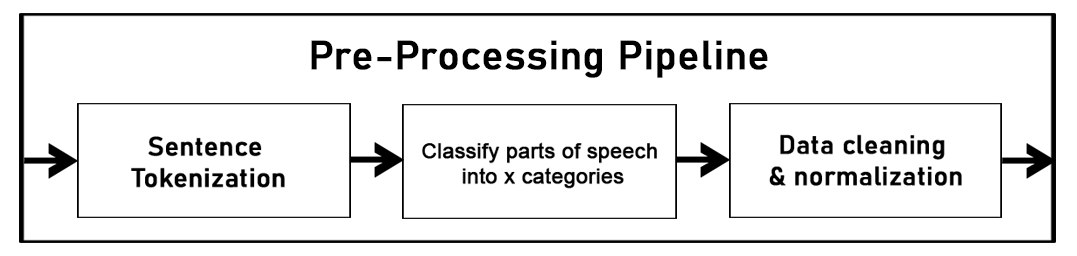
list**[**k**]** **=** list**[**n**];**

list**[**n**]** **=** value**;**

**}**

**return** list**;**

* 1. **Pre-processing**

****

**2.3.1. Sentence Tokenization**

Tokenization este procesul de a delimita cuvintele dintr-un text si posibil a le clasifica, folosit in analiza lexicala. Acesta deobicei delimiteaza pe baza unei reguli, algoritmul folosit pentru model este *Whitespace Tokenizer* dar se pot implementa mai multi algoritmi de tokenizare pentru model, pentru cazul in care modelul va fi aplicat pe date si texte din lumea reala.

Algoritmul *Whitespace Tokenizer* imparte un string intr-o singura lista de string-uri, si fiecare string este delimitat atunci cand gaseste caracterul de spatiu, tab sau newline (coduri ASCII: 32, 09, 13).

ex. I’m home but he is not home.

Token\_list = [“I’m”, “home”, “but”, “he”, “is”, “not”, “home.”]

**2.3.2. Implementare Sentence Tokenization**

Functia whitespace tokenizer echivalenta cu functia Split din c#:

List**<**string**>** tokenizedText **=** **new** List**<**string**>();**

string word **=** ""**;**

**foreach(**char c **in** Text**)**

**{**

**if** **(!**Char**.**IsWhiteSpace**(**c**))**

word **+=** c**;**

**else** **if** **(!**String**.**IsNullOrEmpty**(**word**))**

**{**

tokenizedText**.**Add**(**word**);**

word **=** ""**;**

**}**

**}**

**if** **(!**String**.**IsNullOrEmpty**(**word**))**

tokenizedText**.**Add**(**word**);**

**return** tokenizedText**;**

Acest tokenizer este cel mai potrivit deoarece, in setul de date, fiecare cuvant+tag este delimitat de spatiu, deci lista de token-uri va arata de forma: list = […., “he/pn”, “made/vbd”, “pancakes/nn”, “./.” , ….].

Totusi se prefera despartirea cuvantului de tag, pentru asta se implementeaza urmatoarea structura:

**public** struct WordTag

**{**

**public** string word**;**

**public** string tag**;**

**public** WordTag**(**string word**,** string tag**)**

**{**

**this.**word **=** word**;**

**this.**tag **=** tag**;**

**}**

**}**

, iar functia de despartire si adaugare in lista:

List**<**WordTag**>** wordTags **=** **new** List**<**WordTag**>();**

**foreach** **(**var word **in** Words**)**

**{**

string**[]** separated **=** word**.**Split**(**'/'**);**

string separatedTag **=** separated**[**separated**.**Length **-** 1**];**

string combWord **=** ""**;**

**if** **(**separated**.**Length **>** 2**)**

**{**

**for** **(**int i **=** 0**;** i **<** separated**.**Length **-** 1**;** i**++)**

combWord **+=** separated**[**i**]** **+** "/"**;**

combWord **=** combWord**.**Remove**(**combWord**.**Length **-** 1**);**

**}**

**else** combWord **=** separated**[**0**];**

wordTags**.**Add**(new** WordTag**(**combWord**,** separatedTag**));**

**}**

**return** wordTags**;**

Separam cuvantul de tag prin functia split cu parametrul ‘/’ si obtinem un vector de string-uri. Ultimul string din vector va fi mereu tag-ul cuvantului dar pentru restul vectorului nu exista aceeasi siguranta, de exemplu cuvintele compuse precum “input/output” pot aparea in setul de date si daca acestea nu sunt salvate integral, pot aduce rezultate gresite in procesul de invatare. Se cunoaste ca in cazul cel mai bun vom avea un vector de string-uri de dimensiune = 2 (cuvant necompus+tag) iar in cazul opus, in sectiunea colorata de mai sus, se itereaza vectorul de cuvinte si se concateneaza tot ce apare in vector pana la tag, dupa care se sterge ultimul caracter(care va fi ‘/’) din string-ul *combWord* obtinut, deoarece dupa fiecare cuvant se adauga caracterul ‘/’. La sfarsit se va adauga in lista de structuri, string-ul *combWord* care contine cuvantul compus/necompus si string-ul *separatedTag* care contine tag-ul cuvantului.

**2.3.3. Classify parts of speech into x categories**

Pentru o mai buna predictie, se alege clasificarea partilor de vorbire in x categorii, acestea fiind partile de vorbire de baza pentru a acoperi majoritatea tag-urilor din Brown Corpus. Acest proces nu este unul automat facut de un algoritm, ci este facut in urma unei analize a partilor de vorbire din engleza si alese cele mai potrivite pentru setul de date cu care se lucreaza (Brown Corpus). Pentru x = 10, am ales urmatoarele categorii:

1. Noun (NN) – substantive proprii,
2. Pronoun (PN) - pronume
3. Verb (VB) – verb
4. Adjective (JJ) – adjectiv
5. Adverb (RB) – adverb
6. Preposition (PP) – prepozitie
7. Conjuction (CC) – conjunctie
8. Article/Determiner (AT/DT) – articol & determinant (cel din engleza)
9. End of sentence (.) – sfarsit de propozitie, indiferent de semne ‘!’, ‘?’, “.” etc.
10. Others (OT) – altele precum interjectie, cardinal numeral precum “six”, “two” etc.

Deci in continuare, orice cuvant/string din lista de antrenare sau de testare va avea doar un singur tag din cele 10 mentionate anterior.

**2.3.4. Implementare Classify parts of speech into x categories**

Procedeul principal este de a itera vechea lista cu cuvinte si tag-uri si de a creea/schimba lista in care vechile tag-uri vor face parte dintr-una din cele 10 categorii.

**foreach** **(**var w **in** Words**)**

**{**

int tagIndex **=** GetTagIndexForConversion**(**w**);**

string newTag **=** ConvertBrownTagToHierarchicTag**(**tagIndex**);**

newWords**.**Add**(new** Tokenizer**.**WordTag**(**w**.**word**,** newTag**));**

**}**

Metoda GetTagIndexForConversion, implementare:

int tagIndex **=** **-**1**;**

**for** **(**int i **=** 0**;** i **<** BrownCorpusTags**.**Count**;** i**++)**

**{**

string**[]** splittedTag **=** Word**.**tag**.**Split**(new** Char**[]** **{** '+'**,** '-' **});**

**foreach** **(**string w **in** splittedTag**)**

**{**

**if** **(**Word**.**tag**.**Equals**(**"wql"**)** **||** Word**.**tag**.**Equals**(**"wql-tl"**))**

**{**

tagIndex **=** 55**;**

**return** tagIndex**;**

**}**

**else** **if** **(**Word**.**tag**.**Contains**(**BrownCorpusTags**[**i**]))**

**{**

tagIndex **=** i**;**

**return** tagIndex**;**

**}**

**}**

**}**

**return** tagIndex**;**

Aceasta returneaza un index in functie de unde se afla tag-ul in lista BrownCorpusTags, mai intai va desparti tag-urile compuse si va itera fiecare tag in parte pentru a ii atribuii un index in lista. Se va verifica daca exista tag-ul iterat in lista BrownCorpusTags, daca exista se va returna index-ul din lista, daca nu, se va returna -1. Metoda cauta tag-ul cu functia Contains deoarece unele tag-uri precum “nps$” apar si cu simboluri dar apartin tot in aceeasi categorie unde “nps” catalogat.

Lista BrownCorpusTags contine urmatoarele tag-uri:

"nn"**,** "nns"**,** "nns$"**,** "np"**,** "np$"**,** "nps"**,** "nps$"**,** "nr"**,** "nrs"**,**

"pn"**,** "pn$"**,** "pp$"**,** "pp$$"**,** "ppl"**,** "ppls"**,** "ppo"**,** "pps"**,** "ppss"**,** "wp$"**,** "wpo"**,** "wps"**,**

"vb"**,** "vbd"**,** "vbg"**,** "vbn"**,** "vbz"**,** "bem"**,** "ber"**,** "bez"**,** "bed"**,** "bedz"**,** "ben"**,** "do"**,** "dod"**,** "doz"**,** "hv"**,**

"hvd"**,** "hvg"**,** "hvn"**,** "hvz"**,** "md"**,**

"jj"**,** "jjr"**,** "jjs"**,** "jjt"**,**

"rb"**,** "rbr"**,** "rbt"**,** "rn"**,** "rp"**,** "wrb"**,** "ql"**,** "qlp"**,**

"in"**,** "to"**,**

"cc"**,** "cs"**,**

"at"**,** "ap"**,** "abl"**,** "abn"**,** "abx"**,** "dt"**,** "dti"**,** "dts"**,** "dtx"**,** "be"**,** "beg"**,** "ex"**,** "wdt"**,**

"."

, tag-urile din aceasta lista sunt ordonate in functie de categoriile descrise anterior. Conditia colorata din functie este un caz special deoarece gaseste secventa cu pattern-ul “ql” inainte de a gasi “wql” in lista(“ql” l-am catalogat ca adverb iar daca va cauta cu functia Contains atunci “wql” va fi catalogat ca adverb cand ar trebuii sa fie conjunctie) si de accea se verifica mai intai daca daca tag-ul este egal cu “wql” sau “wql-tl” (aici nu apar tag-uri compuse cu ‘+’ sau cu tag-uri “fw”, “hl” etc.). Tag-urile care nu apar in lista vor avea index = -1, si vor fi catalogate ca tag de Others (altele).

Metoda ConvertBrownTagToHierarchicTag, implementare:

string tag **=** "Tag NOT found! Something went wrong!"**;**

**if** **(**tagIndex **>=** 0 **&&** tagIndex **<=** 8**)**

tag **=** "NN"**;**

**else** **if** **(**tagIndex **>=** 9 **&&** tagIndex **<=** 20**)**

tag **=** "PN"**;**

**else** **if** **(**tagIndex **>=** 21 **&&** tagIndex **<=** 40**)**

tag **=** "VB"**;**

**else** **if** **(**tagIndex **>=** 41 **&&** tagIndex **<=** 44**)**

tag **=** "JJ"**;**

**else** **if** **(**tagIndex **>=** 45 **&&** tagIndex **<=** 52**)**

tag **=** "RB"**;**

**else** **if** **(**tagIndex **>=** 53 **&&** tagIndex **<=** 54**)**

tag **=** "PP"**;**

**else** **if** **(**tagIndex **>=** 55 **&&** tagIndex **<=** 56**)**

tag **=** "CC"**;**

**else** **if** **(**tagIndex **>=** 57 **&&** tagIndex **<=** 69**)**

tag **=** "AT/DT"**;**

**else** **if** **(**tagIndex **==** 70**)**

tag **=** "."**;**

**else**

tag **=** "OT"**;**

**return** tag**;**

Aceasta este metoda care, in functie de index-ul returnat de metoda anterioara, va decide noul tag pentru cuvant.

ex. pentru setul exemplu = (“cuvant\_exemplu”, “nps$”),

GetTagIndexForConversion va returna 6,

ConvertBrownTagToHierarchicTag va returna “NN”

Setul nou: exemplu 🡺 (“cuvant\_exemplu”, “NN”)

Aici adaug graf in python cu noile tag-uri si frecventele lor

* 1. **Hidden Markov Model**
  2. **Decoder**
  3. **Evaluation**

End. Se adauga mai tarziu aici

**Bibliografie**

Brown Corpus manual: <http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM>

Brown Corpus wiki: <https://en.wikipedia.org/wiki/Brown_Corpus>

Brown Corpus data-set download: <https://archive.org/details/BrownCorpus>

Cross-Validation informatii: <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>

Cross-Validation exemple: <https://blog.quantinsti.com/cross-validation-machine-learning-trading-models/>

Cross-Validation exemple implementari: <https://scikit-learn.org/stable/modules/cross_validation.html>

Fisher-Yates Shuffling algoritm: <https://en.wikipedia.org/wiki/Fisher%E2%80%93Yates_shuffle>

Fisher-Yates Shuffling implementari algoritm: <https://stackoverflow.com/questions/273313/randomize-a-listt>

File Reader in c# informatii: <https://docs.microsoft.com/en-us/dotnet/standard/io/how-to-read-text-from-a-file>

Tokenization (Analiza lexicala) wiki: <https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization>

Parti de vorbire in gramatica limbii engleze (1): <https://www.grammar.cl/english/parts-of-speech.htm>

Parti de vorbire in gramatica limbii engleze (2): <http://www.butte.edu/departments/cas/tipsheets/grammar/parts_of_speech.html>

Parti de vorbire in gramatica limbii engleze (3): <https://www.englishclub.com/grammar/parts-of-speech.htm>

Parti de vorbire in gramatica limbii engleze (4): <https://www.english-grammar-revolution.com/parts-of-speech.html>