**Etichetarea părţii de vorbire**

**Bărbulescu Adrian, ISM 244/1**

Se adauga mai tarziu aici ..

Font – Times New Roman

Heading – 28 + bold

Subheading – 20 + bold

Main Body Text – 18

Bibliography text - 16

Links - 14

Captions – 12

**III. Consideratii practice**

1. **Descriere generala proiect**

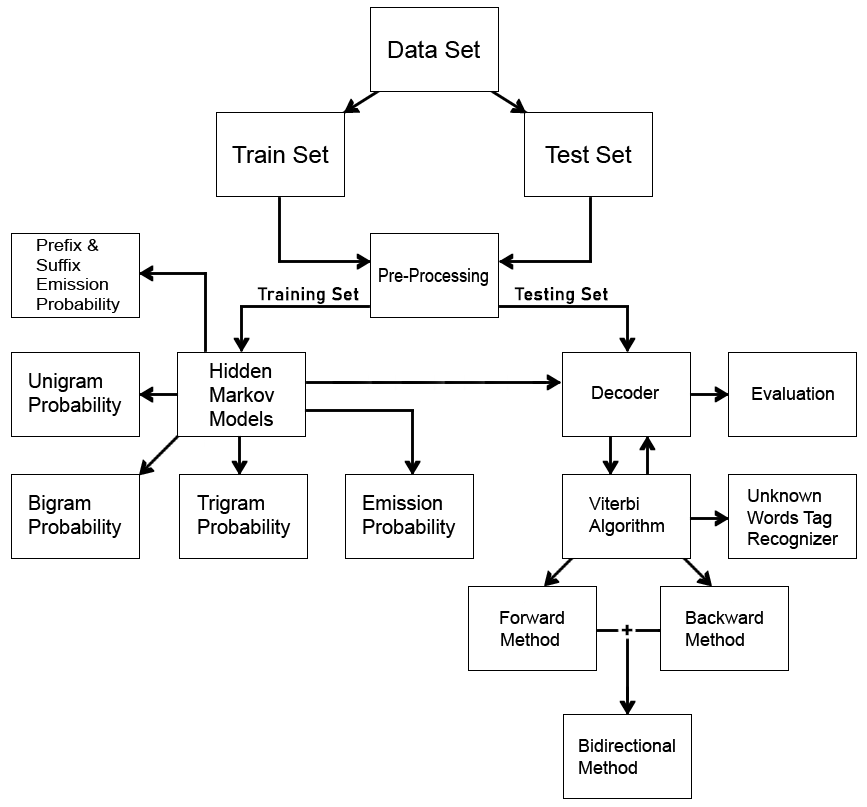


Figura 1 - Diagrama bloc funcţional

**2. Descriere separata bloc**

* **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 de stiinta – 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”.

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 e 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$ …

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

Mai multe exemple si informatii despre tag-uri pot fi gasite in manualul pentru Brown Corpus.

* **Train & Test Set -**
* **Pre-processing -**
* **Hidden Markov Model -**
* **Prefix & Suffix Emission Probability -**
* **Unigram Probability -**
* **Bigram Probability -**
* **Trigram Probability -**
* **Emission Probability -**
* **Decoder -**
* **Viterbi Algorithm -**
* **Unknown Words Tag Recognizer -**
* **Forward Method -**
* **Backward Method -**
* **Bidirectional Method –**
* **Evaluation -**

End. Se adauga mai tarziu aici

**Bibliografie**

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

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

Brown Corpus Download Samples: <https://archive.org/details/BrownCorpus>