Necesitatea Lematizarii.

**Lematizarea** reprezintã procesul prin care fiecare formã ocurentã într-un text este încadratã lemei sale. [...] Lematizarea formelor este procesul prin care se realizeazã funcþia inversã: fiind datã o formã, se determinã lema (lemele) în a cãrei (cãror) paradigmã apare forma în cauzã; concomitent cu determinarea acestor leme, se regãsesc si atributele aferente formei.

Căutarea se poate face atât folosind forma bază cât şi oricare dintre formele derivate (forme flexionare). Se pot căuta cuvinte precum mersului, sunt, luni, place, etc. Acolo unde este cazul, vă sunt sugerate şi alte forme ale cuvântului, cu şi fără diacritice.

Procesul de lematizare va furniza, pentru fiecare forma urmatoarele rezultate:

- lema, adica reprezentantul paradigmei cuvantului;

- clasa lexico-gramaticala a cuvantului (substantiv, adjectiv, verb etc.);

- subclasa lexico-gramaticala, daca este cazul(substantiv propriu sau comun, etc);

ln cazul cuvintelor flexibile, se stabilesc in plus celelalte determinari, dependente de clasa morfologica:

- pentru substantive si adjective: gen, numar, caz;

-pentru verbe: mod,timp,persoana, numar;

- pentru pronume: felul pronumelui, gen, numar, caz, persoana;

- pentru numerale: felul numeralului, gen, caz (unde exista).

Lematizarea este strans legata de stemming.

Diferența este că un Stemmer funcționează pe un singur cuvânt , fără cunoștințe de context, și, prin urmare , nu poate discrimina între cuvinte care au sensuri diferite , în funcție de parte de vorbire .

In concluzie un Stemmer este destul de usor de implementat, ruleaza mai repede, dar are o precizie mai scazuta fata de lematizare, de aceea nu poate inlocui lematizarea

Pentru o aplicatie ca cea la care lucram noi, tratarea computerizata a textelor este ceva ce nu poate lipsi. Module precum Lematizarea fac posibila dezvoltarea aplicatiilor de acest fel, dar deasemeanea si diverse aplicatii ce folosesc prelucrarea automata a textelor.

Ce am facut noi?

Pentru a face o statistica si pentru a vedea daca lematizarea este intradevar utila sau nu vom procedat astfel.

Vom lua un text normal pe care il vom lematizat conform algoritmului implementat de colegii nostrii.(Vom avea ca input outpu-ul de la echipa 3). Apoi vom separa cuvintele, acestea fiind separate astfel: cuvantul din text/ forma lematizata. Dupa ce vom crea un algoritm care sa faca aceasta separare, vom continua prin a crea 100 de propoziitii cu aceste cuvinte. Acestea vor fi luate la distante nu foarte mari intre ele(la nivel de text initial), si de asemenea fiecare cuvant ce urmeaza sa fie pus in propozitie trebuie sa aiba o pozitie mai mare in textul initial decat cuvantul curent.

Propozitiile vor avea o lungime cuprinsa intre 5 si 15 cuvinte.

La final vom verifica , automat, prin intermediul unui algoritm, cate din aceste propozitii folosesc cuvinte lematizate si cate nu. Astfel vom obtine , per pagina un procentaj al propozitiilor ce folosesc lematizare si un procentaj al propozitiilor ce nu folosesc lematizare.

In cele din urma vom aduna aceste procentaje pe un esantion de 10 fisiere pentru a vedea daca este nevoie sau nu de lematizare.

Rezultatele experimentului:

* Propozitii ce folosesc lematizare: 89,8 %
* Propozitii ce nu folosesc lematizare: 10,2 %