**Importanta Lematizarii**

*Partea a doua*

In aceasta parte a proiectului vom verifica cum afecteaza lematizare modul in care sunt identificate conceptele.

Mai exact, pe un text nelematizat, folosindu-se de rezultatul de la M2, identificam un anumit numar de concepte cu o anumita frecventa totala.

Pe acelasi text, de data aceasta lematizat, folosindu-se de rezultatul de la M3 de data aceasta, identificam un alt numar de concepte cu o alta frecventa totala (exista situatii cand cele doua identificari ar putea da acelas rezultat).

Urmatorul pas consta in identificare conceptelor care apar in a doua situatie, dar nu apar in prima situatie. Acestea sunt conceptele noi, ce apar in urma lematizarii. Pe un esantion de o suta de fisiere, vom obine un procentaj, care demonstreaza in cat la suta din cazuri apar noi concepte in urma lematizarii.

Pe esantionul nostru de o suta de fisiere am ajuns la concluzia ca in 46% din fisiere apar concepte noi dupa lematizare.

(Nota: multe din aceste fisiere nu au concepte identificate nici inaite si nici dupa lematizare, ceea ce afecteaza corectitudinea statisticii daca acele fisiere sunt goale. Daca nu, si suntem intr-o situatie in care chiar nu exista concepte, atunci statistica nu este afectata.)