Clasificator bayes multinomial

Analizarea discursurilor candidaților la funcția de Președinte al României.

Proiect realizat de: Militaru Ștefan-Octavian Cucu Mihnea

Conținut

[Scopul Proiectului 2](#_Toc182250783)

[Implementarea proiectului 2](#_Toc182250784)

[Obținerea datelor de antrenament 2](#_Toc182250785)

[Calcularea probabilităților „a priori ” 3](#_Toc182250786)

[Determinarea „cuvintelor cheie” 3](#_Toc182250787)

[Calcularea probabilităților condiționate 3](#_Toc182250788)

[Clasificarea utilizând modelul creat 5](#_Toc182250789)

[Testarea modelului 5](#_Toc182250790)

[Bibliografie 8](#_Toc182250791)

# Scopul Proiectului

Întrucât alegerile pentru funcția de Președinte al României sunt doar la un pas distanță, am ales ca temă pentru acest proiect la materia Probabilități și Statistică analizarea limbajului utilizat de candidați în discursurile acestora. Scopul programului realizat de noi este să detecteze cu un nivel de precizie nu foarte ridicat (întrucât implementarea unui clasificator Bayes este destul de naivă) originatorul unui text dat, utilizându-se de probabilitatea apariției anumitor cuvinte cheie în discurs.

# Implementarea proiectului

## Obținerea datelor de antrenament

Primul pas în realizarea proiectului a fost selectarea „categoriilor” în care va sorta clasificatorul Bayes datele primite. Am selectat pentru acestea cei 5 candidați care au o șansă reală la ajungerea în turul al doilea al alegerilor prezidențiale și anume Marcel Ciolacu, Nicolae Ciucă, Elena Lasconi, George Simion și Mircea Geoană.

Pentru fiecare dintre aceștia am căutat în jur de 5-6 discursuri (sau fragmente din discursuri întrucât unii candidați nu au fost documentați suficient de bine de către sursele media) susținute în ultimii câțiva ani. Temele discursurilor variază, întrucât nu am dorit ca menționarea unor cuvinte specifice acelor teme să „dea de gol” originea unui discurs, iar clasificarea acestora să fie strict bazată pe modul de formulare a propozițiilor și pe obiectivele principale pe care candidații doresc să le îndeplinească odată aleși de popor.

Sursele discursurilor vor fi adăugate în bibliografia documentației.

După ce acest pas a fost realizat, este nevoie să împărțim setul de date de antrenament în bucăți cu care putem lucra mai ușor. Am realizat acest proces prin utilizarea unui obiect de tip dicționar din limbajul Python unde reținem pentru fiecare candidat numărul aparițiilor unui anumit termen în toate discursurile purtate. Implementarea acestui pas este realizată în funcția „analiză discurs”.

## Calcularea probabilităților „a priori ”

Cu ajutorul datelor obținute la pasul anterior, putem calcula mai multe probabilități care să ne ajute să ne construim modelul matematic. În primul rând, doar din dicționarul asociat unui anumit candidat putem afla probabilitatea ca un cuvânt să se afle în discursul său, iar utilizând dicționarul „d\_total” care cumulează toate datele despre toți candidații, putem afla probabilitatea ca acel cuvânt să se afle în întreg setul de date.

## Determinarea „cuvintelor cheie”

Cele mai importante date care ne vor ajuta să realizăm clasificarea unui discurs sunt cuvintele cheie. Acestea sunt cuvinte alese în mod intenționat deoarece apar în discursul unui candidat mult mai des decât în discursurile celorlalți.

Determinarea lor se realizează cu ajutorul listei „lista\_ponderi” care va conține 5 liste de obiecte de tip tuplu, unde primul element este un cuvânt prezent în discursuri, iar cel de-al doilea vom stoca un număr care reprezintă de câte ori mai des este utilizat cuvântul de către un candidat decât media utilizării acestuia de către ceilalți patru. Spre exemplu, dacă Marcel Ciolacu utilizează cuvântul „proiect” de 10 ori, iar ceilalți candidați îl folosesc doar de 4-5 ori, ponderea cuvântului pentru Ciolacu va fi de 2, iar pentru ceilalți va fi în jur de 0,8.

„lista\_ponderi” va cuprinde așadari 5 liste pentru fiecare candidat, care vor fi sortate în funcție de ponderile cele mai mari și care vor determina cuvintele cheie. Dintre aceste cuvinte, vor fi incluse doar cele care au o pondere mai mare de 10, adică apar mai des de 10 ori la măcar unul dintre candidați.

## Calcularea probabilităților condiționate

O probabilitate condiționată este probabilitatea ca un eveniment să se întâmple, având deja informații că un alt eveniment s-a întâmplat.

• P(A | B): Probabilitatea ca evenimentul A să se întâmple, având deja informația că B s-a întâmplat.

În contextul nostru, vrem să calculăm probabilitatea ca un discurs să aparțină unui anumit candidat, dat fiind că un cuvânt cheie apare în acel discurs.

Pentru a clasifica un discurs pe baza cuvintelor cheie, folosim teorema lui Bayes, care este definită astfel:



(nu știam să scriu in Latex ☺)

Aceasta poate fi citită astfel:

• P(Candidat i | Cuvant): Probabilitatea ca discursul să aparțină candidatului **i**, dat fiind că un anumit cuvânt apare în discurs.

• P(Cuvant | Candidat i): Probabilitatea ca un cuvânt specific să apară într-un discurs al candidatului **i** (aceste valori le ai din datele anterioare).

• P(Candidat i): Probabilitatea a priori ca un discurs să aparțină candidatului **i** (presupunem că fiecare candidat are șanse egale, deci este 1/5 în cazul nostru cu 5 candidați).

• P(Cuvant): Probabilitatea generală ca acel cuvânt să apară într-un discurs, indiferent de candidat.

Funcția calculeaza\_probabilitati\_conditionate din cod folosește teorema lui Bayes pentru a calcula probabilitățile condiționate pentru fiecare candidat, dat fiind un cuvânt cheie.

Extragem probabilitatea generală a cuvântului (p\_cuvant = probabilitati[0]).

Calculăm probabilitatea a priori pentru fiecare candidat (probabilitate\_prior), presupunând că fiecare candidat are șanse egale (1/5).

Pentru fiecare candidat (de la Candidat\_1 la Candidat\_5):

• Extragem probabilitatea ca un cuvânt să apară într-un discurs al candidatului respectiv (p\_cuvant\_din\_candidat).

• Aplicăm teorema lui Bayes pentru a calcula probabilitatea condiționată:



• Salvăm rezultatul într-un dicționar.

Returnăm dicționarul cu probabilitățile condiționate pentru fiecare candidat.

## Funcțiile clasifica\_discurs și clasifica\_discurs\_nou

Cele două funcții, clasifica\_discurs și clasifica\_discurs\_nou, sunt utilizate pentru a identifica care dintre cei 5 candidați este cel mai probabil să fi rostit un discurs pe baza analizei cuvintelor cheie.

Funcția clasifica\_discurs(probabilitati) primește ca parametru o listă de probabilități care include: probabilitatea generală ca un cuvânt să apară într-un discurs și probabilitățile ca acel cuvânt să fie folosit de fiecare dintre cei 5 candidați. Folosind teorema lui Bayes, calculează probabilitățile condiționate pentru fiecare candidat și determină candidatul cel mai probabil pe baza valorii maxime. Returnează atât candidatul prezis, cât și toate probabilitățile condiționate.

Funcția clasifica\_discurs\_nou(cuvinte\_cheie\_discurs\_nou, date\_candidati) primește o listă de cuvinte cheie dintr-un discurs nou și un dicționar care conține probabilitățile pentru fiecare cuvânt cheie pentru toți candidații. Inițializează un scor pentru fiecare candidat, iar pentru fiecare cuvânt cheie din discursul nou verifică dacă acesta există în date\_candidati. Dacă da, folosește clasifica\_discurs pentru a calcula probabilitățile condiționate și actualizează scorul fiecărui candidat în funcție de acestea. Se returnează candidatul cu cel mai mare scor, indicând al cui este cel mai probabil discursul analizat, împreună cu scorurile detaliate pentru fiecare candidat.

## 

## Clasificarea utilizând modelul creat

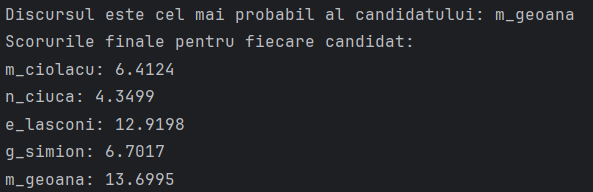
În acest subcapitol vom discuta funcționalitatea modelului creat la pașii anteriori. Acesta va citi dintr-un fișier un discurs care nu a fost inclus în datele de antrenament, iar cu ajutorul funcției „analiza\_discurs” va obtine un dicționar al cuvintelor din discursul nou. Apoi, programul va itera prin cuvintele cheie determinate la pasul anterior și va calcula probabilitatea apartenenței la o anumită categorie pe baza acestora. Aceste probabilități vor fi agregate și, la final, se va calcula un scor probabilistic care va determina apartența textului. Scorurile vor crește proporțional cu lungimea discursului, așa că vom avea o clasificare mai precisă pentru textele mai lungi.

## Prelucrarea input-ului de test

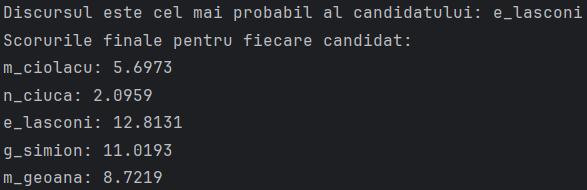
Se prelucrează un discurs nou pentru a determina al cui candidat este. Mai întâi, deschidem fișierul test\_input.txt (discursul nou) și folosim funcția analiza\_discurs() pentru a extrage cuvintele și a le număra, rezultând un dicționar (d\_test) unde fiecare cuvânt este o cheie, iar valoarea reprezintă frecvența sa în discurs. Apoi, parcurgem fiecare cuvânt din d\_test și, folosind funcția clasifica\_discurs\_nou, comparăm cuvintele cheie din discursul nou cu datele (date\_candidati) extrase anterior din discursurile candidaților. Funcția clasifica\_discurs\_nou utilizează aceste cuvinte cheie pentru a calcula scoruri pentru fiecare candidat, folosind probabilitățile condiționate calculate anterior. La final, determinăm candidatul cu cel mai mare scor, indicând al cui este discursul nou analizat.

# Testarea modelului

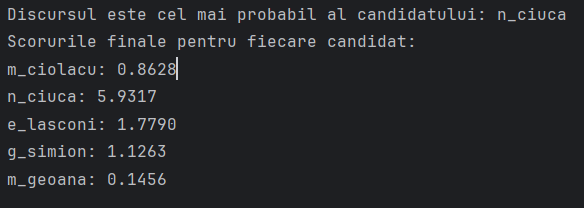
Test 1 (Input discurs al lui Mircea Geoană)



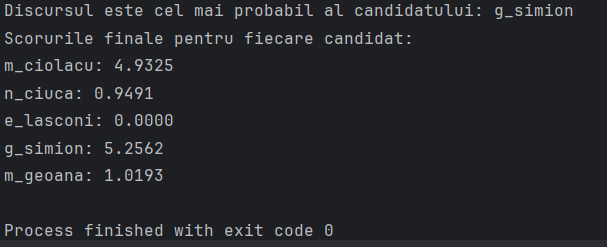
Test 2 (Input discurs al Elenei Lasconi)



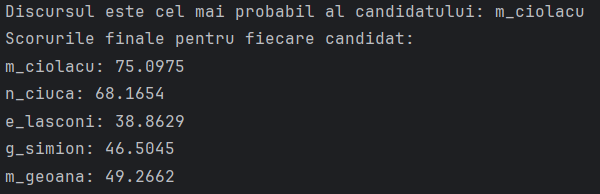
Test 3 (Input discurs al lui Nicolae Ciucă)



Test 4 (Input discurs al lui George Simion)



Test 5 (Input discurs al lui Marcel Ciolacu)



# 

# Bibliografie

Discursuri Marcel Ciolacu:

<https://romanialibera.ro/la-zi/discursul-lui-marcel-ciolacu-la-congresul-psd-2024-un-presedinte-pentru-toti/>

<https://www.psd.ro/video-discursul-presedintelui-psd-marcel-ciolacu-cu-prilejul-aniversarii-15-ani-de-la-aderarea-romaniei-la-uniunea-europeana/>

<https://www.psd.ro/video-discursul-sustinut-de-presedintele-psd-marcel-ciolacu-in-plenul-reunit-al-camerei-deputatilor-si-senatului/>

<https://www.europafm.ro/premierul-marcel-ciolacu-discurs-de-ziua-micii-uniri-sarbatorind-ziua-unirii-suntem-insa-datori-sa-reafirmam-responsabilitatea-noastra-fata-de-generatiile-viitoare/>

<https://economedia.ro/premierul-marcel-ciolacu-discurs-de-prezidentiabil-in-fata-companiilor-americane-romania-poate-deveni-in-urmatorii-4-5-ani-una-din-marile-forte-economice-ale-continentului-poate-sa-aiba-un.html>

<https://gov.ro/ro/stiri/discursul-sustinut-de-premierul-marcel-ciolacu-in-plenul-parlamentului-inaintea-votului-de-investitura>

Discursuri Nicolae Ciucă:

<https://gov.ro/ro/stiri/discursul-premierului-desemnat-nicolae-ciuca-in-plenul-parlamentului>

<https://www.senat.ro/StiriSenatDetaliu.aspx?ID=D50B9424-62E3-4A38-BAA2-BC601D440F0F>

<https://gov.ro/ro/stiri/discursul-premierului-nicolae-ionel-ciuca-la-ia-i-cu-prilejul-sarbatoririi-unirii-principatelor-romane>

<https://www.bursa.ro/nicolae-ciuca-armata-romana-o-perioada-nu-a-beneficiat-de-finantarea-necesara-pentru-a-avea-un-ritm-coerent-de-dotare-65974353>

Discursuri Mircea Geoană:

<https://www.caleaeuropeana.ro/video-interviu-mircea-geoana-la-summitul-nato-de-la-washington-sprijinul-nato-o-garantie-suplimentara-ca-ucraina-si-r-moldova-isi-pot-indeplini-calea-europeana/>

<https://www.digi24.ro/stiri/actualitate/politica/ca-viitor-presedinte-al-romaniei-mircea-geoana-l-a-invitat-pe-trump-in-romania-as-fi-onorat-sa-ne-intalnim-cat-de-curand-posibil-2998373>

<https://hotnews.ro/mircea-geoana-dupa-ce-ministrul-digitalizarii-a-anuntat-ca-a-descoperit-zeci-de-conturi-nu-exista-ferme-de-troli-in-romania-este-clar-1830750>

<https://www.g4media.ro/mircea-geoana-nici-psd-nici-usr-si-nici-macar-pnl-nu-l-ar-propune-pe-bolojan-premier-efectiv-il-folosesc-pentru-ca-are-o-reputatie-buna-ciuca-mai-degraba-l-ar-pune-premier-pe-iohannis.html>

<https://www.libertatea.ro/stiri/mircea-geoana-interviu-integral-55-intrebari-powered-by-libertatea-5071258>

<https://www.euronews.ro/articole/mircea-geoana-anunta-oficial-candidatura-la-alegerile-prezidentiale-2024>

Discursuri Elena Lasconi:

<https://adevarul.ro/stiri-interne/societate/elena-lasconi-acuzata-de-un-discurs-jignitor-la-2393357.html>

https://www.digi24.ro/stiri/actualitate/politica/scandalul-fermei-de-troli-lasconi-geoana-minte-cum-respira-nu-eu-trebuie-sa-cercetez-ci-institutiile-statului-2989609

<https://www.g4media.ro/video-interviul-integral-cu-elena-lasconi-anticoruptia-e-proiectul-meu-in-romania-exista-un-razboi-real-influenta-ruseasca-voi-avea-in-jurul-meu-oameni-mai-destepti-decat-mine-oameni-care-sunt.html>

<https://www.euronews.ro/articole/usr-si-a-lansat-programul-de-guvernare-elena-lasconi-romania-merita-mai-mult-deca>

<https://podul.ro/articol/24532/video-elena-lasconi-discurs-memorabil-la-depunerea-candidaturii-ii-reprezint-pe-romanii-care-de-35-de-ani-se-simt-tradati-si-pacaliti-de-aceasta-clasa-politica-nu-intru-in-dezbateri-cu-candidatii-prorusi-rusii-ne-au-facut-foarte-mult-rau-noi-suntem-pentru-vest-mie-imi-plac-blugii-imi-place-coca-cola-intotdeaua-ne-am-dorit-sa-fim-cu-americanii>

<https://www.euronews.ro/articole/elena-lasconi-este-noul-presedinte-usr>

<https://www.euronews.ro/articole/lasconi-critica-dur-sistemul-de-sanatate-si-educatie-romania-nu-este-o-tara-sarac>

<https://www.euronews.ro/articole/elena-lasconi-despre-congresul-psd-un-festival-al-dezinformarii-al-minciunilor-si>

Discursuri George Simion:

<https://www.gandul.ro/politica/exclusiv-george-simion-discurs-mobilizator-la-roma-sa-castigam-impreuna-toate-bataliile-i-a-facut-un-dar-giorgiei-meloni-ce-a-discutat-cu-elon-musk-20113700>

<https://www.mediafax.ro/politic/george-simion-discurs-la-cpac-in-mexic-sa-fim-luptatori-nu-marionete-22469190>

<https://stirileprotv.ro/stiri/actualitate/george-simion-candidatul-aur-la-presedintie-fata-in-fata-cu-andreea-esca-de-la-ora-19-00-in-platoul-stirilor-pro-tv.html>

<https://georgesimion.ro/cum-e-oltenia-e-toata-romania-o-tara-uitata-video/>

<https://georgesimion.ro/suntem-inca-stapani-in-tara-noastra-video/>

<https://georgesimion.ro/romania-e-sub-jugul-psd-si-pnl-e-nevoie-de-o-schimbare-e-nevoie-ca-romanii-sa-fie-din-nou-stapani-la-ei-in-tara/>

<https://georgesimion.ro/planul-simion-pentru-transport-feroviar-punem-romania-inapoi-pe-sine-video/>

<https://georgesimion.ro/planul-simion-pentru-energie-romanii-trebuie-sa-fie-stapani-la-ei-in-tara-video/>

<https://georgesimion.ro/educatia-e-singura-sansa-de-a-construi-un-viitor-mai-bun-fiecare-copil-merita-sa-aiba-sanse-egale-video/>

<https://georgesimion.ro/planul-simion-pentru-turism-mergeti-la-talpa-tarii-mergeti-la-motoarele-economice-sa-facem-din-romania-tara-turismului/>