Graphical user interface

Description automatically generated**Segmentarea Topică a Documentelor Text**

Figura , [IBM](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)

Înainte de a adresa tematica acestui articol, trebuie discutată problema pe care încearcă să o soluționeze. Un clișeu, dar și un adevăr este faptul că 90% din datele create in totalitate de umanitate au fost create în ultimii doi ani, după cum se poate observa în *figura 1*.

Însă, ce este rar menționat, este că mai bine de 80% din aceste date sunt nestructurate - majoritatea datelor generate sunt sub formă de fișiere video, text, imagini fișiere 3D și altele.

Acest fapt creează o dificultate în a folosi aceste date, deoarece, spre deosebire de datele tabulare, nu există o cale simplă de a sorta, segmenta sau a filtra aceste informații. Acest fapt este datorat lipsei unei etichete care poate fi folosită pentru filtrare. Lipsa unei astfel de ordini duce la inaccesibilitatea datelor pentru *user*-ul final.

**Problema cu Datele Text**

Datele Text sunt, pe departe, cea mai răspândită formă de date și, adesea, cea care conține cea mai multă informație. Există zicala că o poză valorează o mie de cuvinte, însă - pragmatic vorbind - nu se învață algebră sau inginerie din imagini. Totuși, nu vorbim doar de documente text de tip tehnic - majoritatea datelor text curent generate sunt sub formă de articole, *review*-uri, *comment*-uri, documente istorice și multe alte exemple, postate fără etichetare sau o altă modalitate de a discerne informațiile din aceste documente. Problema ar putea fi rezolvată prin implementarea unei convenții de etichetare încă de la începutul fiecărei platforme, dar în momentul în care o convenție este gândită, deja există mulți *terabytes* de date text ce trebuie reprocesate. Și, chiar în cazul în care o convenție a fost implementată de la început, nu am avea o garanție că aceste convenții s-ar aplica datelor deja generate sau datelor ce urmează a fi.

În această privință, acest articol intenționează să demonstreze că, atunci când avem de-a face cu datele text, problema întâmpinată nu este chiar atât de dificilă. În multe cazuri, *Machine Learning-*ulpoate rezolva aceste probleme sau discerne aceste informații, fără a fi nevoiți să recurgem la un proces ce ar avea, în caz contrar, un cost ridicat în termeni de timp și resurse.

**Sursele de Date**

Articolul curent menționează folosirea a 3 surse de date, toate prezentând problema unei lipse de structură, făcând filtrarea imposibilă.

Ar fi dificil pentru orice persoană să înțeleagă lungimea acestor texte doar după o numărătoare de cuvinte. Deci, pentru a fi mai ușor inteligibil cât de mare este volumul de text, intenționăm să folosim un punct de referință simplu - o comparație cu lungimea Bibliei.

Sursele de date sunt următoarele:

* *NPR Articole de Știri, 9.266.936 cuvinte, de 11.47 mai lung decât Biblia;*
* *Spotify Million Song Dataset, 12.653.383 cuvinte, de 15.67 mai lung decât Biblia;*
* *Women’s Clothing Review, 1.221.308 cuvinte, de 1.51 mai lung decât Biblia.*

După cum se poate observa, cele 3 surse de date sunt de origini total diferite, toate prezentând aceeași problemă - nu există o etichetare, deci nu avem posibilitatea filtrării. Acest fapt, combinat cu volumul mare de date text, duce la o dificultate ridicată în a obține informații utile din datele respective, în timp util.

Cea mai evidentă soluție ar fi crearea unui *count vectorizer sau mai avansat Term Frequency-Inverse Document Frequency embeding,* și folosirea unui algoritm clasic de segmentare pe rezultatul acestor procese, însă s-ar crea o segmentare fără a ne permite o înțelegere a particularităților acelor segmente. Astfel, rezultatul ar fi segmentarea datelor inițiale și apariția necesității parcurgerii întregului volum, pentru a discerne informații, în segmente care prezintă exact aceeași problemă. Concluzionăm că problema resurselor și timpului necesare nu a fost rezolvată, ci doar împărțită în probleme mai mici.

Totuși, există algoritmi de segmentare care nu prezintă această problemă, algoritmii folosiți în acest articol fiind Alocarea Latent Dirichlet și Factorizarea Matricelor Non-Negative.

Ambii algoritmi se bazează pe aceleași asumpții:

* documentele s-ar segmenta în jurul unor cuvinte sau serii de cuvinte ce definesc aceste segmente;
* fiecare topic are un cuvânt sau o serie de cuvinte ce au o frecvență sau o probabilitate mai mare de a fi folosite decât acele cuvinte ce nu definesc topicul.

**NPR**

Chart

Description automatically generatedSetul de Date este format din 11.991 de articole de știri neetichetate. Știm că sunt articole de știri dar, neavând o etichetă, nu avem cum să fim siguri dacă sunt despre politică sau despre meteo.

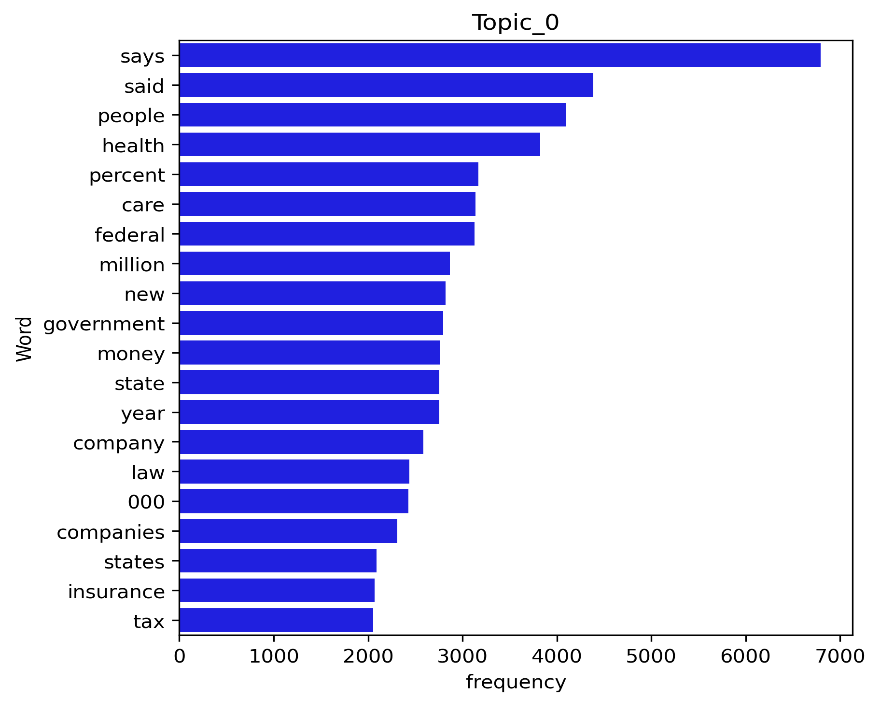
Știm care sunt cele mai des folosite cuvinte: cuvintele de stop din limba engleză, dar nu putem să le folosim pe acestea pentru a categorisi articolele, după cum poate fi observat în *figura 2*. Și, chiar eliminând aceste cuvinte de stop, este dificil să determinăm un subiect, fapt vizibil în *figura 3*. Putem, astfel, concluziona că un simplu filtru pe cuvinte cheie nu poate fi folosit pentru a segmenta articolele, deci, recurgem la metodele de categorisire menționate anterior.

Figura

Începem prin asumarea unui număr de șapte segmente. Nu există un număr corect de segmente, dar un număr mai mare va crea segmente mai specifice, un număr mai mic va crea segmente mai generale.

Astfel, rezultă un număr de șapte segmente, fiecare având o frecvență diferită pentru cuvinte diferite. De exemplu, pentru primul topic, *(Topic\_0)*, putem observa în *figura 4* că cele mai ridicate frecvențe se găsesc la cuvintele de natură financiară. Se observă că algoritmul nu oferă o denumire topică. Numele segmentul poate fi determinat pe baza cuvintelor cu frecvența cea mai mare pentru fiecare topic. 

Figura

Obținem următoarele rezultate:

* Topic\_0: [*tax insurance states companies 000 law company year state money government new million federal care percent health people said says*] - putem clasifica acest topic ca ***Finance;***
* Topic\_1: [*attack state military war news department country according reported president russia security npr reports told government people says police said*] - putem clasifica acest topic ca fiind ***International Nwes,*** din punctul de vedere al SUA;

Figura

* Topic\_2: [*local little land small way year make world home time day city new years just water people food like says*] - putem clasifica acest topic ca și ***National or Local News****;*
* Topic\_3: [*brain time years research new don percent just care drug children like disease medical patients study women health people says*] - putem clasifica acest topic ca fiind ***Medical Research****;*
* Topic\_4: [*presidential just voters political vote donald party new people republican election white house obama state campaign clinton president said trump*] - putem clasifica acest topic ca fiind ***Presidential Election****;*
* Topic\_5: [*black says world ve said going story years don life music way really new think know time people just like*] - putem clasifica acest topic ca fiind ***Cultural News****;*
* Topic\_6: [*children work science kids make really way schools don university education time new think just like people students school says*] - putem clasifica acest topic ca fiind ***Education****.*

Pentru a testa asumțiile făcute, au fost selectate mai multe articole în mod aleatoriu.

Cultural News

Order in the court — but maybe not in movie theaters. With all the talk lately about politics and the judiciary — fights over Supreme Court vacancies, the President complaining about”” judges — I’ve been thinking about the judges I’ve seen on screen, and how their depiction might have intersected with public opinion through the years. Ce

Finance

Updated at 2:30 p. m. ET, for years, Puerto Rico has grappled with an debt crisis, watching as its bills have grown to more than $70 billion. Including what the U. S. territory owes to pension funds, that debt exceeds $120 billion. Now, Puerto Rico’s struggle with its creditors has stepped into U. S. federal court, where an unprecedented case

Testele de mai sus tind să confirme asumpțiile făcute.

**Spotify**

S-a putut observa că aceste metode de segmentare funcționează pentru articole de știri, dar pot fi testate în circumstanțe diferite, pe versurile cântecelor. Totuși, ar trebui limitate așteptările a ceea ce pot face acești algoritmi; așteptările nu ar trebui să fie ca algoritmii să determine genul muzicii, ci, după cum sugerează numele, să determine topicul sau subiectul cântecelor care adesea nu coincid cu un gen de muzică.

A picture containing chart

Description automatically generatedSetul de date Spotify are un total de 57.650 versuri de cântece; vom vedea dacă acestea pot fi segmentate în funcție de topic.

Figura

Chart

Description automatically generatedDupă cum se putea aștepta, cele mai des întâlnite cuvinte în acest set de date sunt cuvinte englezești de stop - *figura 5*; dacă le ignorăm pe acestea, putem observa că cele mai des întâlnite cuvinte sunt de natură sentimentală. Astfel, putem presupune că segmentarea topică va fi în relație cu un sentiment.

Același algoritm ca în cazul precedent a fost rulat peste setul de date Spotify, asumând același număr de șapte segmente. Putem observa în *figura 6* că cele mai întâlnite cuvinte pentru al treilea topic au o conotație religioasă, deci putem asuma un topic de natură religioasă. Ca și înainte, algoritmul nu îi asumă o denumire segmentului, ci un număr; denumirea este implicită din cuvintele cu frecvența cea mai mare în fiecare topic.

Figura 6

Astfel, denumirea topicelor a fost dedusă după cum urmează:

* Topic\_0: *[wanna like hey know girl come let love want got don gonna yeah baby oh]* - putem clasifica acest topic ca ***Subject of a Courtship****;*
* Topic\_1: [*wind blue time dream rain day come sky eyes light sun away ll night like*] - putem clasifica acest topic ca ***Nature****;*
* Topic\_2: [*feel away life want like way heart say time ve ll just don know love*] - putem clasifica acest topic ca ***Love***. Se poate observa că algoritmul distinge între Subiectul actului de curte și sentimentul de dragoste;
* Topic\_3: [*free born heaven man sing die jesus soul oh let come life world lord god*] - putem clasifica acest topic ca ***Divine***;
* Topic\_4: [*way day ll long town good little ve got just old said home man la*] - putem clasifica acest topic ca ***Locational***;
* Topic\_5: [*la santa gimme music di ba happy ha doo roll dance rock da christmas na*] - putem clasifica acest topic ca **Holidays**;
* Topic\_6: [*nigga chorus fuck man shit money em just cause yaain know don got like*] - putem clasifica acest topic ca ***Rap Music***.

Ca și înainte, asumpțiile făcute trebuie testate cu unele selecții aleatorii:

***Love***

*All my love*

*Came to nothin' at all my love*

*When I woke up to find*

*You were no longer mine*

*All my love*

*Throw away after all this time*

*Now there's no place for me*

*In the future you see*

*I don't understand you*

*I've done all I c*

*Song: All My Love by Cliff Richard*

***Locational***

*They're really rockin' Boston*

*In Pittsburgh, P. A.*

*Deep in the heart of Texas*

*And 'round the Frisco Bay*

*All over St. Louis*

*And down in New Orleans*

*All the cats want to dance with*

*Sweet little sixteen*

*Sweet little sixteen*

*Song: Sweet Little Sixteen by Chuck Berry*

Asumpțiile par a fi corecte. Acum putem verifica popularitatea topicelor în baza de date, după cum poate fi observat în *figura 7*.

Chart, bar chart

Description automatically generated

Figura 7

***Women’s Clothing***

Chart, bar chart

Description automatically generatedNu, această denumire nu este începutul unei glume. Din punct de vedere tehnic, este, mai degrabă, o tragedie – motivul e că, în calitate de comerciant de haine pentru femei, ai presupune că e relativ direct modul în care înțelegi dorința clientelor când vine vorba de produsele vândute. Desigur, e relativ ușor să măsori satisfacția per total, bazându-te pe un sistem de rating, după cum poate fi observat în *figura 8*.

Figura

Text

Description automatically generatedSe poate observa că, per total, clientele sunt mulțumite de produsele cumpărate, dar nu ai cum să determini care este interesul clientelor, dacă nu te folosești de textul *review*-urilor.

Într-un astfel de mediu, se poate observa că se pot folosi exact aceleași unelte ca și în cazurile prezentate anterior. În *figura 9*, sunt ilustrate cele mai des întâlnite cuvinte, dar acestea nu ne oferă topicurile de interes; ca și în cazurile precedente, trebuie să aplicăm Alocarea Latent Dirichlet, cu o asumare de 7 topicuri de interes. Rezultatele sunt următoarele segmente - primele 20 de cuvinte după frecvența folosirii:

Figura

* Topic\_0: *[don, ordered, went, got, jeans, did, pants, try, love, bought, just, online, price, fit, retailer, tried, saw, size, sale, store*] - putem clasifica acest topic ca ***Sale****;*
* Topic\_1: [*petite, bit, right, nice, little, great, short, just, love, hips, look, fabric, flattering, long, like, length, skirt, size, fit, waist*] - putem clasifica acest topic ca ***Fit***;
* Topic\_2: [*long, colors, cute, nice, black, look, bought, fall, like, looks, jacket, comfortable, jeans, perfect, soft, wear, color, sweater, love, great*] - putem clasifica acest topic ca și ***Comfort***;
* Topic\_3: [retailer, run, like, love, bit, fits, lbs, big, little, wear, usually, runs, petite, medium, fit, xs, ordered, large, small, size] - putem clasifica acest topic ca ***Small Sizes****,* algoritmul pare să distingă între ***Fit*** și ***Small Sizes***;
* Topic\_4: [*looks, colors, material, blouse, pretty, bit, sheer, soft, really, underneath, little, bra, nice, love, wear, color, white, shirt, like, fabric*] - putem clasifica acest topic ca ***Material and Color***;
* Topic\_5: [*cut, cute, loved, beautiful, thought, wanted, work, model, material, looks, fit, way, didn, looked, really, fabric, look, just, dress, like*] - putem clasifica acest topic ca ***Appearance****;*
* Topic\_6: [*work, gorgeous, fabric, true, quality, recommend, summer, dresses, fit, compliments, size, fits, comfortable, beautiful, flattering, great, perfect, wear, love, dress*] - putem clasifica acest topic ca ***Dresses***. Acesta pare a fi singurul topic ce se suprapune cu un tip de haine, dar nu este identic cu categoria acestora.

Ca și în cazurile precedente, trebuie testate asumpțiile printr-o selecție aleatorie:

Appearance

I love natural colors and nature themes, but this top was a disappointment. I ordered both the small and the XS, because retailer can run very large. The small was too big and the XS did not hang correctly, the arms hit at a weird spot and the neckline wouldn't sit right. Overall, the fit was boxy and unflattering. The fabric is pretty, but the style is well.... boring. There was just no wow factor for me. I think the fabric deserved a much better design.

Material and Color

I'm not usually a fan of simple crewneck sweaters, but this one has some nice added details that make it a keeper. The dark red oxblood color is nice (it manages to be a red sweater without feeling overly holiday), the detailing at shoulder/chest is interesting yet subtle enough to not be distracting and doesn't feel bulky at all, the dropped shoulder seams are done correctly and without the body becoming boxy. It does feel like it might run a tad bit large, but only very slightly.

Asumpțiile par a fi adevărate, însă această investigație se poate duce și mai departe. Avem un rating pentru produsele vândute, deci se poate investiga și distinge care sunt topicurile de interes pentru produsele cu rating pozitiv și cele cu rating negativ - se pot obține informații valoroase despre cliente.

Se începe cu convenția că toate produsele cu un rating de 4 stele sau mai mult sunt considerate produse cu rating pozitiv, iar restul sunt considerate produse cu rating negativ. Se separă datele în *review*-uri pozitive și negative, iar asupra fiecărui set de date se aplică separat același algoritm ca și în exemplele anterioare. În acest caz, au fost asumate doar patru topicuri. Se pot observa în *figura 10* topicurile discutate în *review*-uri pozitive; același lucru poate fi observat legat de cele negative în *figura 11*.

Figura

Produsele cu review-uri pozitive tind să aibă topicul legat de:

* *Well-fitting small sizes* (Mărimi mici cu un *fitting* bun);
* *Appearance* (aspect);

Figura 11

* *Casual Style* (stil *casual*);
* *Price to Quality Ratio* (raport preț-calitate).

În același timp, produsele cu *review*-uri negative tind să aibă conectat cu:

* *Tops that do not fit or look well* (bluze care nu se potrivesc bine sau nu arată bine);
* *Dresses made of cheap fabric* (rochii din material ieftin);
* *Tops with a bad Color or Material* (bluze cu o culoare sau material nesatisfăcătoare);
* *Items of clothing that are labeled as a small size but are larger than expected* (haine care au mărimi mici, dar au o dimensiune mai mare decât cea așteptată).

Astfel, printr-un algoritm de segmentare, s-a putut observa perspectiva clientelor atât în privința produselor care sunt pe placul lor, cât și a celor care nu sunt.

**Concluzie**

Când este luat în considerare faptul că majoritatea datelor generate în mod curent sunt nestructurate și o bună parte a acestora este sub formă de text, atât companiile, cât și indivizii se vor confrunta cu mari dificultăți în a încerca să extragă informații utile din acestea. Totuși, cel puțin în cazul datelor de tip text, există anumite unelte care pot ajuta în această privință și, în ciuda dificultăților întâmpinate, se pot obține informații valoroase din aceste date.

**Referințe**

* [cod și imagini](https://github.com/alex-parvu/Articles/tree/main/02.Text%20Document%20Clustering)
* [Date NPR](https://www.kaggle.com/datasets/gauravduttakiit/npr-data)
* [Date Spotify](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset)
* [Date Women’s Clothing](https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews)
* [Esimări IBM](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)