Graphical user interface

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

Figura 1 sursă: [IBM](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)

Înainte de a adresa topicul acestui articol, trebuie discutată problema ce încearcă a o rezolva. Este deja un clișeu, dar este adevărat că 90% din datele create în totalitate de umanitate, au fost create în ultimii doi ani, după cum se poate observa in figura.1.

Însă, ce este rar discutat, este faptul că mai bine de 80% din aceste date sunt nestructurate. Majoritatea datelor generate în ziua de azi sunt sub forma 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 categorisi, segmenta sau a filtra aceste informații. În multe cazuri acest fapt se datorează lipsei unei etichete ce poate fi folosita pentru filtrare. Datorită acestei lipse de ordine datele devin inaccesibile pentru *user*-ul final.

**Problema cu Datele Text**

Datele Text sunt pe departe cea mai răspândită forma de date și, adesea, cea care conține cea mai multă informație. Exista zicala că o poză valorează o mie de cuvinte, însă pragmatic, nu se învață algebră sau inginerie din imagini, dar nu vorbim doar de documente text de tip tehnic. Majoritatea datelor text generate în ziua de azi sunt sub formă de articole, *review*-uri, comment-uri, documente istorice și multe alte exemple, postate fără etichetare sau o modalitate de a discerne informațiile din aceste documente într-un mod eficient. Problema ar putea fi rezolvată prin implementarea unei convenții de etichetare încă de la începutul fiecărei platforme, dar, adesea, în momentul în care o convenție este gândită, deja există mulți *terabytes* de date text ce trebuie reprocesate. Și, chiar în cazul in 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 dovedească 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 nevoit sa recurgi la un proces ce ar avea altfel un cost ridicat în timp și resurse.

**Sursele de Date**

In acest articol au fost folosite 3 surse de date, toate prezentând problema unei lipse de structura, astfel făcând filtrarea imposibila.

Ar fi dificil pentru orice persoana sa înțeleagă lungimea acestor texte doar după o numărătoare de cuvinte. Astfel pentru a fi mai ușor înțeles cat de mare este volumul de text, intenționăm sa folosim un punct de referință simplu, anume 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 problema, nu exista o etichetare, astfel ne având posibilitatea filtrării. Acest fapt combinat cu volumul mare de date text duce la o dificultate ridicata in a obține informații utile din datele respective, in timp util.

Cea mai evidenta soluție ar fi crearea unui *count vectorizer (or Term Frequency-Inverse Document Frequency),* si sa folosim un algoritm clasic de segmentarea pe rezultatul acestor procese. Însă, asta ar crea o segmentare fără a ne permite o înțelegere a particularităților acelor segmente. Astfel rezultatul ar fi segmentarea datelor inițiale, si a problemei de a fi necesare parcurgerea întregului volum pentru a discerne informații, in segmente care prezinta exact aceeași problema. Astfel problema in materie de resurse si timpul necesar pentru acest fapt nu au fost rezolvate, ci doar împărțite in probleme mai mici.

Însă, exista algoritmi de segmentare ce nu prezinta aceasta problema. Algoritmii folosiți in acest articol fiind Alocarea Latent Dirichlet si Factorizarea Matricelor Non Negative.

Ambii algoritmi se bazează pe aceleași asumpții:

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

**NPR**

Setul de Date este format din 11.991 de articole de știri neetichetate. Știm ca sunt articole de știri dar, neavând o eticheta nu avem cum sa știm daca sunt despre politica sau despre meteo.

Chart

Description automatically generatedȘtim care sunt cele mai des folosite cuvinte, anume cuvintele de stop din limba engleza, neputând sa le folosim pe acestea pentru a categorisi articolele, după cum poate fi observat in figura.2. Si chiar eliminând aceste cuvinte de stop, este dificil sa determinam un subiect, după cum se poate observa in figura.3. Putem astfel conclude ca un simplu filtru pe cuvinte cheie, nu poate fi folosit pentru a categorisi articolele. Astfel recurgem la metodele menționate anterior.

Figura 2

Începem cu asumare unui număr de șapte segmente. Nu exista 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 rezultam cu un număr de 7 segmente, fiecare având o frecventa diferita pentru cuvinte diferite. De exemplu pentru primul topic, (Topic\_0), putem observa in figura.4, ca cele mai ridicate frecvente se găsesc la cuvintele de natura financiara. Se observa ca algoritmul nu oferă o denumire topica. Numele segmentul poate fi determinat, pe baza cuvintelor cu frecventa cea mai mare pentru fiecare topic.

Figura 3



Astfel obținem următoarele rezultate:

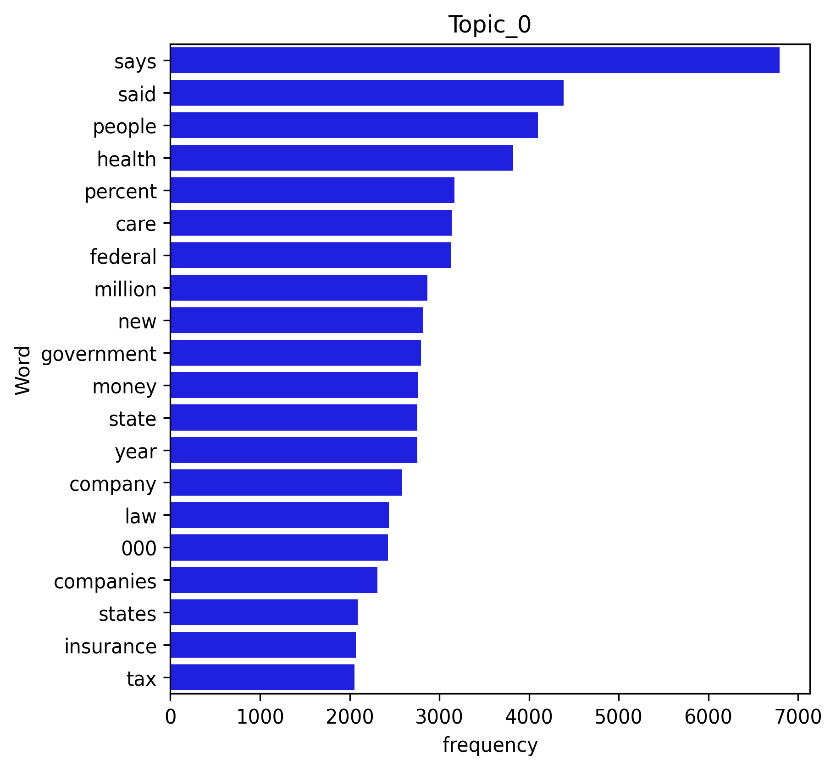
* Topic\_0: [*tax insurance states companies 000 law company year state money government new million federal care percent health people said says*]. Putem clasifica acest topic ca ***Finance,***
* **Topic\_1: [*attack state military war news department country according reported president russia security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

Figura 4

* 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 asumtiile facute, au fost selectate mai multe articole in mod aleatoru.

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 sa confirme asumpțiile făcute.

**Spotify**

A picture containing chart

Description automatically generatedSa putut observa ca aceste metode de segmentare funcționează pentru articole de știri. Însă pot fii încercate in circumstanțe diferite, în cazul de față pentru liricele de cântece.. Însă ar trebui limitate așteptările a ceea ce pot face acești algoritmi. Așteptările nu ar trebui sa fie ca algoritmii să determine genul muzicii, ci după cum sugerează numele ar determina topicul sau subiectul cântecelor care adesea nu coincide cu un gen de muzică. Setul de date Spotify are un total de 57.650 lirice de cântece. Vom vedea, daca acestea pot fi segmentate in funcție de topic.

Figura 5

Chart

Description automatically generatedDupă cum se putea aștepta cele mai des întâlnite cuvinte in acest set de date sunt cuvinte englezești de stop, după cum se poate observa in figura.5. Însă daca le ignoram pe acestea, putem observa ca cele mai des întâlnite cuvinte sunt de natura sentimentală. Astfel putem asuma ca segmentarea topică va fi in relație cu un sentiment.

Același algoritm ca in cazul precedent a fost rulat peste setul de date Spotify, asumând același număr de șapte segmente. Putem observa in figura.6, ca cele mai întâlnite cuvinte pentru al treilea topic au o conotație religioasa, astfel putem asuma un topic de natura religioasă. Ca si înainte, algoritmul nu ii asuma o denumire segmentului ci un număr, denumirea este implicită din cuvintele cu frecvența cea mai mare in fiecare topic.

Figura 6

Astfel denumirea topicelor au fost deduse 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 ca algoritmul distinge intre Subiectul actului de curte si 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 si 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 in baza de date, după cum poate fi observat in figura.7.:

Chart, bar chart

Description automatically generated

Figura 7

**Women’s Clothing**

Nu aceasta denumire nu este începutul unei glume. Din punct de vedere tehnic este mai degrabă o tragedie. Motivația fiind ca in calitate de comerciant de haine pentru femei ai asuma ca ii relativ direct modul in care înțelegi dorința clientelor când vine vorba de produsele vândute.Chart, bar chart

Description automatically generated Desigur ii relativ ușor să măsori satisfacția per total, bazându-te pe un sistem de rating după cum poate fi observat in figura.8.

Figura 8

Se poate observa ca per total, clientele sunt mulțumite de către produsele cumpărate. Dar nu ai cum sa înțelegi care este interesul clientelor daca nu te folosești de textul review-urilor.

Text

Description automatically generatedIn astfel de mediu se poate observa ca se pot folosii exact același unelte ca si in cazurile prezentate anterior. In figura.9 se pot observa cele mai des întâlnite cuvinte, dar acestea nu ne dau topicurile de interes. Ca si in cazurile precedente trebuie sa aplicăm Alocarea Latent Dirichlet cu o asumare de 7 topicuri de interes. Rezultatele fiind următoarele segmente împreună cu primele 20 de cuvinte după frecventa folosirii:

Figura 9

* 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 ***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 considere ca femeile fac o distincție intre ***Fit*** si ***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 si in 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 really 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 se poate duce acesta investigație si mai departe . Având un rating pentru produsele vândute; astfel se poate investiga si sa distinge care sunt topicurile de interes pentru produsele cu rating pozitiv si cele cu rating negativ, astfel se pot obține informații valoroase despre cliente.

Se începe cu convenția ca toate produsele cu un rating de 4 stele sau mai mare, sunt considerate produsele cu rating pozitiv iar restul sunt considerate produsele cu rating negativ. Se separa datele in *review*-uri pozitive si negative. Iar asupra fiecărui set de date se aplica separat același algoritm ca si in cazurile anterioare. In acest caz au fost asumate doar patru topicuri. Se poate observa in figura.10 topicurile discutate in *review*-uri pozitive. Același lucru poate fi observat legat de cele negative in figura.11.

Figura 10

Se poate observa ca produsele cu review-uri pozitive tind sa aibă topicul legat de:

* *Well-fitting small sizes* (Mărimi mici ce se potrivesc bine)
* *Appearance* (aparențe)
* *Casual Style* (stil *casual*)
* *Price to Quality Ratio* (raport preț calitate)

Figura 11

In același timp produsele cu review-uri negative tind sa aibă topicul legat de:

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

Astfel printr-un algoritm de segmentare s-a putut observa perspectiva clientelor atât in privința produselor ce le plac cât și a produselor ce nu le plac.

**Concluzie**

Când este luat in considerare faptul ca majoritatea datelor generate in ziua de azi sunt nestructurate, si o bună parte din aceste date sunt sub forma de text, companiile cat si indivizii se vor confrunta cu mari dificultăți in a încerca sa obțină informații utile din acestea. Însă cel puțin in cazul datelor de tip text exista anumite unelte ce pot ajuta in aceasta privință, si in ciuda dificultăților confruntate pot obține informații valoroase din aceste date.

**References:**

* [cod și images](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)
* [Estimări BM](https://www.slideshare.net/MichaelBeatty/ibm-cloud-storage-cleversafe)