UNIVERSITATEA POLITEHNICA BUCURESTI
FACULTATEA DE AUTOMATICA SI CALCULATOARE
CATEDRA DE AUTOMATICA

TEZA DE DOCTORAT
REZUMAT

SISTEM AUTOMAT DE ANALIZĂ ŞI PRELUCRARE SEMANTICĂ PENTRU TEXTE SCRISE ÎN LIMBA ROMÂNĂ

Autor: Ing. Alexandru Catalin Cosoi

Comisia de Doctorat

Presedinte	Prof.dr.ing.Adina Florea		
Conducator de doctorat	Prof.dr.ing.Valentin Sgarciu		
Referent	Prof.dr.ing.Victor Patriciu		
Referent	Prof.dr.ing.MihaelaOprea		
Referent	Prof.dr.ing.Radu Dobrescu		

Bucureşti

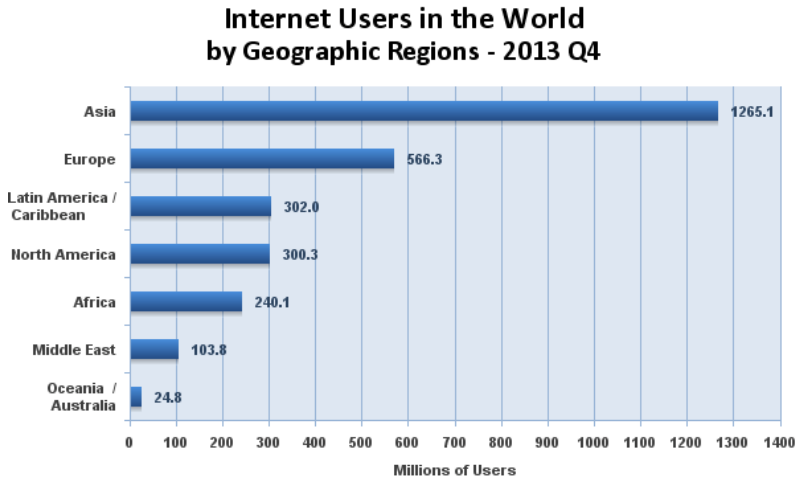
2014

CUPRINS

Introducere

Importanţa aplicaţiilor de prelucrare a limbajului natural a crescut pe măsură ce cantitatea de date disponibile utilizatorilor s-a mărit. Acest lucru se datorează în cea mai mare parte apariţiei şi expansiunii Internet-ului şi a trecerii unui număr din ce în ce mai mare de utilizatori la conexiuni de mare viteză. La sfârşitul anului 2012 erau peste două miliarde de utilizatori de Internet, ceea ce reprezintă 34.3% din populaţia globală.

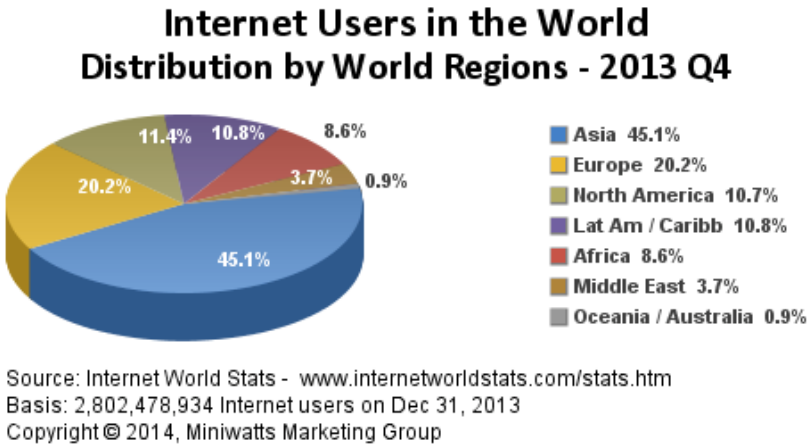
În mod evident, monitorizarea manuală a acestor informaţii este o sarcină imposibilă. Dacă de exemplu am dori să facem un rezumat despre tot ce s-aţ scris în presa online despre un anumit subiect dintr-o anumită zi sau de-a lungul unei săptămâni, ar trebui să extragem mai întâi toate articolele care dezbat subiectul respectiv şi apoi să le citim pe fiecare în parte urmând ca apoi să putem crea rezumatul dorit. Dacă este vorba despre o căutare după un anumit cuvânt, am avea un oarecare avantaj pentru că ar trebui să citim doar articolele care conţin cuvântul respectiv (deşi ne-ar scăpa toate articolele care s-ar referi la acel cuvânt prin sinonime sau metafore). Dacă însă căutarea se face după o anumită idee, atunci ar trebui să parcurgem toate articolele din toate ziarele şi revistele din ziua respectivă sau din săptămâna respectivă şi abia apoi să fim capabili să scriem rezumatul lor.



Aplicaţiile procesării limbajului natural se înscriu în trei mari categorii:

1. aplicaţiile bazate pe text, dintre care amintim:
    a. clasificarea documentelor (şi respectiv găsirea documentelor legate de anumite subiecte);
    b. regăsirea informaţiei (căutarea unor cuvinte-cheie sau concepte);
    c. extragerea informaţiei (legate de un anumit subiect, deci de un anumit cuvânt-cheie);
    d. inţelegerea textelor (care presupune o analiză profundă a structurii acestora);
    e. traducerea automată şi traducerea asistată de calculator dintr-o limbă în alta;
    f. alcătuirea de sinteze;
    g. achiziţia de cunoştinţe.

2. aplicaţiile bazate pe dialog, care implică comunicarea între om şi maşină, aplicaţii cum ar fi sistemele de învăţare, sistemele de interogare şi răspuns la întrebări, rezolvarea problemelor, controlul (bazat pe limba vorbită) al unui calculator, etc.
3. procesarea vorbirii (Este important să facem distincţia între problemele de recunoaștere a vorbirii și cele de inţelegere a limbajului. Astfel, trebuie să remarcăm încă de la început faptul că un sistem de recunoaștere a vorbirii nu folosește nici un element de înţelegere a limbajului. Recunoașterea vorbirii se ocupă numai de identificarea cuvintelor vorbite provenind de la un semnal dat, nu și de înţelegerea mesajului, adica a modului în care aceste cuvinte sunt folosite în procesul de comunicare. Pentru a deveni un sistem de înţelegere a limbajului, un dispozitiv de recunoaștere a vorbirii trebuie să furnizeze intrarea sa unui sistem de înţelegere a limbajului natural, operaţie care produce un așa-numit "sistem de înţelegere a limbajului vorbit". Caracteristica de bază a oricarui sistem de inţelegere este aceea că el realizează o reprezentare a înţelesului propoziţiilor într-un limbaj de reprezentare, care poate fi utilizat în vederea unei procesări ulterioare).



Această lucrare dorește să exploreze principalele tehnici de prelucrare a limbajului natural și posibilitatea utilizării acestora pentru texte scrise în limba română. Un interes deosebit se va arata în **capitolul doi**, metodelor de clasificare semantică precum indexarea semanticii latente sau a grafurilor de context și adaptarea acestora (în funcţie de necesităţi) pentru specificul limbii române.

În termeni matematici, clasificarea semantică este definită ca fiind partiţionarea unui set de documente într-un număr de clase semantice. Aceste clase identifică o mulţime de documente care vorbesc (în cea mai mare parte) despre același subiect. De-a lungul timpului, metodele care au avut cel mai mare succes în utilizare au fost cele statistice, însă limbajul natural nu este un domeniu static, iar sinonimia și polisemia sunt elemente care pot creea probleme metodelor statistice. Un cuvânt poate avea mai multe înţelesuri în același timp iar același concept semantic poate avea mai multe reprezentări. Mai mulţi algoritmi de clasificare semantică vor fi analizaţi în acest capitol, analizându-se rezultatele și defectele fiecăruia.

Capitolul trei va trata problema analizei automate a sentimentelor. Majoritatea textelor disponibile pe Internet pot fi împărţite în doua mari categorii: fapte şi opinii. Faptele sunt expresii obiective despre entităţi şi evenimente şi de asemenea proprietăţile acestora. De cele mai multe ori, opiniile sunt subiective, expresii ce reprezintă sentimentele utilizatorilor, părerile lor despre entităţi, eventimente sau despre proprietăţile acestora. Conceptul de "opinie" este unul foarte larg, însă capitolul trei se va concentra doar pe determinarea sentimentelor pozitive sau negative dintr-un articol. Majoritatea articolelor despre procesarea automată a textelor s-au concentrat pe regăsirea şi extragerea informaţiilor, căutarea inteligentă pe web, clasificarea de texte, clustering şi aşa mai departe, însă puţină atenţie a fost îndreptată spre procesarea opiniilor. Cu toate acestea, opiniile sunt cele mai importante de fiecare dată când trebuie luată o decizie şi dorim să aflăm şi părerile celorlalţi.

Capitolul patru va analiza tehnicile de sumarizare automată. Scopul unui rezumat este simplu şi evident: de a facilita identificarea unui subiect din mai multe articole şi selectarea materialelor cele mai relevante pentru subiectul de interes. În general, crearea unui rezumat implică o oarecare familiaritate cu subiectul dezbătut. Extragerea elementelor cheie pe care autorul articolului a dorit să le sublinieze necesită de regulă antrenament şi experienţă. În general, se face o distincţie clară între scopul unui rezumat şi tipul de rezumat folosit, însă din nefericire, terminologia în acest domeniu este incă insuficient dezvoltată pentru a permite acest lucru. O descriere mai generală împarte rezumatele în două mari subclase: rezumatele indicative, care indică subiectele dezbătute în documente şi pot anunţa cititorul/utilizatorul despre documentele care dezbat un anumit subiect şi rezumate informative care descriu subiectele dezbătute în lucrare.

Capitolul cinci va prezenta partea aplicativă a acestei cercetări si de asemenea va pune la dispoziţie şi un set complex de resurse sintactice şi semantice pentru limba română pentru viitoare cercetări. Ca si aplicaţie demonstrativă, va fi prezentată de asemenea si o platforma de monitorizare media ce va oferi în mod automat aceleaşi servicii ce acum sunt oferite de diferite agenţii de media. Serviciile de monitorizare media implică monitorizarea tuturor canalelor media, împărţirea pe subiecte a articolelor, extragerea rezumatelor (a unuia sau a mai multor articole), posibilitatea de a cauta după anumite cuvinte cheie şi posibilitatea măsurării opiniei generale din articolele respective (dacă este vorba despre un ton pozitiv sau negativ), lucru foarte important pentru companii sau pentru campaniile politice.

Iar ultimul capitol, **capitolul 6**, va prezenta concluziile lucrarii, posibilităţile de dezvoltare ulterioară şi de asemenea contribuţia personală.

De asemenea, resursele sintactice din această lucrare au fost concepute în aşa fel încât să poată fi folosite atât singure, cât şi utilizând framework-uri deja existente, precum NLTK. Scopul acestor resurse este să uşureze munca cercetătorilor români prin oferirea de resurse necesare studiului algoritmilor de procesare a limbajului natural pentru limba română. Resursele vor fi disponibile gratuit pe internet şi vor putea fi accesate de oricine doreşte să continue aprofundarea acestor cunoştinţe prin continuarea rezultatelor prezentate aici.

Clasificare Semantica

În termeni matematici, clasificarea semantică de texte este definită ca fiind partiţionarea unui set de documente într-un număr de clase semantice. Aceste clase identifică o mulţime de documente care vorbesc (în cea mai mare parte) despre acelaşi subiect. De-a lungul timpului, metodele care au avut cel mai mare succes în utilizare au fost cele statistice, însă limbajul natural nu este un domeniu static, iar sinonimia şi polisemia sunt elemente care pot crea probleme metodelor statistice. Un cuvânt poate avea mai multe înţelesuri şi în acelaşi timp acelaşi concept semantic poate avea mai multe reprezentări.

Faţă de sinonimie, polisemia s-a dovedit a fi cea mai problematică. De exemplu, verbul "a juca" poate apărea în numeroase contexte şi poate determina apartenenţa textului sau documentului din care face parte în mai multe clase semantice. Pe de altă parte, dacă două cuvinte sunt sinonime, înseamnă că în interiorul aceleiaşi propoziţii, cuvintele sunt interschimbabile fără ca propoziţia să-şi piardă înţelesul. Însă existenţa mai multor sinonime pentru acelaşi concept poate duce la diminuarea frecvenţelor apariţiei aceluiaşi cuvânt, ducând implicit la modificarea ponderilor în diverse metode statistice.

În acest capitol vor fi tratate principalele tehnici de indexare şi clasificare semantică cum ar fi analiza şi/sau indexarea semanticii latente, TF-IDF (term frequency - inverse document frequency), analiza/indexarea probabilistică a semanticii latente, teorema lui Bayes şi reţelele bayesiene, reţelele semantice, reţelele de context şi deasemenea vor fi prezentate îmbunătăţiri aduse acestor metode observate de-a lungul implementării şi testării.

Analiza Semanticii Latente

Analiza Semanticii Latente (ASL) reprezintă o metodă folosită pentru determinarea înţelesului cuvintelor şi pasajelor prin procesarea unor baze de date de text. ASL crează o reprezentare a cuvintelor şi a submulţimilor de cuvinte folosite - cum ar fi o propoziţie, un paragraf sau un eseu - luat din corpusul iniţial sau nou, într-un spaţiu semantic multidimensional.

Deşi ASL este similară cu modelul reţelelor neuronale, este bazată pe descompunerea valorilor singulare, o tehnică matematică de descompunere a unei matrice, similară analizei factorilor, şi este aplicabilă corpusurilor de text ce se apropie de volumul şi relevanţa limbajului folosit de oameni.

Reprezentarea înţelesului pasajelor şi al cuvintelor derivată din ASL este capabilă să simuleze o varietate de fenomene cognitive umane, de la construirea experimentală a unui vocabular până la categorisirea de cuvinte, corespondenţa dintre o frază şi un cuvânt, înţelegerea unui discurs şi evaluarea calităţii unui eseu. Câteva dintre aceste simulări vor fi prezentate mai jos.

De exemplu, pentru o metodă practică pentru caracterizarea înţelesului cuvintelor, ştim că ASL produce o măsură a relaţiei cuvânt-cuvânt, cuvânt - pasaj şi pasaj - pasaj care este bine corelată cu multiple fenomene umane cognitive ce implică similaritatea semantică. Aceste corelaţii demonstrează asemănarea dintre ceea ce extrage ASL şi

modul în care reprezentarea oamenilor a înţelesului reflectă ceea ce au auzit sau au citit, și de asemenea modul în care reprezentarea înţelesului este reflectată în modul în care autorii textelor își aleg cuvintele. Ca și o consecinţă practică a acestei corespondenţe, ASL permite aproximarea judecăţii umane asupra similarităţii dintre cuvinte și prezicerea obiectivă a consecinţei similarităţii paragrafelor, ambele elemente importante în procesarea de discursuri.

ASL este o metodă statistică automată pentru extragerea și crearea relaţiilor dintre cuvinte și utilizarea acestora în pasaje de text. ASL nu este o metodă tradiţională de procesare a limbajului natural sau un program bazat pe inteligenţă artificială; nu folosește dicţionare construite de către un om, baze de date de cunoștinţe, reţele semantice, gramatici, parsere sintactice, morfologii sau orice altceva din această gamă, și primește ca și date de intrare doar text împărţit în paragrafe sau propoziţii.

Primul pas în aplicarea acestei metode este reprezentarea textului sub forma unei matrice în care fiecare rând reprezintă un cuvânt unic iar fiecare coloană reprezintă un pasaj (sau un document). Fiecare celulă conţine frecvenţa cu care cuvântul de pe linia respectivă se regăsește în pasajul de pe coloana respectivă. Apoi, fiecare celulă este supusă unei transformări, prin care frecvenţa este ponderată printr-o funcţie ce exprimă importanţa cuvântului atât în pasajul analizat cât și pentru analiza unui discurs în general.

Pasul următor constă în descompunerea valorilor singulare ale matricei. Acesta este un caz particular de analiză a factorilor, sau mai bine zis, generalizarea matematică în care analiza factorilor este un caz special. În cazul SVD (singular value decomposition – descompunerea valorilor singulare - DVS), matricea iniţială este descompusă în produsul a trei matrice. O matrice descrie entităţile liniare sub forma unor vectori derivaţi din valorile ortogonale ale factorilor, a doua matrice descrie același lucru dar pe coloane, iar cea de-a treia este o matrice diagonală ce conţine valorile de scalare, astfel încât atunci când se înmulţesc cele trei matrice, se obţine matricea iniţială. Se poate demonstra matematic că orice matrice se poate descompune perfect utilizând nu mai mult decât cea mai mică dimensiune a matricei originale.

Descompunerea valorilor singulare (DVS) este una dintre cele mai puternice metode în algebra liniară. Probleme importante precum calculul rangului unei matrice, determinarea bazelor ortogonale pentru subspaţii liniare, calculul proiecţiilor ortogonale pe subspaţii liniare, problema celor mai mici pătrate, sunt rezolvate cel mai bine folosind DVS.

Descompunerea valorilor singulare ale unei matrice $A \in R^{m*n}$ $ este definită astfel: exista două matrice ortogonale $U \in R^{m*n}$ si $V \in R^{m*n}$ astfel încât:

$$U^T AV = \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$$

unde $\Sigma_1 = diag(\sigma_1, \sigma_2 \dots \dots \sigma_r) \in R^{r*r}$ cu $\sigma_1 \geq \sigma_2 \dots\dots \geq 0$.

Numerele reale positive σ_i sunt numite valori singulare ale lui A și sunt întotdeauna ordonate crescător. Coloanele $u_j \in R^m$ ale matricei ortogonale U se numesc valori singulare la stânga ale matricei A. Coloanele matricei ortogonale V se vor numi

vectori singulari la dreapta ai matricei A. Calculul descompunerii valorilor singulare are la bază următorul rezultat:

Valorile singulare nenule σ_i *ale unei matrice* $A \in R$ *sunt rădăcinile pătrate pozitive ale valorilor proprii nenule ale matricei simetrice pozitiv semidefinite:*

$$B = A^T A$$

i.e. dacă $\lambda_1 \geq \lambda_2 \ldots \ldots \geq 0$ sunt cele r valori proprii nenule ale lui B, atunci:

$$\sigma_1 = \sqrt[2]{\lambda_i}$$

Teorema de mai sus sugerează o procedură pentru calculul valorilor singulare ale unei matrice A, date folosind algoritmul QR simetric pentru calculul valorilor proprii ale matricei $B = A^T A$. Această procedură nu este recomandată datorită posibilei slabe condiţionări numerice a matricei B.

Aplicarea SVD/DVS în căutarea informaţiei a fost iniţial propusă de un grup de cercetători de la Bellcore, respectiv de către Deerwester şi colegii săi în 1990 şi numită în acest context indexare semantică latentă. ASL a fost comparată cu căutarea standard din spaţiul vectorial pe mai multe colecţii de documente. S-a ajuns la concluzia că ASL dă rezultate mai bune, în special în cazul căutărilor în care este necesar să se întoarcă un procent cât mai mare din documentele relevante din colecţie. Pe de altă parte, din cauza falselor apariţii simultane a termenilor, poate duce uneori la o precizie mai scazută.

Rezultatele ASL depind şi de colecţia de documente. Într-o colecţie heterogenă, documentele pot folosi termeni diferiţi pentru a face referire la același subiect, iar ASL poate identifica similaritatea semantică dintre documente ce aparent nu se aseamană. Dar într-o colecţie cu un vocabular omogen, ASL este mai puţin folositoare.

Ca și aplicaţii ale acestei metode, se poate enumera compararea documentelor în spaţiul conceptelor (clasificarea de documente, gruparea datelor, etc), căutarea de documente similare în mai multe limbi, după analiza în prealabil a traducerilor documentelor (cross language retrieval), găsirea de relaţii între termeni/cuvinte (polisemie și sinonimie) și de asemenea, având o mulţime de termeni ce trebuiesc a fi căutaţi, să fie transpuși într-un spaţiu de concepte și apoi să poată căuta documente similare (regăsirea informaţiei – information retrieval)

Unul din dezavantajele majore pe care le are indexarea semanticii latente este că de fiecare dată când se dorește inserarea unui document nou, baza ortogonală se schimbă și este implicit nevoie să se refacă descompunerea în valori singulare. Această operaţie este destul de costisitoare din punct de vedere computaţional și are o complexitate calculată la $O(m^2 n + n^3)$ ceea ce presupune un efort considerabil pentru simpla operaţie de adăugare de documente noi. Acest lucru complică adăugarea de documente noi odată ce au fost deja invăţate un set de documente, însă având în vedere că această metodă nu necesită un calcul al frecvenţelor stabilit apriorii adăugării altor documente, este mult mai ușoară și mai convenabilă învăţarea tuturor

documentelor în același timp. Din acest motiv, LSI este mult mai bine folosită atunci când există o separare inițială clară și concisă iar corpusul de antrenare este stabil și fix.

De asemenea, LSI/LSA a fost o perioadă destul de îndelungată o metodă denigrată deoarece nu ia în considerare ordinea în care apar cuvintele în text și implicit a relaţiilor sintactice sau a morfologiei, ceea ce poate duce la diverse erori în unele situaţii, cu toate ca este capabilă sa extragă înţelesurile corecte în majoritatea situaţiilor.

Un alt dezavantaj al acestei metode este ca a fost patentată de Telcordia, lucru ce a dus la cercetarea și dezvoltarea altor metode a.i. cercetătorii sa-și poata continua munca fără a încălca vreun patent.

Retele de context

Așa cum a fost prezentat mai sus, un pas intermediar în LSA este crearea unei matrice de corespondenţă între termeni (cuvinte) și documente, care este defapt un tabel de căutare ponderată a frecvenţelor tuturor cuvintelor din întreaga colecţie de documente. În LSI, această matrice este interpretată ca un spaţiu vectorial multidimensional.

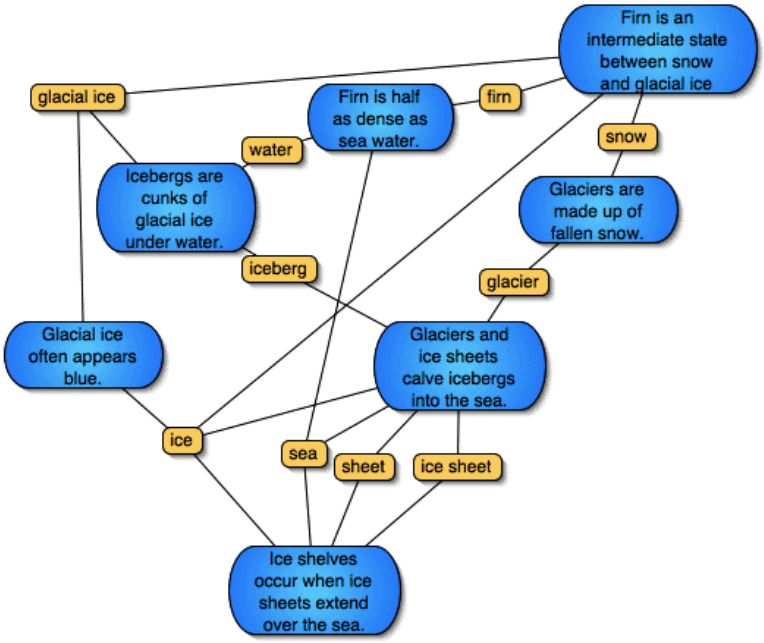
O interpretare alternativă a acestei matrice este utilizarea ei ca și graf bipartit cu noduri termeni sau documente unde fiecare valoare nenulă în matrice corespunde unei muchii ce conectează un nod-termen de un nod-document. În acest model, fiecare document are câte o legatură către toţi termenii conţinuţi în el (către toate cuvintele). Frecvenţele ponderate din matrice corespund ponderilor atribuite muchiilor din graf. Acest graf poartă numele de graf de context (sau contextual).

Exemplu practic:

- Glacial Ice often appears blue
- Glaciers are made up of fallen snow
- Firn is an intermediate state between snow and glacial ice
- Ice shelves occur when ice sheets extend over the sea
- Glaciers and ice sheets calve icebergs into the sea
- Firn is half as dense as sea water
- Icebergs are chunks of glacial ice under water

Nod	Termen	Aparitii
a	Glacial ice	3
b	Ice	5
C	Glacier	2
D	Snow	2
E	Firn	2
F	Ice sheet	2
G	Sea	3
H	Water	2
I	Iceberg	2
J	Sheet	2

Pe baza documentelor din prima lista, se construieşte tabelul de mai sus în care nodurile reprezintă termeni (notate în cazul de mai sus cu a,b,c,d…) iar documentele din care fac parte vor fi notate cu 1,2,3…. De asemenea, se vor păstra şi frecvenţele de apariţie a fiecărui termen cu scopul stabilirii ponderilor în graful bipartit. Această matrice poate fi reprezentată apoi într-un graf de context. Pentru exemplul prezentat mai sus, graful rezultat este prezentat în figura de mai jos:



Cu toate că această metodă este destul de similară cu alte metode din IR (information retrieval) cum ar fi de exemplu grafurile de concept, grafurile (reţelele) de context nu înglobează nici un fel de informaţie gramaticală sau ierarhică privind relaţiile dintre termeni (cuvinte). Structura sa este determinată doar din apariţia listei de cuvinte în colecţia de documente. Fiecare muchie din acest graf are o anumită pondere, care este direct proporţională cu alegerea ponderării globale sau locale utilizată în generarea matriciei de adiacenţă. Singura constrângere legată de ponderarea muchiilor este că acestea trebuie să se regasească în intervalul (0,1).

Putem căuta în colecţia de documente reprezentată prin acest graf prin oferirea unei "energii" nodului de început şi prin oferirea posibilităţii ca această energie să se poată disipa în graf de la nod la nod prin conexiunile sale, urmând câteva reguli simple. Energia totală în orice nod din graf este legată atât de numărul de căi prin care se poate ajunge la acel nod, cât şi prin valoarea ponderilor prin care se ajunge la el. Intuitiv, documentele care au foarte mulţi termeni (foarte multe cuvinte rare) rar vor fi semantic înrudite. De asemenea, se obţin rezultate similare ca în cazul analizei semantice latente, având în vedere că o căutare după un anumit termen se poate termina sau poate trece printr-un document care nu conţine acel termen, dar este suficient de înrudit cu alte documente din listă.

Deoarece energia iniţială se disipă în timp ce se răspândeşte în reţea, (de aici şi constrângerea asupra muchiilor grafului), rezultatele fiecărei căutări sunt localizate în majoritatea cazurilor într-o singură zonă din graf, lucru foarte important pentru performanţa algoritmului, oferindu-i scalabilitate.

În exemplul prezentat mai sus, o căutare după cuvântul “iceberg” ar incepe prin activarea nodului corespunzător acestui cuvant, adica nodul i. Acest nod va primi energia inițială E, care este apoi distribuită nodurilor cu care este conectat, după următorul algoritm:

procedure = energize(energy E, node n_k)
energy(n_k := energy(n_k) + E
E':= E / degree of n_k
if (E' =>T)
for each \= node n_k in N_k
E" := E' * e_{jk}
energize(E", n_j)

Unde N_k este mulțimea tuturor nodurilor vecine lui n_k, e_k este ponderea muchiei ce unește nodurile n_j și n_k , T este o valoare constantă și reprezintă valoarea prag de activare. Se poate observa că acest algoritm este defapt o parcurgere în adancime a grafului. Dacă muchiile ar fi fost sortate în ordinea crescătoare a ponderilor, parcurgerea s-ar fi putut desfășura în mod euristic optimal. Algoritmul poate fi de asemenea modificat să efectueze o parcurgere în lățime.

CNG este cu mult mai rapid decât LSI, este mult mai ușor de înțeles, adăugarea de noi documente în listă nu este atât de complicată ca la LSI și nu necesită efectuarea pașilor de descompunere în valori singulare. De asemenea, modul de reprezentare al datelor este unul foarte intuitiv și ușor de înțeles. Ca și dezavantaje putem menționa faptul că nu există încă suficient de multă documentație și de asemenea, CNG nu se descurcă la fel de ușor cu query-urile complicate.

Term Frequency - Inverse Document Frequency

Algoritmul TF-IDF (term frequency - inverse document frequency) este folosit în general în domenii ca extragerea informațiilor sau text mining. De regulă, este vorba de o pondere determinată statistic ce evaluează importanța unui cuvânt în interiorul unui document dintr-o colecție de documente. Importanța acestuia este crescută direct proporțional cu numărul de apariții într-un document, și este ponderată de frecvența apariției acestui cuvânt în întreg corpusul de documente.

Diferite variații ale acestei metode sunt folosite de regulă de motoarele de căutare ca și metodă principală de ranking a relevanței documentelor pentru categoriile definite de către utilizator. Un exemplu de funcție de ranking (poate cel mai incipient exemplu) este prin însumarea ponderii tf-idf a fiecărui cuvânt din expresia căutată. Foarte multe modele de ranking pornesc de la acest exemplu de bază.

Să presupunem că avem un set de documente în limba română și vrem să căutăm care sunt acele documente cele mai relevante pentru expresia “vacă maro”. O metodă bună de început este eliminarea din analiză a tuturor documentelor ce nu conțin ambele cuvinte. O primă preselecție scoate din analiză documentele inutile, însă ramân în continuare suficiente documente. Următorul pas ar fi numărarea de câte ori apare fiecare cuvânt în fiecare document și însumarea acestor valori. Numărul de apariții ale unui cuvant în interiorul unui document, va purta numele de frecvență a cuvântului. (term frequency). De asemenea, este foarte importantă elimineara cuvintelor puțin

importante. Dacă de exemplu expresia de căutat era “acea vacă maro”, atunci el ne va crea probleme, deoarece “acea” este un cuvânt destul de frecvent. De asemenea, “acea” nu este un cuvânt cheie potrivit cu toate că este foarte des întâlnit în documente, iar “vacă” şi “maro” sunt foarte potrivite, deşi se regăsesc mult mai rar. De aceea, este nevoie şi de un factor care să ţină legătura între frecvenţa cuvintelor şi frecvenţa apariţiei într-o colecţie de documente (inverse document frequency).

Acest *inverse document frequency factor* va micşora ponderea termenlor care se regăsesc foarte frecvent în colecţia de documente şi o va creşte acelor cuvinte care se regăsesc mai rar. Notăm cu tf(i,j) frecvenţa cuvântului I din documentul J.

$$tf(i,j) = \frac{n_{ij}}{\Sigma n_{kj}}$$

unde n_{ij} este numărul de apariţii al cuvântului i în documentul j iar suma respectiva este suma tuturor frecvenţelor cuvintelor din documentul j.

Inverse document frequency factor se defineşte ca logaritmarea raportului dintre numărul total de documente şi numărul de documente în care se regăseşte cuvântul i. Bineînţeles, dacă un cuvânt nu se regăseşte în nici un document, se practică împărţirea la 1 + d(i), însă este o situaţie care nu ar trebui să se întâmple niciodată având în vedere că lista de cuvinte face parte chiar din documentele în care se caută.

$$idf(i) = \frac{|D|}{(|d|t_i \in D|\square)}$$

Ponderea finală se obţine prin înmulţirea celor doi factori:

$$tfidf_{ij} = tf_{ij} * idf_i$$

Regresie Logistică

În statistică, regresia logistică (câteodată numit model logistic sau model logit) este utilizată pentru estimarea probabilităţii apariţiei unui eveniment, prin aşezarea datelor într-o curbă logistică. Ca multe alte metode de regresie, foloseşte multe variabile predictoare, atât numerice cât şi categoriale. De exemplu, probabilitatea ca o persoană să facă un atac de cord într-o anumită perioadă de timp ar putea fi prezisă daca s-ar cunoaşte vârsta, sexul şi masa corporală a persoanei.

$$f(z) = \frac{1}{1 + e^{-z}}$$

Dacă z ar reprezenta un set de factori de risc, atunci f(z) ar reprezenta probabilitatea unui anumit rezultat, având în vedere factorii de risc z.

Această funcţie primeşte valori între minus infinit şi infinit şi întoarce valori între 0 şi 1 sub forma de probabilităţi. Dacă valoarea întoarsă este mai mare decât o anumită valoare prag, atunci se poate considera că documentul aparţine de categoria B. Algoritmul necesită un vocabular, şi fiecărui cuvânt din vocabular îi este asociată o pondere.

Algoritmul poate de asemenea învăţa atât în timpul testării cât şi înainte de testare fiind o fază specială de antrenare. La fel ca şi în cazul celorlalte metode mai puţin LSA şi CNG, algoritmul nu poate determina relaţiile semantice între cuvinte, ci doar similaritatea între documente pe baza stasticilor apariţiilor cuvintelor fiind încadrat în categoria algoritmilor de tip bag of words.

Teorema lui Bayes

Thomas Bayes (1702-1761) a fost un matematician britanic şi în acelaşi timp un pastor prezbiterian, cunoscut a fi formulat o teoremă care îi poartă şi numele: Teorema lui Bayes (adesea numită legea lui Bayes), publicată post-mortem sub forma unui eseu *Essay Towards Solving a Problem in the Doctrine of Chances (1764)*, de către prietenul său Richard Price în *"Philosophical Transactions of the Royal Society of London*. Teorema lui Bayes este una din teoremele fundamentale ale teoriei probabilităţii, care determină probabilitatea apartenenţei evenimentelor şi a obiectelor la o anumită grupă. În cazul filtrelor spam bazate pe teorema lui Bayes de exemplu (numite şi filtre bayesiene), pentru determinarea probabilităţii apartenenţei unui anumit mesaj la spam, sunt utilizate dicţionarele create în timpul "învăţării" filtrului. De regulă programul "învaţă" analizând arhivele de email-uri, selectate în prealabil manual. Când dicţionarele sunt create definitiv, probabilitatea apartenenţei unui nou mesaj la spam este calculată prin normalizarea şi însumarea probabilităţii fiecărui cuvânt în parte. Prin urmare, adunând informaţii statistice despre rata de apariţie a unor diferite cuvinte şi structuri în mesajele de tip spam sau în mesajele legitime, filtrul compară apoi noile mesaje cu aceste modele şi le clasifică corespunzător. Formula de calcul al probabilităţilor în cazul filtrării antispam poate fi vazută mai jos:

$$PR(S|W) = \frac{PR(W|S) * PR(S)}{PR(W|S) * PR(S) + PR(W|H) * PR(H)}$$

În mod evident, clasificarea bayesiană se poate face pe mai mult de două categorii. În "Introducing Syntactic Features into a Bayesian Classifier" in 2008, am arătat că în funcţie de numărul de categorii şi de cantitatea de text din fiecare document, rezultatele obţinute în urma unei clasificari bayesiane sunt în general bune sau foarte bune, însă dacă se introduc de asemenea şi informaţiile sintactice pentru fiecare cuvânt, clasificarea se poate îmbunătăţi cu până la 10% (bineînţeles, acolo unde rata de clasificare permite acest lucru).

Sistemul care s-a folosit în lucrarea respectivă a constat într-un filtru bayesian clasic şi algoritmul Link Grammar pentru determinarea părţilor sintactice din propoziţiile analizate. Informaţia sintactică a fost adaugată prin aducerea cuvintelor într-o formă canonică de forma “parte sintactică_cuvânt”. Testele s-au efectuat pe un total de 25 de categorii semantice, formate din diferite domenii şi subdomenii ştiinţifice şi conţinând câteva zeci de exemple. În mod evident, erorile de clasificare au fost mult mai mari în cazul subdomeniilor dintr-un acelaşi domeniu ştiinţific comparat cu documentele din domenii diferite, însă introducerea părţilor sintactice a adus îmbunătăţiri în ambele situaţii. Ca şi dezavantaje, timpul de antrenare s-a mărit iar baza de date de cuvinte s-a dublat, lucru de aşteptat, având în vedere canonizarea cuvintelor.

O alta metodă de filtrare bayesiană constă în utilizarea reţelelor bayesiene. Modelul reţelelor Bayesiene, introdus de Judea Pearl [1988], porneşte de la modelul probabilistic Bayesian, dar elimină numărul enorm de calcule necesare prin

considerarea caracteristicilor de modularitate și de cauzalitate ale domeniului problemei. Ideea de bază a reţelelor Bayesiene este aceea că, pentru a descrie domeniul problemei, nu este necesar să se considere probabilităţile tuturor perechilor de evenimente (fapte) posibile. Cele mai multe evenimente sunt independente între ele şi interacţiunile dintre acestea nu trebuie considerate, deoarece nu există. Reţele Bayesiene, numite și reţele de încredere aparţin familiei modelelor grafice probabilistice. Modelele grafice reprezintă grafuri în care nodurile sunt variabile aleatoare și arcurile sunt condiţii de propagare. Aceste structuri sunt utilizate pentru reprezentarea cunoașterii despre un anumit domeniu incert. În mod particular, fiecare nod dintr-un asemenea graf reprezintă o variabilă aleatoare, în timp ce liniile dintre noduri reprezintă dependenţe probabilistice între variabilele aleatoare corespondente. Aceste dependenţe condiţionale din graf sunt în general estimate folosind metode statistice de calcul. Deci, reţelele Bayesiene combină principii din teoria grafurilor, teoria probabilităţilor, stiinţa calculatoarelor și din statistică.

O reţea baeyesiană poate fi folosită pentru predicţia sau analiza problemelor din lumea reală şi din sistemele naturale complexe, unde corelaţiile statistice pot fi găsite între variabile sau aproximate folosind opiniile experţilor. Aceste reţele au o vastă arie de aplicabilitate, cum ar fi ajutorul în luarea deciziilor în domenii ca medicină, inginerie, resurse naturale și managementul decizional.

Ca şi aplicativitate în clasificarea semantică exemplele sunt multiple iar construcţia depinde foarte mult despre ce problema este vorba. O subproblemă care poate fi rezolvată cu ajutorul reţelelor bayesiene este împarţirea documentelor intr-un set de categorii prestabilit. Un alt exemplu este prezentat in unde este descris un sistem bazat pe reţele bayesiene capabil sa poarte conversaţii și să descifreze expresii complexe. De asemenea, o dată ce sunt parcurse noţiuni precum "analiza semanticii latente" sau "grafurile de context", reţelele bayesiene reprezinta incă o alternativă ușor de înţeles pentru aceste metode.

Retele Semantice

O reţea semantică utilizează grafuri pentru reprezentarea cunoștinţelor. Conceptul de reţea semantică a fost introdus aproximativ în perioada 1965-1970. Scopul primelor reţele semantice a fost reprezentarea cuvintelor din limbajul natural. O reţea semantică are forma unui ansamblu de noduri și arce, orientate și etichetate. Nodurile reprezintă obiectele. Un obiect poate fi un concept abstract sau particular, un atribut, ș.a.m.d. Arcele sunt utilizate pentru a reprezenta legăturile care există între aceste obiecte.

Reţelele semantice sunt reprezentări geometrice utilizate pentru reprezentarea cunoştinţelor şi organizarea lor în noduri şi conexiuni. Implementările pe calculator a reţelelor semantice au apărut mai întâi pentru diverse aplicaţii de inteligenţă artificială și traducere automată, dar chiar și implementări mai rudimentare au fost folosite foarte mult timp în filozofie, psihologie și lingvistică.

Câteva din reţelele semantice prezentate mai sus au fost dezvoltate în mod explicit să implementeze câteva din sistemele cognitive umane, însă altele au avut ca scop doar îmbunătăţirea performanţelor diverșilor algoritmi de procesare a limbajului natural. Însă, o dată cu apariţia reţelelor, multe din soluţiile computaţionale au fost apropiate

de cele care implementau sisteme cognitive. De exemplu, diferenţa între reţelele de definire şi cele aserţionale este similară cu distincţia între memoria semantică şi memoria episodică.

Concluzii parţiale şi Adaptarea la specificul limbii române

Majoritatea tehnicilor prezentate în acest capitol sunt în general statistice, aşa că sunt destul de puţine îmbunătăţiri care pot fi aduse algoritmilor existenţi. Rezultatele sunt în general la fel de bune dacă corpusul de antrenare este la fel de bun şi la fel de complex, ceea ce ne face să concluzionam că dacă reuşim să avem un corpus de antrenare exhaustiv, atunci implicit şi rezultatele vor fi pe măsură.

De asemenea, aşa cum se poate observa şi din exemplele de mai sus, gradul de complexitate şi necesarul de resurse este direct proporţional cu numărul de cuvinte din corpusul de antrenare şi în mod implicit cu dimensiunea corpusului. De asemenea, în funcţie de timpul la care se face referire sau în care se petrece acţiunea oferă un grad de variaţie destul de larg atât verbelor cât şi celorlalte cuvinte prin variaţia diacriticelor, a genului si a persoanei. De exemplu, variaţia [acest, această, acesta, aceştia, acestea, acești, aceste], cu toate că reprezintă mereu acelaşi lucru (un substitut pentru un anumit obiect/persoană), variaza în funcţie de număr şi sex.

În funcţie de problema care se dorește a fi rezolvată, se pot efectua diferite operaţii de preprocesare. Aşa cum am văzut în "Introducing Syntactic Features into a Bayesian Classifier" inclusiv rolul sintactic al fiecărui cuvânt poate avea un rol esenţial în clasificarea documentelor.

Când vorbim de clasificarea semantică şi a modului în care poate fi ea aplicată, trebuie să fim foarte atenţi la problema pe care trebuie să o rezolvăm şi abia pe urmă să decidem ce algoritm se potrivește mai bine, sau ce corpus de antrenare este necesar. De exemplu, majoritatea algoritmilor prezentaţi mai sus se bazează pe un model de învăţare supervizată, care în mod evident duce la rezultate mult mai bune decât în cazul unei antrenări nesupervizate, însă la cantitatea din ziua de astăzi de documentaţie şi de informaţii, problema pregătirii corpusului de antrenare se complică exponential.

Să luăm următorul exemplu: dorim să creăm un sistem automat de monitorizare media, care este capabil să:

- Preia din presa online toate articolele cu o anumita frecvenţă prestabilită. Statistic, frecvenţa optimă este din oră în oră
- Cum multe ziare vor acoperi acelaşi subiect, sistemul va trebui să fie capabil să creeze clustere de articole care tratează acelaşi topic/categorie
- Majoritatea publicaţiilor oferă o preclasificare a articolelor în categorii precum economic, politic, etc, însă acest lucru este insuficient în cadrul unui sistem de monitorizare media, deoarece se dorește o filtrare la nivel de topic și nu la nivel de categorie
- Pentru fiecare topic, sistemul este capabil să determine daca este vorba despre un topic pozitiv sau negativ. Acest lucru este foarte important în cazul review-ului de produse, al politicii, etc

- De asemenea, pentru fiecare topic în parte, pe baza articolelor din ziare diferite care ating același subiect, să se creeze rezumatul

La volatilitatea știrilor din ziua de azi, crearea unui corpus de antrenare la nivel de categorie (politic, financiar, tabloid, etc) este un lucru trivial, însă nu se poate spune același lucru si despre clasificarea la nivel de topic.

O bună metodă de creare a unui corpus de antrenare la nivel de topic s-a dovedit a fi Similaritatea Jaccard, care cu toate că pare destul de simplă, s-a dovedit în testele mele, în special pe zona clasificării articolelor din presa online, a avea rezultate foarte bune.

Distanţa Jaccard dintre două documente este: $d = \frac{|A \cap B|}{|A \cup B|}$, unde A este mulţimea cuvintelor din primul document iar B este mulţimea cuvintelor din al doilea document. Rezultatele sunt în general foarte bune, dar în capitolul 5 vom arata diferite metode de îmbunătăţire, prin adaugarea de semidistanţe Jaccard (numărul de apariţii al fiecărui cuvânt, dimensiunea documentelor, etc), care ne vor duce gradul de acurateţe care în mod normal la distanţa Jaccard clasică este undeva între 70-80 la sută. în cazul testelor mele, în gama 98-100 la sută. Însă și distanţa Jaccard clasică poate fi folosită daca dorim doar crearea unui corpus de antrenare.

Analiza Sentimentelor

Majoritatea textelor disponibile pe Internet pot fi împărţite în două mari categorii: fapte și opinii. Faptele sunt expresii obiective despre entităţi și evenimente și de asemenea proprietăţile acestora. De cele mai multe ori, opiniile sunt subiective, expresii ce reprezintă sentimentele utilizatorilor, părerile lor despre entităţi, evenimente sau despre proprietăţile acestora. Conceptul de "opinie" este unul foarte larg, însă acest capitol se va concentra doar pe determinarea sentimentelor pozitive sau negative dintr-un articol. Majoritatea articolelor despre procesarea automată a textelor s-au concentrat pe regăsirea și extragerea informaţiilor, căutarea inteligentă pe web, clasificarea de texte, clustering și așa mai departe, însă puţină atenţie a fost îndreptată spre procesarea opiniilor. Cu toate acestea, opiniile sunt cele mai importante de fiecare dată când trebuie luată o decizie și dorim să aflăm și părerile celorlalţi. Acest lucru este adevărat atât pentru indivizi, cât și pentru organizaţii.

Unul din multele motive pentru lipsa studiului opiniilor este că o mare perioadă de timp a fost foarte puţin text ce exprima opiniile oamenilor disponibil pe Internet. Inainte de utilizarea web-ului la o scara largă, în momentul în care cineva trebuia să ia o decizie, și-ar fi întrebat prietenii sau familia. Când o organizaţie dorește să facă acest lucru, va comanda sondaje și cercetări de piaţă. Dacă luăm în considerare creșterea explozivă a Internetului și a cantităţii de conţinut generat de către utilizatori în ultimii ani, atunci putem să concluzionăm că lumea s-a schimbat. Internetul a schimbat modul în care oamenii își exprimă opiniile și părerile. Acum își pot expune părerile direct pe site-urile vânzătorilor și își pot prezenta întreaga experienţă pe care au avut-o cu produsul respectiv, astfel încât următorii cumpărători să fie avizaţi. Această nouă lume oferă o sursă nemărginită de informaţii ce poate prezenta beneficii substanţiale unei game foarte largi de utilizatori. Acum nu mai este nevoie să suni un

prieten dacă ai nevoie de un sfat în vederea achiziţionării unui produs ci poţi avea acces la informaţiile dorite imediat.

Cu toate acestea, găsirea de surse de opinii valide şi monitorizarea lor pe internet încă poate cauza numeroase probleme datorită numărului foarte mare de surse, iar fiecare dintre acestea poate avea un număr foarte mare de opinii. De cele mai multe ori, opiniile sunt ascunse in discuţii lungi pe blog-uri sau forum-uri. Este destul de dificil pentru o persoană să găsească surse relevante, să extragă propoziţiile şi frazele ce conţin opinii, să le citească, să le sumarizeze şi să le organizeze într-o formă utilizabilă. În mod evident, necesitatea unor sisteme ce pot extrage singure opiniile din texte și să sumarizeze aceste opinii, este destul de mare. Analiza sentimentelor, cunoscută de altfel şi ca "mineritul sentimentelor", s-a născut din această necesitate. Având în vedere aplicaţiile multiple pe care le poate avea, atât în mediul industrial cât şi în cel academic, acest domeniu a explodat în ultimii ani. La momentul actual, există cel puţin 20-30 de companii care oferă analiza sentimentelor pentru industrie.

Clasificarea sentimentelor sau clasificarea sentimentelor la nivel de document încearcă să stabilească o părere generala asupra sentimentelor ce reies din text cu privire la subiectul dezbătut. Dacă de exemplu ne gândim la evaluarea unui produs, o astfel de aplicaţie ar determina în mod automat daca este o evaluare pozitivă sau negativă. Celălalt pas se referă la identificarea unei opinii la nivel de propoziţie şi stabilirea tipului de opinie, respectiv negativă sau pozitivă.

Analiza sentimentelor bazată pe caracteristici: acest model caută obiectele asupra cărora sunt exprimate sentimentele dintr-o propoziţie, urmând ca apoi să se determine daca sunt pozitive sau negative. Un obiect poate fi reprezentat printr-un produs, un serviciu, un individ, o organizaţie, un eveniment, un subiect, etc. De exemplu, în cazul unei evaluări de produs, se poate determina la ce componente ale produsului se face referire în cadrul evaluării şi dacă exprimă o părere pozitivă sau negativă. De exemplu, propoziţia "durata de viaţă a bateriei acestei camere este prea scurtă" este un comentariu ce se referă la durata de viaţă a bateriei unei camere, iar comentariul este unul negativ. Acest gen de aplicaţii sunt necesare în foarte multe situaţii: de exemplu, ca o companie să-şi îmbunătăţească produsele, trebuie să fie la curent cu toate evaluările produselor curente.

Evaluarea unui obiect se poate face prin două metode principale: direct sau prin comparaţie. Analiza directă sau opiniile directe oferă în mod direct sentimentele negative sau pozitive asupra subiectului dezbătut, fără a fi necesară menţionarea altor obiecte. Analiza prin comparaţie înseamnă comparaţia obiectului analizat cu alte obiecte similare, eventual concurente. De exemplu, următoarea propoziţie reprezintă o opinie directă: "calitatea acestei camere este slabă", iar următoarea propoziţie reprezintă o opinie pozitivă exprimată prin comparaţie: "Calitatea pozelor este mai bună decât a celor făcute cu camera x". Este evidentă necesitatea determinării acestor tipuri de propoziţii şi situarea produsului analizat în comparaţie cu concurenţa.

Așa cum este ușor de înţeles de ce opiniile și sentimentele sunt foarte importante de monitorizat, la fel de ușor de înţeles este și de ce este foarte important pentru companii să răspândească opinii pozitive în propriul interes, create sau nu de oameni. Acest fenomen se numeşe "opinion spam" și de regulă se traduce prin oferirea de comentarii negative produselor concurenţei și oferirea de opinii pozitive produselor

proprii. Acest gen de opinii sunt inutile pentru un evaluator și este foarte importantă eliminarea acestora din lista opiniilor ce vor ajunge la utilizator.

Sentiment Analysis (Analiza sentimentelor) sau Opinion Mining (mineritul opiniilor) este studiul computaţional al opiniilor, sentimentelor și emoţiilor exprimate într-un text. De exemplu, în următoarea evaluare, se prezintă un set de păreri personale despre iPhone:

"Mi-am cumpărat noul iPhone acum câteva zile. Este un telefon foarte bun. Ecranul tactil este foarte incitant. De altfel, inclusiv calitatea vocii este foarte bună. Cu toate că bateria nu rezistă foarte mult, nu mă deranjează foarte tare lucrul acesta. Cu toate acestea, mama s-a suparat pe mine ca nu i-am spus înainte sa-l cumpăr. I s-a părut scump și voia să ne întoarcem la magazin sa-l returnăm."

Putem observa că există atât păreri pozitive cât și negative. De asemenea, se poate remarca că fiecare propoziţie face referire la un anumit obiect despre care este enunţată o opinie. Se mai poate observa și că în textul de mai sus, majoritatea opiniilor sunt pozitive, iar cele care sunt negative sunt atribuite altei persoane.

În general, opiniile pot fi exprimate în legătură cu orice (un produs, un serviciu, un individ, etc). Fiecare obiect poate avea la rândul său un set de atribute (sau proprietăţi, iar daca obiectul respectiv este compus din mai multe componente, atunci fiecare componentă poate avea la rândul său un set de atribute sau proprietăţi. In mod evident, această relaţie poate continua. Se poate concluziona că un obiect poate fi descompus într-un arbore de componente, subcomponente și proprietăţi ale acestora. O definiţie corectă ar fi: "un obiect este o entitate care poate fi un produs, o persoană, un eveniment, o organizaţie, sau un subiect. Obiectului îi este asociată o pereche (T,A) unde T este o ierarhie de componente, subcomponente, s.a..m.d iar A este un set de atribute." În exemplul de mai sus, telefonul este obiectul iar câteva din componentele sale sunt bateria și ecranul. De asemenea, setul adiacent de atribute este format din calitatea vocii, mărime și greutate. Bateria are propriul set de atribute: durata de viaţă și dimensiunile. Astfel, un obiect poate fi reprezentat sub formă arborescentă ce are ca rădăcină obiectul în sine. Fiecare nod care nu este rădăcină este ori o componentă ori o subcomponentă a obiectului iar fiecare muchie este relaţia dintre acestea. De asemenea, fiecare nod are asociat un set de atribute. O opinie se poate lega de orice componentă din acest arbore.

Definirea opiniilor explicite si implicite

O opinie explicită referitoare la caracteristica *f* este o opinie exprimată vis a vis de *f* într-o propoziţie subiectivă. O opinie implicită referitoare la caracteristica *f* reprezintă o opinie asupra *f* subînţeleasă în cadrul unei propoziţii obiective. Propoziţia următoare ilustrează exemplul clasic al unei opinii pozitive explicite: *Calitatea sunetului pentru telefonul acesta este extraordinara.* Comparativ, această propoziţie ne oferă un exemplul unei opinii negative implicite: "Difuzorul telefonului s-a defectat in doua zile". Astfel, deși această propoziţie prezintă o realitate obiectivă, implicit indică și o opinie negativă referitoare la difuzorul telefonului.

Definirea propoziţiei în care sunt exprimate opinii: O propoziţie în care sunt exprimate opinii este o propoziţie care exprimă în mod explicit sau implicit opinii pozitive sau negative. Aceasta poate fi o propoziţie de natură subiectivă sau obiectivă.

Asa cum se poate observa, noţiunea de propoziţie subiectivă şi cea de propoziţie în care sunt exprimate opinii nu se suprapun perfect, deşi în multe cazuri, propoziţiile în care sunt exprimate opinii reprezintă de fapt o subcategorie a propoziţiilor subiective.

Abordările utilizate pentru clasificarea acestor două tipuri de propoziţii este totodată similară. Metoda prin care se determină daca o propoziţie este de natură subiectivă sau obiectivă poartă denumirea de clasificare a subiectivităţii.

În mod evident şi documentele pot fi evaluate în ce măsură prezintă sau nu opinii. Până acum, am considerat din start documentele ca fiind unele în care sunt prezentate opinii. În practică însă trebuie stabilită şi această caracteristică a documentului. De exemplu, forumurile de discuţii dispun şi de numeroase pasaje cu intrebări şi răspunsuri care nu conţin opinii. Firul logic pe care îl vom aplica mai departe consideră că un document conţine propoziţii în care sunt exprimate opinii, atunci şi documentul respectiv poate fi considerat drept unul care prezintă această caracteristică. Această definiţie nu este totuşi exhaustivă, deoarece există şi cazuri precum o ştire, în care sunt preluate opiniile altor persoane însă numai cu titlu de informaţie. Nu ar fi în intregime corect să catalogăm documentul respectiv drept unul subiectiv.

Cel mai corect în cazul de faţă ar fi să conchidem că documentul respectiv prezintă într-o manieră obiectivă o serie de opinii care aparţin unei alte persoane, alta decât autorul documentului. O definiţie mai clară a tipologiei documentelor poate fi realizată pe baza intenţiei autorului, respectiv dacă acesta intenţionează sau nu să exprime o serie de opinii proprii prin intermediul textului. Evaluările de produse corespund acestei categorii de documente. Dacă o propoziţie conţine sau nu opinii personale este mai uşor de stabilit, în timp ce într-un text obişnuit se regăsesc atât propoziţii care conţin opinii cât şi altele fără.

Prin acest lucru se dovedeşte cât de necesară este aplicarea unei forme de sumarizare asupra rezultatelor obţinute în urma procesului de căutare (“mining”), deoarece valoarea adăugată oferită de o simplă listare a opiniilor identificate şi a elementelor acestora este minimă pentru utilizator. În vederea atingerii obiectivului principal şi anume acela de a transforma limbajul natural nestructurat în date structurate, ne putem folosi cu succes de evidenţierea rezultatelor sub forma celor 5 elemente esentiale. Datele astfel descompuse pot fi păstrate în baze de date ce permit interogări ulterioare. În cazul bazelor de date astfel construite se pot aplica tehnici de vizualizare pentru a putea beneficia de rezultate în formatul cel mai elocvent extragerii de informaţii relevante, care de obicei se numesc sumarizări structurate şi prezintă datele sub forma de grafice.

Sumarizarea struturată bazată pe opinii reprezintă o metodă simplă de a produce o sumarizare la nivel de caracteritici ale opiniilor la nivelul unui obiect sau a unor obiecte aflate în competiţie/opoziţie.
Presupunând că dorim să realizăm sumarizarea unei evaluării pentru un telefon mobil, denumit în continuare Telefon Mobil 1, rezumatul va arăta aşa cum este ilustrat în tabelul de mai sus. În cadrul figurii, cuvântul “telefon” reprezintă telefonul în sine

(rădăcina reprezentării ierarhice a obiectului). 125 de evaluari erau de natură pozitivă, în timp ce doar 7 erau de natură negativă cu privire la performanţa telefonului.

Calitatea sunetului şi mărimea telefonului sunt două caracteristici ale obiectului la care s-a făcut referire, prima dintre ele primind 120 de evaluări pozitive şi doar 8 negative. Propoziţiile care conţin evaluări venite de la o singură persoană pot fi corelate cu alte propoziţii sau chiar cu un text de evaluare, care oferă comentarii pozitive sau negative vis a vis de o caracteristică a produsului. Folosind acest beneficiu, clienţii pot afla relativ uşor care este părerea clienţilor actuali faţă de produsul în cauză. Dacă persoana care citește evaluarea este interesată într-o masura mai mare de evaluarea unei anumite caracteristici ale produsului, atunci ea poate urmări link-ul din cadrul evaluării individuale pentru a afla mai multe păreri şi de la alti clienţi. Un astfel de rezumat al informaţiei poate fi vizualizat folosind graficele de tip bara.

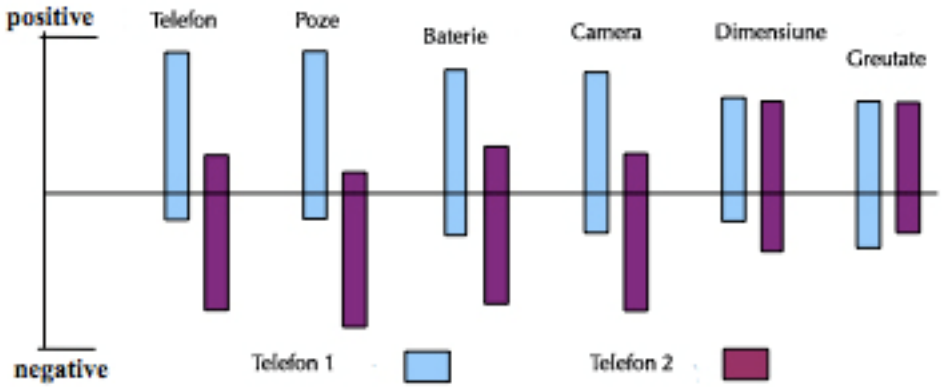


Figura de mai sus ilustrează un rezumat al opiniilor în cadrul unei evaluări pentru un telefon mobil. Fiecare bară situată deasupra axei X mediane indică numărul de opinii pozitive înregistrate de către caracteristica respectivă a produsului, iar bara situată dedesuptul axei X ne arată câte opinii negative s-au înregistrat cu privire la aceeaşi caracteristică. Cu siguranţă, putem folosi şi alte metode de vizualizare, cum ar fi prezentarea procentul de opinii pozitive, cuantumul opiniilor negative fiind calculat drept unitatea minus procentul de opinii pozitive.

Faţă de telefon în sine, majoritatea persoanelor şi-au exprimat o părere pozitivă faţă de primul dintre produse şi una de natură negativă faţă de cel de-al doilea telefon. După cum se poate observa, vizualizarea permite utilizatorilor să observe cum performează un telefon faţă de un competitor de-a lungul unei liste de caracteristici. În afară de aceste tehnici de vizualizare prezentate sunt posibile multe alte tipuri particularizate în funcţie de obiectivele urmărite sau de contextul în care se doresc a fi folosite. Totodată, şi sumarizarea care nu vizează texte în care sunt exprimate opinii poate fi foarte folositoare, după cum se va putea observa din exemplele urmatoare.

Sumarizarea bazată pe apariţiile unei anumite caracteristici (“feature buzz”) se referă la frecevenţa relativă a menţiuniilor cu privire la o anumită caracteristică a obiectului. Această tehnică poate indica managementului unei companii care sunt cu adevarat aspectele importante care îi preocupă pe clienţii lor. De exemplu, în cadrul unui studiu asupra mediului online bancar, caracteristica cea mai des menţionată de către clienţi a fost securitatea tranzacţiilor. Sumarizarea bazată pe apariţiile unui anumit obiect

(“object buzz”) vizează cuantificarea frecvenţei de menţionare a diferite produse sau mărci concurente. Această metodă este utilă deoarece oferă informaţii privind gradul de popularitate al produselor sau mărcilor respective la nivelul pieţei în care activează.

De asemenea, având în vedere că momentul când s-a realizat exprimarea opiniei este cuantificat ca fiind unul dintre cele cinci elemente de bază, o altă analiză care poate fi implementată se referă la monitorizarea în timp a schimbarilor survenite la nivelul fiecărui aspect. Monitorizarea trendurilor (“trend tracking”) poate fi obţinută prin adăugarea dimensiunii temporale în cadrul analizelor de mai sus.

Raportarea ce include și monitorizarea trendurilor este extrem de utilă pentru utilizatorii finali, deoarece le permite să fie la curent cu schimbările survenite în timp. Toate aceste rezumate pot fi realizate și vizualizate extrem de ușor deoarece sunt rezultatele interogării unei baze de date fără a mai fi nevoie de acţiuni suplimentare, dovedind încă o dată valoarea adusă de structurarea informaţiei care conţine opinii în funcţie de cele cinci elemente prestabilite.

Clasificarea subiectivităţii și a sentimentelor

Atunci când discutăm despre analiza sentimentelor, clasificarea este procesul la care se va face referire cu precădere. Astfel, clasificarea sentimentelor vizează crearea unei diferenţieri la nivelul documentelor în care sunt exprimate opinii în funcţie de tipul opiniei: pozitivă sau negativă. Această activitate este cunoscută și sub numele de clasificare a sentimentelor la nivelul documentelor, deoarece ia în considerare întreg documentul drept unitate de bază pentru analiză. Cercetarea existentă în domeniu pornește de la premisa ca documentul este deja considerat ca fiind unul în care sunt exprimate opinii. În mod firesc, același procedeu de clasificare a sentimentelor poate fi aplicat și la nivelul propoziţiilor, însă în această situaţie literatura de specialitate nu consideră ca fiecare propoziţie drept una în care sunt exprimate opinii. Metoda prin care se stabilește dacă într-o propoziţie sunt exprimate sau nu opinii, poartă denumirea de clasificare a subiectivităţii, pentru că ulterior, propoziţiile despre care se consideră că exprimă opinii să fie clasificate în funcţie de tipul opiniei exprimate, pozitivă sau negativă, folosind metoda de clasificare a sentimentelor la nivelul propoziţiei.

Analiza sentimentelor bazată pe caracteristici

Deși clasificarea textelor în care sunt exprimate opinii la nivelul documentului sau a propoziţiei se dovedește a fi utilă în multe cazuri, aceasta clasificare nu oferă gradul de detaliere necesar în cazul unor alte aplicaţii. Caracterizarea unui document drept unul în care sunt exprimate opinii pozitive vis a vis de un obiect evaluat, nu înseamnă că autorul documentului are o părere pozitivă despre toate trăsăturile obiectului respectiv. Aceeași regulă aplicându-se și în cazul în care opinia generală desprinsă la nivelul textului este una negativă. În general, în cazul unui text în care sunt exprimate opinii cu privire la un obiect, persoană, serviciu, etc, autorul include atât referinţe pozitive sau negative cu privire la trăsăturile obiectului, deși opinia desprinsă la nivel general poate să fie ori pozitivă, ori negativă.

Pe de altă parte, atunci când vine vorba de identificarea obiectului evaluat, această informaţie este de o reală valoare și în cazul textelor informale, așa cum sunt cele disponibile online. Deși în general atunci când se face evaluarea unui produs se cunoaște obiectul la care se face referire, de-a lungul evaluării există și posibilitatea de a se face referi la alte obiecte pentru comparaţii, ceea ce îngreunează relativ procesul de identificare a tuturor obiectelor relevante pentru textul respectiv.

În ceea ce privește extragerea cu exactitate a perioadei în care a fost redactat textul, această informaţie este mai ușor de reperat în cazul informaţiilor postate online, unde orice postare este înregistrată și se menţionează ora exactă la care aceasta a devenit disponibilă publicului, însă destul de dificilă în cazul altor tipuri de documente. Modalitatea prin care sunt extrase aceste trei tipuri de informaţii la nivelul unui text, respectiv persoana care emite opinia, obiectul sau obiectele la care se fac referire și perioada la care s-a expus opinia, poarta denumirea de Named Entity Recognition (NER).

Una dintre resursele ce va fi disponibilă pe site este o listă, în format XML, ce va conţine un set de obiecte, și pentru fiecare obiect, un set de criterii ce îl definesc. Exemplu:

```
<title>Obiecte și componente</title>
<continut>

<obiect>
        <nume>telefon</nume>
        <caracteristica>ecran</caracteristica>
        <caracteristica>baterie</caracteristica>
        <caracteristica>greutate</caracteristica>
        <caracteristica>dimensiune</caracteristica>
</obiect>

<obiect>
        <nume>baterie</nume>
        <caracteristica>greutate</caracteristica>
        <caracteristica>durată de viaţă</caracteristica>
</obiect>

<obiect>
        <nume>laptop</nume>
        <caracteristica>ecran|display</caracteristica>
        <caracteristica>greutate</caracteristica>
        <caracteristica>fiabilitate</caracteristica>
        <caracteristica>dimensiuni</caracteristica>
        <caracteristica>RAM</caracteristica>
        <caracteristica>hard|hard-disk|HDD</caracteristica>
        <caracteristica>procesor</caracteristica>
        <caracteristica>culoare</caracteristica>
        <caracteristica>design</caracteristica>
</obiect>

</continut>
```

Un format care ar fi mult mai util pentru procesare online ar fi pastrarea într-un format mult mai rapid și ușor de procesat de către limbajele ce se folosesc pentru programarea web. Un format mult mai recomandat ar fi păstrarea în format JSON, ceea ce înseamnă că fișierul va arata astfel:

{"telefon":{"cnt":4,"p1":"ecran","p2":"baterie","p3":"greutate",
"p4":"dimensiune"}}
{"baterie":{"cnt":2, "p1":"greutate", "p2":"durată de viaţă"}}
{"laptop":{"cnt":9, "p1":"ecran|display","p2":"greutate","p3":
"fiabilitate","p4":"dimensiuni","p5":"RAM","p6":"hard|hard-disk|HDD",
"p7":"procesor","p8":"culoare","p9":"design"}}

De asemenea, având în vedere că se lucrează cu cantităţi foarte mari de date, este atât dificil cât și nepractic să păstrăm aceste informaţii într-un fișier text, așa că una dintre propuneri ar fi păstrarea acestor informaţii într-o bază de date. Una din bazele de date foarte recente aparute în industrie și care se împacă nativ formatul JSON, este MongoDB.

MongoDB este capabilă să memoreze cantităţi impresionante de informaţie și datorită modului diferit de reprezentare a datelor, este de asemenea capabilă să răspundă foarte rapid la cereri.

De exemplu, crearea unei colecţii în care să păstrăm informaţii ca cele prezentate mai sus, se face folosind următoarele comenzi.

```
mongo
use mydb
db.elements.save({"telefon":{"cnt":4,"p1":"ecran","p2":"baterie",
"p3":"greutate","p4":"dimensiune"}});
show collections
```

În mod evident, fiecare element al unui obiect poate are anumite proprietăţi care sunt pozitive și altele care sunt negative. De asemenea, nu există o corelaţie foarte clară între elementele unor obiecte diferite și atributele lor. Un atribut asociat unui element și care exprimă o opinie pozitivă poate exprima o opinie negativă pentru chiar același element dar al unui alt obiect și în mod evident pentru elemente diferite.

Dacă avem la dispoziţie ambele resurse, atât descompunerea în elemente pentru fiecare obiect, cât și lista de proprietăţi pozitive și negative pentru fiecare element, atunci putem face o aproximare destul de bună a orientării opiniei din documentul analizat.

Aceste resurse pot fi păstrate separat, ca două resurse separate, sau pot fi păstrate împreună, având la dispoziţie astfel o resursă completă pentru analiza sentimentelor. Această resursă va fi disponibilă pe site.

Concluzii parţiale

Am discutat mai sus despre analiza sentimentelor bazate pe sentimentele asociate fiecărei componente sau subcomponente ale obiectului analizat. Mai este de facut o singură observaţie: în momentul în care facem această analiză, este foarte important să ne uitam şi la zona geografică, deoarece se pot observă diferenţe de gusturi. Cel mai elocvent exemplu este vânzarea de telefoane iphone 6 şi iphone 6 plus: dacă în Statele Unite şi în Europa iphone 6 a stat de aproape 6 ori mai bine decât iphone 6 plus, însă nu acelaşi lucru se poate spune şi despre Asia, unde iphone 6 plus a fost un succes.

Motivul este foarte simplu: dacă în Statele Unite şi în Europa review-urile au fost oarecum spre negativ datorită ecranului foarte mare comparativ cu toate generaţiile precedente de iphone, în Asia, acest lucru a fost mai degrabă apreciat.

Dacă ar fi să facem această analiză la nivel mondial şi să coroborăm rezultatele cu cele de vânzări, atunci am putea concluziona că iphone 6 este net superior faţă de iphone 6 plus, iar ca principal argument ar fi dimensiunea foarte incomodă a telefonului şi a faptului ca este mai degrabă o tabletă decât un telefon şi poate în unele situaţii inclusiv preţul poate fi un argument. Spre exemplu, dacă un telefon iphone valorează în Statele unite suma de X USD, atunci în România acesta va valora X EUR, iar în UK X GBP şi având în vedere diferenţa de curs valutar, atunci diferenţa de preţ este chiar una substanţială.

Putem concluziona că analiză sentimentelor la nivel mondial poate avea rezultate eronate sau poate chiar controversate. Ca sa-l citãm pe Mark Twain: "There are three kinds of lies: lies, damned lies and statistics" ceea ce face foarte grea analiza sentimentelor la nivel mondial fără să ne uitam si la zona geografică.

Sumarizarea

O dată cu dezvoltarea rapidă a Internet-ului, din ce în ce mai multă informaţie este disponibilă on-line. Această explozie de date a evoluat însă într-o supraîncărcare cu informaţii. Nu există timp suficient pentru a putea parcurge toate materialele disponibile şi de regulă deciziile critice trebuie făcute pe baza datelor disponibile la acel moment de timp. Ca şi raspuns la această problemă, numarul de cercetători interesaţi de sumarizare a crescut considerabil.

O dată cu creşterea acestui interes, se poate observa deasemenea şi o creştere a bugetului dedicat cercetării acestui domeniu. Statele Unite (DARPA - Defense Advanced Research Projects Agency), Uniunea Europeana şi ţările din zona Pacificului au identificat sumarizarea textului ca fiind un domeniu de cercetare critic şi încep să investească în dezvoltarea sa. Deasemenea, se poate remarca un interes şi din partea sectorului privat, în special industria telecom care a dezvoltat o serie de tehnici în acest sens: Prosum - BT - data mining pentru explorarea unor colecţii mari de date, Context - Oracle - filtre web pentru găsirea documentelor online, Inxight - AltaVista Discovery.

Aşa cum am arătat şi în introducere, nivelul comunicaţiilor în interiorul unei companii a crescut considerabil în ultimii ani, iar şansele să crească şi mai mult în următorii ani

sunt foarte mari. Pe langă această creştere a volumului comunicaţiilor, se poate observa un număr copleşitor şi în creştere de potenţiale surse de informare, majoritatea fiind în formă scrisă, ca de exemplu, reţelele de socializare.

Resursele şi timpul limitat forţează managerii să-şi limiteze documentarea la doar un mic subset din toată informaţia disponibilă. Acest lucru poate afecta performanţa şi calitatea deciziilor pe care aceştia le iau. De exemplu, în companiile de success se investesc mai mulţi bani în monitorizarea presei şi a altor surse de documentare faţă de companiile cu performanţe mai slabe. Accesul la informaţie în timp util poate aduce multe beneficii pentru deciziile luate în cadrul unei companii.

Scopul unui rezumat este simplu şi evident: de a facilita identificarea unui subiect din mai multe articole şi selectarea materialelor cele mai relevante pentru subiectul de interes. În general, crearea unui rezumat implica o oarecare familiaritate cu subiectul dezbătut. Extragerea elementelor cheie pe care autorul articolului a dorit să le sublinieze necesită de regulă antrenament şi experienţă. În general, se face o distincţie clară între scopul unui rezumat şi tipul de rezumat folosit, însă din nefericire, terminologia în acest domeniu este incă insuficient dezvoltată pentru a permite acest lucru. O descriere mai generală împarte rezumatele în două mari subclase: rezumatele indicative, care indică subiectele dezbătute în documente şi pot anunţa cititorul/utilizatorul despre documentele care dezbat un anumit subiect şi rezumatele informative care descriu subiectele dezbătute în lucrare.

Rezumatul manual este întotdeauna influenţat de trecutul persoanei care îndeplineşte această sarcină, de personalitatea acestuia şi de dispoziţia sa. Părerile personale ale persoanei care extrage rezumatul şi interesele sale sunt de asemenea responsabile pentru subiectivismul rezumatului rezultat. Putem extrapola această situaţie şi putem afirma că un articol poate avea mai multe rezumate în funcţie de persoana care l-a extras şi dispoziţia acestuia la acel moment de timp.

Un alt factor ce influenţează cercetarea în domeniul sumarizarii automate este lipsa unei măsurători obiective a calităţii unui abstract sau a unui rezumat. Cu toate acestea, Morris et all afirmă că o metodă destul de obiectivă poate fi măsurarea înţelegerii textului iniţial comparat cu varianta rezumată a acestuia.

Având în vedere interesul pentru acest domeniu, la momentul actual se doreşte eliminarea subiectivităţii umane şi găsirea de algoritmi performanţi şi capabili să efectueze această sarcină riguros şi consecvent.

Imediat după ce s-a început cercetarea în domeniul traducerii automate, s-a remarcat un interes crescând pentru posibilitatea extragerii automate de rezumate ale documentelor. Accentul s-a pus mai mult pe rezumatele indicative, adica pe rezumatele ce îi confera cercetatorului un indiciu referitor la ce documente sunt mai importante decat altele pentru un anumit subiect decât pe rezumatele informative, scopul fiind acela de a oferi un rezumat ce poate substitui citirea unui document.

O altă ipoteză investigată a fost că extrasul unui document (selectarea propoziţiilor cheie din document) ce ar putea fi considerat deasemenea rezumatul documentului. Cu toate ca această ipoteză a fost recunoscută o perioadă îndelungată de timp drept extract automat, în mod informal, este denumit tot rezumat automat.

Orice cercetare în domeniul sumarizării automate trebuie să înceapă cu abordările clasice ale lui Luhn și Edmundson, trebuie să parcurgă abordările bazate pe corpus/conţinut și să exploateze structura discursurilor, să treacă prin abordările ce utilizează o bază de date de cunoștinţe, să exploreze divere metode de evaluare și în mod evident să sublinieze problemele nerezolvate.

Metodele originale ale lui Luhn atribuiau ponderi fiecărei propoziţii din document în funcţie de diferite criterii statistice, în special fiind vorba despre o funcţie legată de frecvenţa cuvintelor ce apar într-o propoziţie. Cuvintele foarte comune sunt eliminate (de, pe, care, unde, sus...etc) cu toate că au cea mai mare frecvenţă atât în document cât și în fiecare propoziţie. Eliminarea se face pe baza unui dicţionar iar pentru restul cuvintelor au asignate ponderi în funcţie de frecvenţa acestora, stabilindu-se astfel o funcţie ce exprimă înţelesul (subiectul, semnificaţia) propoziţiilor.

Extrasele ce foloseau aceste metode au oferit rezultate suficient de bune încât să încurajeze aprofundarea cercetării în acest domeniu. În același timp, o metodă pur statistică pentru determinarea rezumatelor unui document s-a dovedit neadecvată și implicit, au apărut și alte metode de-a lungul timpului.

În decursul ultimilor ani au fost prezentate o multitudine de abordari pentru sumarizarea automata, însă acestea se pot împărţii în trei mari cateogrii, în funcţie de nivelul de procesare: abordări de suprafaţă, abordări la nivel de entitate și abordări la nivel de discurs.

Abordările la nivel de suprafaţă au tendinţa să reprezinte informaţia printr-un set de euristici ce vizează anumite caracteristici. Aceste caracteristici sunt apoi combinate cu ajutorul unui grup de funcţii într-un set finit de variabile predefinite din care se pot extrage anumite concluzii mult mai ușor.

Aceste euristici pot urmării:

- termenii tematici - prezenţa unor termeni importanţi determinaţi pe bază statistică
- localizarea - poziţia în text, poziţia în paragraf, etc
- fundalul - prezenţa unor termeni din titlu și subtitlu în interiorul documentului
- cuvintele de evidenţiere sau fraze de evidenţiere: în concluzie, cercetarea noastra a evidenţiat, important, bonus, în particular, etc

Abordările la nivel de entitate contruiesc o reprezentare internă pentru text, modelând entităţile din text si legăturile dintre acestea. Aceste abordări urmăresc să găsească tipare între legaturile dintre entităţi și se folosesc de teoria grafurilor pentru a determina ceea ce este important. Relaţiile dintre entităţi includ:

- similaritatea - suprapunerea de dicţionar
- proximitatea - distanţa dintre unităţile de text
- înrudirea - cuvinte legate prin apariţia acestora în expresii comune
- legăturile de dicţionar - sinonimie, hipernimie, etc
- coreferenţierea - propoziţiile care fac o referenţiere la un anumit subiect
- relaţiile logice - aprobarea, contradicţia, consistenţa, etc
- relaţiile sintactice

- relaţiile bazate pe înţeles

Abordările la nivel de discurs modelează întreaga structură a textului în relaţie cu informaţia transmisă. Aceastra structură poate include:

- formatul documentului
- legaturile dintre subiectele de discuţie pe masură ce sunt descoperite
- structura retorica a textului

Crearea unui rezumat poate fi descompusă în 3 părţi principale: analizarea textului primit la intrare, transformarea acestuia într-o reprezentare pe scurt și sintetizarea acestuia într-un rezumat corespunzator.

Sumarizarea mai poate fi catalogată și în funcţie de metoda folosită: (1) folosind procesarea limbajului natural pentru construcţia unei hărţi de cunoștinţe a unui document pe baza căreia se poate genera un extract; și (2) utilizarea unui algoritm pentru extragerea celor mai importante propoziţii și fraze din document. Procesarea limbajului natural necesită cunoștinţe sintactice, semantice și pragmatice. Cunoștinţele sintactice conţin regulile de structurare al unui limbaj iar cunoștinţele semantice si pragmatice sunt specifice domeniului analizat. Din nefericire, utilizarea singulară a sintacticii poate produce uneori interpretări eronate datorită expresiilor combinate și a metaforelor.

Sumarizarea automată acoperă mai multe sub-probleme, ce vor fi dezbătute de asemenea în acest capitol. Pentru limba română, primele încercări de sumarizare automată datează din anii 70 prin lucrările lui Erika Nistor și Eliza Roman și utilizau algoritmul lui Luhn. Testele realizate de aceștia au constat în procesarea a 75 de pagini de text, reprezentând 21 de articole sau capitole scurte din cărţi din mai multe domenii (știinţă medicală, matematică, istorie și muzică).

Lucrările prezintă rezultatele experimentelor și discută dificultăţile întâlnite în timpul experimentelor, în special datorate limbii române. Metoda lui Luhn aplicată pentru limba română implica la momentul respectiv împărţirea fiecărei propoziţii în segmente despărţite de termeni de legătură (termeni comuni sau cuvinte-stop), la nu mai mult de 4 termeni nesemnificativi distanţă. Luhn puncta fiecare segment utilizând pătratul numărului de elemente semnificative despartite de termenii de legătură. Acest tip de segmentare este puţin semantic comparat cu alte metode de segmentare a textului. Nistor si Roman au observat necesitatea unei liste de cuvinte stop, însă lipsa resurselor lingvistice in format electronic le-au creat destule probleme in anii 70.

În articolele ulterioare Roman si Nistor încearcă împărţirea textului în propoziţii nucleu. Aceste transformări urmăresc recomandările lui Chomsky de a împărţi frazele în propoziţii nucleu. În timp ce Chomsky a găsit 9 categorii de propoziţii nucleu, Nistor si Roman au încercat un studiu asemănător pentru limba română și au aplicat aceste rezultate pentru sumarizarea textelor în limba română, descoperind astfel 8 categorii. Lăsând la o parte aportul indubitabil al acestei cercetări, evaluarea calităţii acestor rezultate (atât teoretice cât și practice) cu privire la limba română rămâne la latitudinea lingviștilor români.

O dată ce textul a fost împărţit in propoziţii nucleu, avand 8 forme de baza stabilite de către autori, abstractul textului poate fi apoi contruit in mod automat. În mod evident, primele întrebări care apar sunt: care sunt cele mai eficiente cuvinte cheie? cele mai frecvente propoziţii nucleu? care sunt cele mai frecvente cuvinte? Autorii ajung la concluzia că acele cuvinte cu cea mai mare frecvenţă sunt deasemenea şi subiectele propoziţiilor nucleu cu cea mai mare frecvenţă. Acest lucru poate fi tradus prin faptul că metoda lui Luhn poate fi combinată cu această transformare și se poate obţine astfel un rezumat din cele mai frecvente propoziţii nucleu.

În următoarele lucrări, cei doi autori își lărgesc noţiunea de rezumat, ceea ce reprezintă o realizare majoră pentru perioada respectivă. Astfel, Nistor și Roman (1979) încearcă să identifice acele propoziţii dintr-un text care se referă la un subiect apriori specificat de către utilizator. Așa cum scriu și cei doi autori, două idei se pot evidenţia: (1) utilizarea de mulţimi fuzzy și funcţia acestora de a măsura similaritatea semantica dintre documente/text și cuvinte cheie și (2) utilizarea dicţionarului disponibil sau construit special pentru acest scop.

Concluzia principală a acestui studiu este ca relaţia dintre un set de documente și rezumatele acestora nu mai este biunivocă. Poate exista un singur document și mai multe rezumate care indică către el. Acesta este unul dintre motivele pentru care autorii își denumesc metoda “rezumare dinamică”. La momentul actual, sunt puţine referinţe în literatura de specialitate care se referă la sumarizarea în limba română.

Începem explorarea acestui domeniu prin prezentarea mai întâi a tehnicilor existente de sumarizare, și inclusiv a diverselor tehnici de preprocesare a textelor a priori sumarizării și vom încheia cu câteva aplicaţii practice pentru limba română. Este destul de greu să ne imaginăm viaţa de zi cu zi fără sumarizare. Titlurile ştirilor sunt de regulă un rezumat al întâmplărilor descrise în conţinut. De multe ori, prima frază este deja suficientă pentru a ne face o idee despre întregul document. În cazul articolelor ştiinţifice, abstractul sau rezumatul care se completeaza la început prezintă foarte pe scurt subiectul dezbătut și concluziile care se vor regăsi la final.

Segmentarea

Segmentarea textului reprezintă procesul de împărţire a textului sursă în paragrafe și propoziţii. Vom numi un document procesat în acest fel "DocSent" provenit din cuvintele englezești *document* și *sentence*. Această împărţire în paragrafe şi propoziţii este realizată printr-o parcurgere secvenţială a textului, când se încearcă detectarea semnelor de punctuaţie care pot arăta sfârșitul unei propoziţii. Acestea pot fi semne simple, precum punctul (.), semnele întrebării (?) și exclamării (!), sau semne compuse, precum elipsa (...) sau combinaţii de semne ale întrebării și exclamării (?!, !?). O problemă care poate apărea la acest nivel o reprezintă abrevierile, care de obicei se termină cu punct. În cazul în care algoritmul nu este pregătit pentru această situaţie, el va considera că propoziţia se sfârșește după abreviere (Mr.) sau chiar în interiorul ei (U.S.A.). Pentru a depăși acest inconvenient, se folosește un corpus de abrevieri disponibil on-line, care a servit la verificarea existenţei acestora în text. Astfel, ele vor fi substituite astfel încât să nu existe confuzii cu privire la sfârșitul real al propoziţiilor. După efectuarea împărţirii în propoziţii, abrevierile înlocuite anterior vor fi readuse la forma iniţială, textul fiind păstrat în acest fel nealterat.

WebSumm

WebSumm funcţionează în două moduri: (1) ca un instrument de filtrare, ordonare şi navigare printre titlurile si subtitlurile textelor întoarse de motoarele de căutare, şi (2) ca un instrument de navigare în adâncime ce foloseşte extragerea de text şi sumarizarea pentru a analiza informaţia din setul de documente.

Algoritmul descris de Mani si Bloedorn poate fi împărţit în trei paşi:

Mai întâi se identifica toate asemănările şi diferenţele între toate documentele implicate, acţiune ce se poate efectua utilizând o serie multiplă de algoritmi. Unul din avantajele determinării similarităţilor şi diferenţelor dintre aceste documente, este faptul că similarităţile ne arată ce este comun în fiecare document corelat cu subiectul dezbătut în toate documentele, diferenţele ne arată ceea ce este unic în fiecare document. Dacă presupunem ca documentele sunt ordonate în ordine cronologică, diferenţele din ultimul document pot indica ce este nou într-un set de documente ce dezbat acelaşi subiect. În mod evident, atat similarităţile cât şi diferenţele pot ajuta foarte mult la construirea unui rezumat.

Se poate apoi determina care sunt cei mai importanţi termeni în cazul unei căutări, acest lucru pornind de la ipoteza că localizarea şi ordinea liniară au un rol foarte mare în determinarea termenilor importanţi. Această sarcină poate fi înfăptuită utilizând un algoritm de ierarhizare (ranking) pe paragrafe. Cu toate acestea, alegerea dimensiunii unei ferestre cu scopul alegerii paragrafelor potrivite poate fi o problemă, chiar daca sunt folosite ferestre de dimensiune fixă sau ferestre cu dimensiune variabilă în funcţie de discurs. Daca dimensiunea ferestrei este prea mică, atunci observatorul ar putea obţine o mulţime mai mare de ferestre mici care oferă puţină informaţie în sine, dar impreuna pot fi foarte relevante. Daca însă dimensiunea ferestrei este prea mare, atunci e posibil să conţină prea multă informaţie irelevantă. În loc să se folosească ferestre de dimensiune fixă sau să se utilizeze un efort mult prea mare pentru a utiliza ferestre de dimensiune dinamică, Mani şi Bloedorn recomandă utilizarea unui algoritm ce conferă diferite ponderi elementelor din text (unde termenii respectivi corespund nodurilor din graf). De asemenea, idenfiticarea acestor regiuni din documente relevante pentru căutarea efectuată, reduce foarte mult spaţiul în care se caută asemănări sau diferenţe.

Se explorează de asemenea un model de reprezentare a textului care ia în considerare şi gradul de conectare dintre termeni. Această reprezentare explorează un model de conectivitate unde puterea legaturilor (în general bazată pe similaritate) între diferite bucăţi de text este utilizată pentru identificarea caracteristicilor importante într-unul sau mai multe documente.

Sumarizarea prin extragerea frazelor cheie

Probabil cea mai întâlnită metodă de sumarizare este prin extragerea celor mai importante paragrafe sau fraze din text şi aşezarea acestora într-o anumită ordine. Voi prezenta în continuare câţiva din aceşti algoritmi şi voi discuta eficacitatea, avantajele şi dezavantajele acestora. În mod normal, sumarizarea unuia sau a mai multor documente conţine mai mulţi paşi:

- Înţelegerea conţinutului documentului sau a documentelor
- Identificarea celor mai importante componente ale articolului sau ale articolelor
- Scrierea sumarului

Având în vedere varietatea informaţiei, este mult mai avantajoasă implementarea unei tehnici de sumarizare independentă de domeniul cu care se lucrează, însă automatizarea paşilor 1 şi 3 este la momentul actual într-un stadiu incipient. De aceea, cel mai frecvent tip de sumarizare la momentul acutal este extrasul. Extrasele sunt în general creeate cu ajutorul euristicilor bazate pe o analiză statistică a apariţiilor unor anumite cuvinte, dorindu-se astfel selectarea acelor fraze sau propoziţii care au cel mai mare potenţial pentru rezumarea documentului. De-a lungul timpului, au fost prezentate diverse metode pentru selectarea și extragerea celor mai concludente propoziţii.

Summarist

Summarist este una din încercările ce a dorit dezvoltarea unui sistem capabil să extragă rezumatul unui document, urmărind definiţia: sumarizare = identificarea subiectului/topic-ului, interpretare şi generare. Trebuie să dezbatem fiecare punct în parte, deoarece conţin o multitudine de subalgoritmi, fiecare antrenaţi pe un corpus mare de date de antrenare.

Înainte de crearea oricărui rezumat, trebuie să determinăm care este subiectul central dezbătut în document. Apoi acest subiect/topic trebuie interpretat și apoi se poate genera rezumatul textului.

- *identificarea* - scopul acestui pas este eliminarea subiectelor inutile pentru crearea unui rezumat al textului. De regulă, se pleacă de la presupunerea că orice text dezbate mai multe subiecte, ceea ce introduce două ipoteze: (1) utilizatorul va dori extragerea celor mai importante \textit{N} subiecte din document sau (2) extragerea celor mai importante subiecte (pentru el) din text.
- *interpretarea* - o dată ce subiectele principale au fost identificate, acestea pot fi listate pentru prezentarea unui extract. Cu toate acestea, în sumarele extrase manual, este nevoie și de un pas de interpretare cu scopul obţinerii a unei compactări mai mari. Într-unul din studiile sale, Daniel Marcu a observat că lucrându-se cu un copus de 10 documente extrase din ziare și cu 14 analiști, se poate obţine un factor de compresie de 2.76. Deasemenea, a observat că un extract este de regulă de trei ori mai lung decât un sumar extras manual, calculul făcându-se pe numărul de cuvinte folosit. Aceste rezultate indică nevoia unei sumarizări ulterioare extragerii informaţiilor importante, cu scopul eliminării informaţiilor care se repetă, reformularea propoziţiilor a.i. să conţină mai multă informaţie condensată, și de asemenea, generalizarea rezumatelor.
- *generalizarea* - are ca scop reformularea extracturilor într-un text coerent, dens și scurt. Dacă se sare această fază, atunci rezumatul va fi compus din simpla extragere a unor fraze cheie din textul iniţial.

Chiar dacă un rezumat extras manual nu se va compara niciodată cu un extras automat, în mod special datorându-se faptului că pentru crearea unui rezumat corect este totuși nevoie de înțelegerea adevăratului sens al documentului analizat, nu se poate totuși contesta necesitatea rezumatelor automate.

Probabil una dintre cele mai frecvente aplicaţii pentru sumarizarea automată sunt aplicaţiile de extragere a informaţiei. Începând cu 1950, mulţi cercetători și-au îndreptat atenţia spre acest domeniu, încercând să categorisească textele pe categorii, și să extragă esenţa documentelor analizate.

Abordările de bază ale metodelor de extragere a informaţiilor (information retrieval) au totuși un set de limitări. De multe ori, orice depășea nivelul cuvintelor era privit cu neîncredere. De aceea, pasul doi din summarist este atât de important: fără înțelegerea deplină a subiectelor dezbătute și posibilitatea reformulării lor, rezumatele nu vor fi mai mult decât propoziţii din text extrase și concatenate alături de seturi de cuvinte care se repetă cu o frecvenţă mai mare.

Cu toate că metodele la nivel de cuvânt au fost destul de bine analizate, există în continuare câteva aspecte care împedică aceste metode de a fi perfect:

- *Sinonimie* - un concept poate fi reprezentat prin mai multe cuvinte
- *polisemie* - un cuvânt poate avea mai multe înţelesuri
- *expresii* - o expresie poate avea un alt înţeles decât cuvintele din care este compusă
- *dependenţa dintre termeni* - majoritatea tehnicilor probabilistice privesc cuvintele ca fiind independente, fără a lua în considerare legăturile dintre ele

De exemplu, Summarist nu calculează doar frecvenţa cuvintelor (o tehnică frecventă în IR (information retrieval) ci calculează și frecvenţa conceptelor (utilizând WordNet și alte resurse) a.i. poate opera la un nivel mai adânc decât cel al cuvintelor. La momentul actual, partea de interpretare este încă la un nivel incipient, deoarece majoritatea atenţiei a fost îndreptată spre extragerea conceptelor.

Fiecare modul este conceput să lucreze independent, deși este posibil ca unele module să aștepte informaţii de la altele. În faza de preprocesare, fiecare cuvânt este scris pe câte o linie. Apoi, fiecare modul deschide fișierul de preprocesare și adaugă ce informaţie este considerată ca fiind relevantă pentru fiecare cuvânt, fiind vorba de multe ori de valori numerice, însă poate fi vorba și de cuvinte (stringuri). Fișierul poate fi deschis și inspectat la orice moment de timp. La sfârșitul fiecărei faze, un modul de integrare combină scorurile precedente într-un scor final, ce se adaugă de asemenea în același fișier.

Interpretarea Subiectului (Topic-ului)

Al doilea pas în sumarizare este interpretarea subiectelor dezbătute în articol. În acest pas, unul, două, sau mai multe topici sunt combinate într-una (sau mai multe) concepte unificatoare. Acest pas poate fi unul simplu, ca de exemplu combinarea unor cuvinte ca "roată", "lanţ", "pedală", "lumină", "șa" într-un singur concept/cuvânt "bicicletă". În mod evident, procesul nu este atât de simplu precum reiese din

exemplul de mai sus, însă nu este o sarcină imposibilă. Cu toate acestea, este unul dintre cele mai dificile probleme din cadrul sumarizării automate, motivul principal fiind faptul că această fuziune de concepte necesită cunoştinţe suplimentare, care de regulă nu sunt incluse explicit în articol.

Există diverse metode ce pot fi folosite la acest pas, iar majoritatea implică un set de indicatori ce au ca și corespondent un concept mai general.

Pentru identificarea subiectelor (topic-urilor), se pornește în general de la premiza că cu cât un cuvant este folosit mai frecvent în text, cu atât este mai important pentru textul analizat. Cu toate că această metodă oferă rezultate destul de bune, indiferent de domeniul analizat, nu se poate descurca însă cu sinonime, expresii complexe sau polisemie. Printr-o listă de concepte ce pot fi exprimate printr-o multitudine de cuvinte se pot depași aceste dificultăţi. Deasemenea, Summarist calculează frecvenţa conceptelor și nu doar a cuvintelor din text, folosindu-se pentru generalizare (determinarea conceptelor părinte) WordNet-ul. Dacă însă nu se găsesc informaţiile necesare în WordNet, atunci se numără cuvintele.

Algoritmul este destul de simplu: se numără de câte ori apare fiecare cuvânt în text, iar numărul de apariţii este adăugat la numărul de apariţii al conceptului reprezentat prin acel cuvânt. Procesul se poate termina o dată ce s-au trecut prin toate cuvintele din document, sau în orice moment în care considerăm că un anumit concept este suficient de bine reprezentat. Acest algoritm va alege cele mai generale concepte prezente în articol.

Rezultate similare se pot obţine și dacă se folosesc semnăturile de topic-uri.

Generarea Sumarului

Ultimul pas în sumarizarea automată este generarea sumarului, ce poate fi rezolvat prin mai multe metode, de la simpla tipărire a propoziţiilor importante din text, până la construirea de propoziţii complexe ce includ cât mai multă informaţie din textul original.

Summarist conţine 3 module de generare a rezumatului, în funcţie de cerinţele utilizatorului. În cazul în care este dorit un extract, se poate elimina complet peste partea de generare. Termenii și/sau propoziţiile extrase în faza de extragere a topic-ului pot fi listaţi direct ca și rezumat al textului iniţial. Cu toate că ar putea părea incoerent, acest extract ar putea conţine suficiente informaţii pentru ca un utilizator să-si dea seama de relevanţa acestui articol.

De multe ori, nu este nevoie de generarea unui rezumat. Simpla listare a subiectelor dezbătute în textul analizat poate fi suficientă pentru utilizator. Summarist este capabil să efectueze și această sarcină.

Summarist conţine un generator rudimentar de rezumate, ce poate concatena propoziţii și poate obţine propoziţii simple din elementele extrase. Urmărind legăturile între cuvinte și între concepte și fiind capabil să formeze propozitii enunţiative sau cauzale.

Cu toate că este departe de a fi perfect, Summarist se regăsește printre cele mai apropiate metode de sumarizare existente. Rezultatele întoarse pot varia de la "foarte bune" până la "acceptabile", în funcţie de tipul rezumatului dorit și de pretenţiile utilizatorului.

De asemenea, Summarist poate fi folosit ori pentru extragerea rezumatelor din texte, dar poate fi folosit și pentru diferite sarcini intermediare, precum extragerea cuvintelor cheie, identificarea subiectelor și determinarea conceptelor principale din text.

Concluzii Parţiale

Având în vedere ca în capitolul 5 tratăm presa din România, o dată ce ne uitam suficient de mult timp în trecut, putem observa cum majoritatea publicaţiilor au adoptat un stil tabloid. Dacă așa cum am văzut până acum, titlul putea fi de multe ori considerat ca un sumar al articolului, în ziua de azi, lucrurile nu mai stau așa.

Să luam ca și exemplu următorul articol din ziuarul Libertatea, publicat pe data de 22 Noimebrie 2014:

Titlu: DEZVĂLUIRI din INTIMITATE. CUM îşi ÎMPART TREBURILE CASNICE Klaus Iohannis şi SOȚIA SA, o GOSPODINĂ DESĂVÂR\C{S}ITĂ

Conţinut: "Marius Vecerdea, antrenorul de tenis al lui Klaus Iohannis, din 2009, a dezvăluit cum î?i impart treburile casnice pre?edintele ales ?i so?ia sa, care se pare că este o gospodină desăvâr?ită.

Klaus Iohannis, preşedintele ales al României va avea sâmbătă seară ocazia să se bucure de rezultatul alegerilor preziden?iale, acasă, la Sibiu, unde va da o petrecere. Evenimentul va avea loc într-un cadru restrâns, pre?edintelui Klaus Iohannis alăturându-i-se prieteni foarte apropiaţi cu care îşi petrece de obicei concediile, potrivit oradesibiu.ro.

"Îmi este imposibil să vă spun exact cum se numește ce gătește Prima Doamnă. Oricum, e genială în bucătăria franţuzească. Gătește o raţă de te lingi pe degete. Apoi, e tare la dulciuri. Nicăieri n-am mâncat bezele și tarte cu fructe mai bune ca ale ei. Are maşină de înghețată acasă și, dacă vă vine să credeţi, își coace singură chiar și pâinea. Totul e ecologic. Iar la final vine șocul. Iohannis spală vasele. Am rămas mască. M-am oferit să-l ajut, dar n-a vrut să audă. E treaba lui dintotdeauna", povestea Vecerdea."

Link către știre: http://www.libertatea.ro/detalii/articol/klaus-iohannis-treburi-casnice-sotie-spalat-vase-515624.html

Așa cum putem observa, există o oarecare legătură între titlul ales și conţinutul articolului, însă de asemenea, putem observa cât de exagerat este și cât de mult denaturează situaţia. Avaând în vedere situaţia în care se afla presa în momentul de faţă (nu doar în România ci chiar la nivel internaţional), dupa ce s-a trecut printr-o scădere considerabilă a bugetelor și multe publicaţii tipărite chiar s-au închis, de multe ori titlurile sunt denaturate a.î. să șocheze și să atragă cititori.

În cazul Summarist ce folosește OPP (optimal position policy) - statistic s-a observat că propoziţia imediat următoare sub titlu este cea care are cea mai mare valoare pentru un rezumat al textului. Acest lucru se poate remarca și in articolul de mai sus, însă folosirea cuvantului "dezvăluiri" (care sunt prezentate apoi in fraza nr. 3) are în continuare ca scop ţinerea cititorului în suspans.

Pe de alta parte, să luăm în continuare exemplul următor:

Titlu: Antivirusul românesc Bitdefender va proteja infrastructura IT a guvernului din Danemarca

Continut: Soluţiile de securitate Bitdefender au fost selectate de Agenţia de Modernizare a Administraţiei Publice din Danemarca pentru protejarea infrastructurii IT a instituţiilor din cadrul guvernului danez, a anunţat producătorul român de soluţii de securitate.

Urmare a selecţiei realizate de agenţia daneză, specialiştii Bitdefender implementează soluţiile Bitdefender GravityZone Security for Virtualized Environments (SVE), GravityZone Security for Endpoints și Cloud Security for Endpoints (Small Office Security) în ministere și alte instituţii publice din Danemarca.

"înainte de a fi selectaţi de agenţia daneză, aveam deja municipalităţi dar şi companii private din Danemarca în portofoliul de clienţi, ce aleseseră soluţiile Bitdefender atât datorită securităţii de top oferite cât şi datorită soluţiilor unice dedicate mediilor virtualizate, ce reduc consumul de resurse şi costurile asociate mentenanţei", a declarat directorul de vânzări pentru Danemarca şi ţările Scandinave în cadrul Bitdefender, Brian Ziegler, într-un comunicat.

Potrivit contractului, Bitdefender furnizează Guvernului Danez protecţie împotriva virușilor, tentativelor de phishing, amerinţărilor de tip ransomware și împotriva altor ameninţări ce vizează terminale de tip PC/laptop indiferent de sitemul de operare, pentru toate tipurile de servere, precum și pentru terminale mobile precum Android și iPhone.

Pachetul cuprinde și o soluţie pentru email precum și o soluţie de scanare a traficului în reţea.

Securitatea infrastructurilor IT virtualizate de mari dimensiuni, aparţinând statului danez e realizată cu soluţia Bitdefender GravityZone Security for Virtualized Environments (SVE), soluţie compatibilă cu toate tipurile de hipervizoare (tehnologie ce stă la baza virtualizării), și care rezolvă provocările tradiţionale legate de mediul cloud, oferind avantaje semnificative în termeni de securitate, consum redus de resurse și un timp mai redus dedicat mentenanţei.

Din consola centrală a GravityZone, instituţiile ce implementează soluţia în cadrul Guvernului Danez pot gestiona toate dispozitivele mobile sau fixe și toate tipurile de clienţi virtuali și servere.

Brandul Bitdefender a fost lansat în 2001, prin redenumirea antivirusului AVX, produs și comercializat de Softwin. în 2007, divizia Bitdefender s-a desprins de grupul de firme Softwin.

Sursca: http://www.zf.ro/business-hi-tech/antivirusul-romanesc-bitdefender-va-proteja-infrastructura-it-a-guvernului-din-danemarca-13603207

În continuare putem observa că primul paragraf este cel mai important și acest lucru se poate observa deoarece majoritatea articolelor respecta piramida inversă.

În teoria și practica jurnalistică, "piramida inversată" este o tehnică de redactare, aplicată în special știrilor, prin care informaţiile esenţiale sunt prezentate la începutul articolului. Celelalte detalii sunt redate descrescător, în ordinea importanţei.

Piramida inversată este formată din:

- Lead – (prima frază/paragraful iniţial, introducere, capul știrii) conţine mesajul esenţial și are rolul principal de a incita cititorul la lectură;
- Paragrafe de susţinere – conţin date care explică și aprofundează introducerea. Ajută la situarea evenimentelor în context. Conţin o serie de date secundare care întregesc imaginea faptului. Background-ul își are locul tot în această zonă;
- Final – este aproape la fel de important ca lead-ul. El trebuie să fie memorabil, să fixeze în mintea cititorului informaţia esenţială.

Adevăratul scop și potenţial al piramidei inversate este captarea atenţiei publicului încă din primele secunde de lectură. Pe online studiile au demonstrat că publicul scanează informaţia. Nici pe print lucrurile nu stau foarte diferit.

În principiu, publicul n-are răbdare sau n-are timp să citească tot articolul ca să afle despre ce e vorba. De regulă, cititorul parcurge titlurile, fotografiile, explicaţiile foto, infografiile, apoi începe să citească. Dacă prima frază nu i-a trezit interesul, sunt mari șanse să nu citească în continuare.

Acum dacă ne uităm la cele doua articole, unul cu tentă tabloidă iar celălalt într-o tentă mai serioasă fiind vorba de un comunicat de presă, prima și ultima propoziţie poate sumariza destul de bine ideea de bază din articol.

Sistem Automat de Monitorizare Media

Partea aplicativă a acestei lucrări este formată din două componente. Prima parte este formată dintr-un website ce va pune la dispoziţie atât majoritatea aplicaţiilor prezentate în această lucrare (atât cod sursă cât și executabil), cât și un set de resurse lingvistice ce pot fi folosite pentru dezvoltarea ulterioară de aplicaţii. Aceste resurse sunt disponibile gratuit și pot fi utilizate de către oricine dorește să continue cercetarea în acest domeniu sau are nevoie de resurse lingvistice pentru diverse actiuni de procesare automată a limbajului natural.

Partea a doua este o aplicaţie ce dorește să demonstreze utilitatea acestor algoritmi, prin implementarea unui sistem automat de monitorizare media.

Deasemenea, partea de monitorizare media nu își dorește să fie un sistem de monitorizare media în adevaratul sens al cuvantului, ci să ofere câteva din funcţionalităţile de bază într-un mod complet automatizat și gratuit. Serviciile de acest gen sunt oferite acum contra cost de o multitudine de firme (în general de către agenţiile de PR), însă de cele mai multe ori selecţia și analiza se face manual prin citirea individuală a fiecărui articol și împărţirea pe categoriile de interes - în general fiind vorba de o clasificare pe bază de cuvinte cheie. Deasemenea, la cerere, se pot creea rezumate, astfel încât clientul să nu fie nevoit să citească toată presa din ziua respectivă, ci doar un rezumat ce conţine strict informaţiile necesare.

Sistemul prezentat va extrage toate articolele din presa online (la momentul actual, peste 53 de publicaţii), le va grupa în funcţie de subiectul dezbătut, le va ordona în funcţie de popularitate, va extrage cuvintele cheie, va măsura pozitivitatea articolului și de asemenea va creea rezumate atât multidocument cât și single-document.

Sistemul automat de monitorizare media folosește un set extins de algoritmi, atât din punctul de vedere al algorimilor de procesare a limbajului natural cât și al procesării în timp real al textelor de pe internet:

Pentru a oferi o imagine de ansamblu a ce se întâmplă în media, știrile trebuie extrase suficient de frecvent pentru a fi la curent cu tot ceea ce se întâmplă (fiind vorba de publicaţii online, se poate întâmpla ca un articol să fie completat pe parcurs, pe măsură ce apar informaţii noi), dar suficient de rar pentru a nu supra-aglomera sistemul. De asemenea, având în vedere că anumite știri vor fi acoperite de către toate publicaţiile, trebuie utilizat un algoritm de clasificare semantică care să fie capabil să grupeze documentele în funcţie de subiectul dezbătut în timp real. Un subiect de maxim interes va fi descris de majoritatea publicaţiilor, așa că de multe ori se dorește așezarea știrilor într-o ordine cronologică, pentru a se putea determina ce publicaţie a fost prima care a publicat știrea respectivă și care a fost raportul gradual al celorlalte publicaţii de-a lungul investigării subiectului dezbătut.

Extragerea articolelor dintr-un set de publicaţii online, cu toate că este o sarcină aparent simplă, ascunde suficiente sarcini încât să nu fie un pas ușor de ignorat.

Varianta folosită în implementarea noastră conţine următorii pași:

- abonarea la feed-urile RSS al tuturor acestor pubicaţii
- parsarea feed-ului atunci când se primește ceva nou
- descărcarea articolelor și verificarea daca este vorba despre un update sau un articol nou. Având în vedere că majoritatea ziarelor de pe internet sunt gratuite, în general fiecare pagină va conţine cel putin un spaţiu publicitar ce trebuie eliminat din analiza textului. Acest lucru se face prin diverși algoritmi euristici și poate diferi de la ziar la ziar.
- inserarea articolelor în baza de date spre analiză

O dată ce avem articolele extrase și grupate în funcţie de subiectul dezbătut, se poate determina daca este un articol ce are o tentă pozitivă sau negativă. în mod evident, algoritmul de analiză a sentimentelor are o tentă generală aici, deoarece va lucra pe un set extrem de diferit de documente, de la știri ce prezintă decese sau dezastre naturale până la știri ce prezintă variaţii de preţ la carburant sau la diferite produse.

Deasemenea, un alt subiect de interes este sumarizarea documentelor. Având în vedere că este vorba despre un sistem de monitorizare media, ne interesează in primul rând sumarizarea multi-document, astfel încât un utilizator al acestui sistem să nu fie nevoit să parcurgă toate articolele pentru a-și face o părere despre subiectul respectiv.

În continuare vom prezenta toate resursele disponibile pentru limba română.

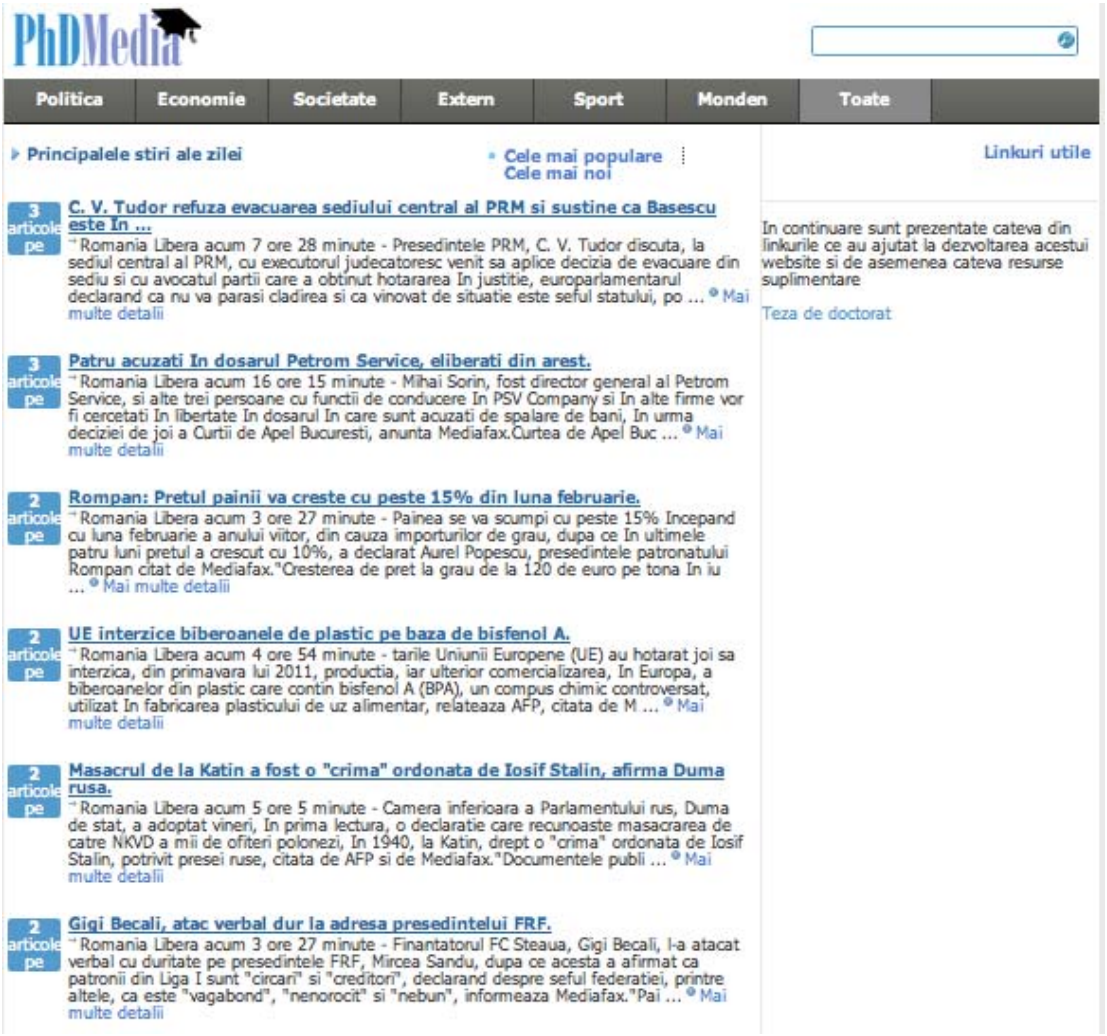
Framework NLP

Așa cum am precizat încă de la început, unul din scopurile principale ale lucrării a fost oferirea unui set de resurse sintactice și semantice pentru limba română, atât cod sursă și librării cât și fișiere de date.

- in primul rând se pot regasi implementări pentru toţi algoritmii descriși în lucrare
- diferite resurse tehnice pentru pre-procesarea textelor scrise in limba roamana (e.g. transformarea unui text scris cu diacritice intr-un text ușor digerabil de algoritmii dezvoltaţi în această lucrare
- întreaga arhitectura (inclusiv codul sursă) al aplicaţiei de monitorizare media

Ca și resurse lingvistice, se pot găsi următoarele elemente:

- expresii de concluzionare - acestea se pot dovedi utile atat pentru algoritmi de sumarizare, cât si pentru algoritmi dedicaţi determinării subiectelor dezbatute în articolele analizate. Multe lucrari de specialitate se folosesc de aceste expresii de concluzionare pentru efectuarea diferitelor task-uri, insa la momentul actual exista foarte putine astfel de resurse pentru limba romana. De aici și inițiativa de a pregăti un set de resurse pentru oricine dorește să continue sau să îmbunătăţească cercetarea în acest domeniu, în special pentru limba română. Exemple de expresii de concluzionare: "în concluzie", "pe scurt', "această lucrare", etc.
- acronime - în analiza unui text, este foarte important să se poată determina că este vorba despre același lucru atunci cand se folosește ori titlul complet al unei expresii și acronimul său. De asemenea, sunt foarte puţine resurse ce conţin acronime dedicate pentru România
- abrevieri - în cazul știrilor publicate în presă sau în articolele ştiinţifice nu este cazul, dar se poate întampla ca într-unul din textele analizate să fie nevoie de înțelegerea abrevierilor pentru a se putea lucra mai ușor.



In interfaţa de bază ofera următoarele funcţionalităţi:

- Clustering semantic pe stirile din ultimile şapte zile şi ordonarea acestora după numărul de elemente din fiecare categorie
- listă cu linkuri utile despre algoritmii folosiţi în acest proiect
- Posibilitatea selectării doar a tipurilor de ştiri de interes (e.g. Politică, Economie, etc)
- Pentru fiecare cluster, se pot afişa toate ştirile din clusterul respectiv
- Butonul \textit{Mai Mult}, returnează prima ştire din cluster, din punct de vedere cronologic

Se poate face de asemenea şi o selecţie după cele mai populare sau cele mai recente ştiri.

Pentru calculul distanţei dintre două texte se pot folosi o multitudine de algoritmi, însă cel mai rapid s-a dovedit a fi o combinaţie între mai mulţi algoritmi rapizi: distanţa Jacard pe toate cuvintele documentului (din care excludem totuşi cuvintele uzuale, distanţa Jacard doar pentru cuvintele care încep cu majusculă, şi un parametru euristic ce depinde de diferenţ de numărul de cuvinte.

O listă incipientă de cuvinte uzuale poate fi regăsită mai jos, însă pe site poate fi găsită o lista completa: *este cele din despre deci deodata când care celor ceea cei celalalt doamna domnul doamnei are tot domnului mi-aş multe multa mi-l mult n-a*

asta chiar dacă dar după pentru în însuși doar aștia între cât îl însă într-o într-un între își și cine ce cum unde când a acest îmi prin acesta acestuia lui un unde pe sus jos acestea acestora aceșteia număr numarul unu una primul prima de o unu doi trei patru cinci șase șapte opt nouă zece una doua treia patra cincea șasea șaptea opta noua zecea oarecare mie să astfel așa.

Extragerea textului corect din presă a fost un efort în sine, deoarece trebuiesc eliminate toate tag-urile HTML, toate scripturile javascript și orice altceva mai poate influența textul final și implicit rezultatele finale. Pentru determinarea celor mai recente articole din presă m-am folosit de feed-urile RSS ale fiecărei publicații, iar următorul script extrage linkurile către toate articolele noi primind ca și date de intrare link-ul către feed-ul RSS.

Alte resurse dezvoltate si adaptate pentru limba română vor fi disponibile pe website si de asemenea tool-uri noi vor fi adaugate in mod frecvent. De asemenea, cele două script-uri prezentate mai sus rulează din februarie 2010, adunând o cantitate foarte mare de date.

Siri

Există foarte multe probleme încă nerezolvate în procesarea limbajului natural, însă în același timp există și foarte multe probleme care au fost rezolvate. Un exemplu de success este Siri, asistentul digital introdus de compania Apple pentru toate dispozitivele mai noi de iPhone 4s. Sistemul a fost initial dezvoltat de Nuance și ulterior cumpărat de Apple. Prima versiune de Siri era capabilă să facă rezervări la diferite restaurante, sa cumpere bilete la film sau să comande un taxi la locația curentă a utilizatorului și totul doar prin oferirea de comenzi vocale în limbaj natural. Ultima versiune este însă capabilă de mult mai multe acțiuni: poate interacționa cu aplicațiile de vreme, mesagerie, email, calendar, contacte, muzica, notițe, browser si hărți. De asemenea, Siri funcționează în următoarele limbi: Engleză, Franceză, Germană, Japoneză, Italiană, Mandarină, Koreană și Cantoneză.

În mod evident, în spatele Siri există un conglomerat de tehnologii din sfera procesării limbajului natural. Dacă în primă fază este vorba doar de transformarea din voce în text, în pașii următori se aplică o mare parte din tehnologiile prezentate în această lucrare.

Ca și mod de funcționare, \în momentul in care utilizatorul transmite o intrebare, Siri înregistrează cerința utilizatorului în format audio și o trimite la serverele sale, unde din fișier audio devine o întrebare în format text.

Pasul următor este înțelegerea sensului întrebării. De exemplu, știm foarte clar că orice informație poate fi prezentată în mai multe feluri. "Aș vrea un croisant", "Exista cumva vreo brutărie prin apropiere?" sau "Mi-ar plăcea niște produse de patiserie franțuzească", toate converg în aceeași direcție. În mod normal, pasul următor ar fi fost împărțirea propoziției în părțile de vorbire corespunzătoare. Însă Siri nu funcționează așa. În loc să modeleze concepte lingvistice, sistemul folosit a fost modelarea obiectelor reale. De exemplu, la cererea "As vrea sa văd un thriller", Siri ar identifica imediat "thirller" ca și gen de film și ar începe să caute într-o listă de filme, \în loc să caute cum este subiectul conectat de verb.

Siri este capabilă să mapeze conţinutul unei întrebări într-un domeniu de posibile acţiuni şi apoi să o aleagă pe cea mai probabilă, bazat pe "înţelegerea ei" asupra legăturilor dintre concepte reale. De exemplu, Siri ştie că un restaurant trebuie să aibe o adresă, o notă, un tip de bucătărie si o gamă de preţ. De asemenea, inclusiv legat de partea de transformare din audio in text. Creatorii Siri au explicat asta printr-un exerciţiu simplu de imaginaţie. Teoretic, ar trebui să ne imaginam un angajat în cadrul unui hotel. O întrebare "closest coffee shop" poate suna la fel de uşor cu "closest call Felicia" daca hotelul este foarte aglomerat şi sunt foarte multe persoane care discută în acelaşi timp. Însă în acelaşi timp, avaând în vedere contextul prezentat, este mult mai probabil ca "closest" să se refere la un loc decât la o persoană şi este de asemenea mult mai probabil ca angajatul sa fie întrebat despre recomandari de cină decât rugat să contacteze o persoană. La fel funcţionează şi Siri, care a fost concepută mai degraba ca un asistent personal decât o persoană cu care ar trebui să ai conversaţii complexe.

Concluzii Parţiale

Aşa cum am vazut de-a lungul tehnicilor explorate, putem observa o schimbare de paradigmă. De-a lungul zecilor de ani pretrecuţi în studierea procesarii limbajului natural, majoritatea tehnicilor dezvoltate au fost concepute pentru a compensa lipsa puterii de procesare de la momentul respectiv. Tehnici care păreau atunci evidente si foarte uşor de implementat nu puteau fi folosite cu adevarat în practică datorită complexitaţii algoritmilor, a cantităţii uriaşe de date necesare şi a puterii de procesare. Însă în ziua de azi, având în vedere progresul uimitor în materie de hardware şi al cloud computing-ului, putem observa cum tehnici foarte simple (poate chiar vechi) sau cel putin suficient de intuitive pot fi folosite cu o rată mult mai mare de succes decât algoritmi foarte complecşi.

În ziua de azi, posibilităţile tehnice sunt mult mai avansate, ceea ce permite utilizarea unor algoritmi atât intuitivi cat şi foarte intenşi din punct de vedere procesare şi în acelaşi timp să obţinem un timp de execuţie foarte bun. Tehnologiile din spatele Siri sunt în general euristice şi se bazeaza pe cantităţi foarte mari de date şi pe un data-center dedicat. Însă în acelaşi timp, rezultatele sunt statistic întotdeauna 77 la sută corecte iar timpul de procesare este insesizabil.

În platforma de pe site, se găsesc atât implementări adaptate pentru limba română a algoritmilor existenţi, cât şi implementari pentru algoritmi mai simplii, cu rezultate mai bune, însă care necesită resurse hardware mai complexe.

De asemenea, o parte din o mare parte din algoritmii folosiţi în platforma de sumarizare ofera rezultate atat de bune deoarece infrastructura pe care rulează este suficientă pentru cerinţele proiectului.

Concluzii finale, contribuţii originale și direcţii viitoare de cercetare

Așa cum am vazut de-a lungul tehnicilor explorate, putem observa o schimbare de paradigmă. De-a lungul zecilor de ani pretrecuţi în studierea procesarii limbajului natural, majoritatea tehnicilor dezvoltate au fost concepute pentru a compensa lipsa puterii de procesare de la momentul respectiv. Tehnici care păreau atunci evidente si foarte ușor de implementat nu puteau fi folosite cu adevarat în practică datorită complexitaţii algoritmilor, a cantităţii uriașe de date necesare și a puterii de procesare. Însă în ziua de azi, având în vedere progresul uimitor în materie de hardware și al cloud computing-ului, putem observa cum tehnici foarte simple (poate chiar și foarte vechi) sau cel puţin suficient de intuitive, pot fi folosite cu o rată mult mai mare de succes decât algoritmi foarte complecși.

În ziua de azi, posibilităţile tehnice sunt mult mai avansate, ceea ce permite utilizarea unor algoritmi atât intuitivi cat și foarte intenși din punct de vedere procesare și în același timp să obţinem un timp de execuţie foarte bun. Așa cum am văzut în capitolul 5, tehnologiile din spatele Siri sunt în general euristice și se bazeaza pe cantităţi foarte mari de date și pe un data-center dedicat. Însă în același timp, rezultatele sunt statistic întotdeauna 77 la sută corecte iar timpul de procesare este insesizabil.

În platforma de pe site, se găsesc atât implementări adaptate pentru limba română a algoritmilor existenţi, cât și implementari pentru algoritmi mai simplii, cu rezultate mai bune, însă care necesită resurse hardware mai complexe.

De asemenea, o parte din o mare parte din algoritmii folosiţi în platforma de sumarizare ofera rezultate atat de bune deoarece infrastructura pe care rulează este suficientă pentru cerinţele proiectului.

Clasificarea documentelor

Majoritatea tehnicilor prezentate în capitolul 2 sunt în general statistice, așa că sunt destul de puţine îmbunătăţiri care pot fi aduse algoritmilor existenţi. Rezultatele sunt în general la fel de bune dacă corpusul de antrenare este la fel de bun și la fel de complex, ceea ce ne face să concluzionam că dacă reușim să avem un corpus de antrenare exhaustiv, atunci implicit și rezultatele vor fi pe măsură.

De asemenea, așa cum se poate observa și din exemplele de mai sus, gradul de complexitate și necesarul de resurse este direct proporţional cu numărul de cuvinte din corpusul de antrenare și în mod implicit cu dimensiunea corpusului. De asemenea, în funcţie de timpul la care se face referire sau în care se petrece acţiunea oferă un grad de variaţie destul de larg atât verbelor cât și celorlalte cuvinte prin variaţia diacriticelor, a genului si a persoanei. De exemplu, variaţia [acest, această, acesta, aceștia, acestea, acești, aceste], cu toate că reprezintă mereu același lucru (un substitut pentru un anumit obiect/persoană), variaza în funcţie de număr și sex.

În funcţie de problema care se dorește a fi rezolvată, se pot efectua diferite operaţii de preprocesare. Așa cum am văzut în inclusiv rolul sintactic al fiecărui cuvânt poate avea un rol esenţial în clasificarea documentelor.

Când vorbim de clasificarea semantică și a modului în care poate fi ea aplicată, trebuie să fim foarte atenţi la problema pe care trebuie să o rezolvăm și abia pe urmă să decidem ce algoritm se potrivește mai bine, sau ce corpus de antrenare este necesar. De exemplu, majoritatea algoritmilor prezentaţi mai sus se bazează pe un model de învăţare supervizată, care în mod evident duce la rezultate mult mai bune decât în cazul unei antrenări nesupervizate, însă la cantitatea din ziua de astăzi de documentaţie și de informaţii, problema pregătirii corpusului de antrenare se complică exponenţial.

Să luăm următorul exemplu: dorim să creăm un sistem automat de monitorizare media, care este capabil să:

- Preia din presa online toate articolele cu o anumita frecvenţă prestabilită. Statistic, frecvenţa optimă este din oră în oră
- Cum multe ziare vor acoperi același subiect, sistemul va trebui să fie capabil să creeze clustere de articole care tratează același topic/categorie
- Majoritatea publicaţiilor oferă o preclasificare a articolelor în categorii precum economic, politic, etc, însă acest lucru este insuficient în cadrul unui sistem de monitorizare media, deoarece se dorește o filtrare la nivel de topic și nu la nivel de categorie
- Pentru fiecare topic, sistemul este capabil să determine daca este vorba despre un topic pozitiv sau negativ. Acest lucru este foarte important în cazul review-ului de produse, al politicii, etc
- De asemenea, pentru fiecare topic în parte, pe baza articolelor din ziare diferite care ating același subiect, să se creeze rezumatul

La volatilitatea știrilor din ziua de azi, crearea unui corpus de antrenare la nivel de categorie (politic, financiar, tabloid, etc) este un lucru trivial, însă nu se poate spune același lucru si despre clasificarea la nivel de topic.

O bună metodă de creare a unui corpus de antrenare la nivel de topic s-a dovedit a fi Similaritatea Jaccard, care cu toate că pare destul de simplă, s-a dovedit în testele mele, în special pe zona clasificării articolelor din presa online, a avea rezultate foarte bune.

Distanţa Jaccard are rezultate în general foarte bune, dar în capitolul 5 vom arata diferite metode de îmbunătăţire, prin adaugarea de semidistanţe Jaccard (numărul de apariţii al fiecărui cuvânt, dimensiunea documentelor, etc), care ne vor duce gradul de acurateţe care în mod normal la distanţa Jaccard clasică este undeva între 70-80 la sută. în cazul testelor mele, în gama 98-100 la sută. Însă și distanţa Jaccard clasică poate fi folosită daca dorim doar crearea unui corpus de antrenare.

Analiza Sentimentelor

Așa cum am observat în capitolul 3, analiza sentimentelor bazate pe sentimentele asociate fiecărei componente sau subcomponente ale obiectului analizat. De asemenea, am observat că în momentul în care facem această analiză, este foarte important să ne uitam și la zona geografică, deoarece se pot observă diferenţe de gusturi. Cel mai elocvent exemplu este vânzarea de telefoane iphone 6 și iphone 6

plus: dacă în Statele Unite și în Europa iphone 6 a stat de aproape 6 ori mai bine decât iphone 6 plus, însă nu același lucru se poate spune și despre Asia, unde iphone 6 plus a fost un succes. S-a observat ca ecranul mai mare este un avantaj pentru piaţa din Asia faţă de cea din Europa sau Statele Unite.

Dacă ar fi să facem această analiză la nivel mondial și să coroborăm rezultatele cu cele de vânzări, atunci am putea concluziona că iphone 6 este net superior faţă de iphone 6 plus, iar ca principal argument ar fi dimensiunea foarte incomodă a telefonului și a faptului ca este mai degrabă o tabletă decât un telefon și poate în unele situaţii inclusiv preţul poate fi un argument. Spre exemplu, dacă un telefon iphone valorează în Statele unite suma de X USD, atunci în România acesta va valora X EUR, iar în UK X GBP și având în vedere diferenţa de curs valutar, atunci diferenţa de preţ este chiar una substanţială.

Putem concluziona că analiză sentimentelor la nivel mondial poate avea rezultate eronate sau poate chiar controversate. Ca sa-l citǎm pe Mark Twain: "There are three kinds of lies: lies, damned lies and statistics" ceea ce face foarte grea analiza sentimentelor la nivel mondial fără să ne uitam si la zona geografică.

Sistem Automat de Monitorizare Media

În capitolul 5, putem observa o schimbare de paradigmă. De-a lungul zecilor de ani pretrecuţi în studierea procesarii limbajului natural, majoritatea tehnicilor dezvoltate au fost concepute pentru a compensa lipsa puterii de procesare de la momentul respectiv. Tehnici care păreau atunci evidente si foarte ușor de implementat nu puteau fi folosite cu adevarat în practică datorită complexitaţii algoritmilor, a cantităţii uriașe de date necesare și a puterii de procesare. Însă în ziua de azi, având în vedere progresul uimitor în materie de hardware și al cloud computing-ului, putem observa cum tehnici foarte simple (poate chiar vechi) sau cel putin suficient de intuitive pot fi folosite cu o rată mult mai mare de succes decât algoritmi foarte complecși.

În ziua de azi, posibilităţile tehnice sunt mult mai avansate, ceea ce permite utilizarea unor algoritmi atât intuitivi cat și foarte intenși din punct de vedere procesare și în același timp să obţinem un timp de execuţie foarte bun. Tehnologiile din spatele Siri sunt în general euristice și se bazeaza pe cantităţi foarte mari de date și pe un data-center dedicat. Însă în același timp, rezultatele sunt statistic întotdeauna 77 la sută corecte iar timpul de procesare este insesizabil.

În platforma de pe site, se găsesc atât implementări adaptate pentru limba română a algoritmilor existenţi, cât și implementari pentru algoritmi mai simplii, cu rezultate mai bune, însă care necesită resurse hardware mai complexe.

De asemenea, o parte din o mare parte din algoritmii folosiţi în platforma de sumarizare ofera rezultate atat de bune deoarece infrastructura pe care rulează este suficientă pentru cerinţele proiectului.

Contribuţii personale

În capitolul 5 am construit un sistem automat de monitorizare media specializat pentru limba română, care implementează următoarele funcţii: topic clustering,

sentiment analysis și sumarizare. Toate aceste funcţii sunt oferite gratuit utilizatorilor, pentru un număr nelimitat de conturi și un număr nelimitat de cuvinte cheie. Majoritatea sistemelor de monitorizare media sunt excesiv de scumpe și acest lucru este nejustificat pentru anul 2014. Cu toate că acest lucru va produce o multitudine de discuţii în această zonă, trebuie luată o inițiativă.

De asemenea, pentru partea de clasificare semantică, există foarte multe probleme ce trebuie rezolvate. În articolele scrise despre clasificarea documentelor am rezolvat probleme concrete, în special din zona de antispam, care au dus mai departe la 4 patente și nenumărate articole despre acest subiect, în special în zona de clasificare documente și împărţirea pe categorii.

În zona de procesare a limbajului natural, în 2010, am creat doua start-up-uri, în care puteam oferi căutarea de restaurante în limbaj natural "aș vrea o ciorbă de burtă ieftină în drumul taberei" sau "vreau o ciorbă de fasole în apropiere" sau oferirea unui algoritm de măsurare a influenţei blogerilor în funcţie de cum se propagă informaţia între bloguri.

Direcţii Viitoare de Cercetare

Având în vedere resursele care vor fi puse la dispoziţie pe site, se poate continua în orice direcţie de cercetare în zona de NLP pentru limba română. În sistemul de monitorizare media dezvoltat, pe partea de clasificare semantică am obtinut rezultate foarte bune folosind un corpus de antrenare generat cu distanţa Jaccard a.î. să putem avea o învăţare supervizată.

Ca și resurse lingvistice disponibile pe site, se pot găsi următoarele elemente:

- expresii de concluzionare - acestea se pot dovedi utile atat pentru algoritmi de sumarizare, cât si pentru algoritmi dedicaţi determinării subiectelor dezbatute în articolele analizate. Multe lucrari de specialitate se folosesc de aceste expresii de concluzionare pentru efectuarea diferitelor task-uri, insa la momentul actual exista foarte putine astfel de resurse pentru limba romana. De aici și inițiativa de a pregăti un set de resurse pentru oricine dorește să continue sau să îmbunătăţească cercetarea în acest domeniu, în special pentru limba română. Exemple de expresii de concluzionare: "în concluzie", "pe scurt', "această lucrare", etc.
- acronime - în analiza unui text, este foarte important să se poată determina că este vorba despre același lucru atunci cand se folosește ori titlul complet al unei expresii și acronimul său. De asemenea, sunt foarte puţine resurse ce conţin acronime dedicate pentru România
- abrevieri - în cazul știrilor publicate în presă sau în articolele științifice nu este cazul, dar se poate întampla ca într-unul din textele analizate să fie nevoie de înţelegerea abrevierilor pentru a se putea lucra mai ușor.

De asemenea, așa cum am precizat încă de la început, unul din scopurile principale ale lucrării a fost oferirea unui set de resurse sintactice și semantice pentru limba română, atât cod sursă și librării cât și fișiere de date.

- în primul rând se pot regasi implementări pentru toţi algoritmii descrişi în lucrare
- diferite resurse tehnice pentru pre-procesarea textelor scrise in limba roamana (e.g. transformarea unui text scris cu diacritice intr-un text ușor digerabil de algoritmii dezvoltaţi în această lucrare
- întreaga arhitectura (inclusiv codul sursă) al aplicaţiei de monitorizare media

Aşadar, se poate merge în mai multe direcţii de cercetare, însă eu personal o v-oi continua în crearea unui asistent bazat pe comunicarea audio.

Referinte

1. T. K. Landauer, P. W. Foltz and D. Laham, Discourse Processes, Pages 259-284, Introduction to Latent Semantic Analysis, 1998
2. G. Golub and W. Kahan, J. Siam, Calculating the Singular Values and Pseudo-Inverse of a Matrix, 1965
3. Scott Deerwester, Susan T. Dumais, George W. Furnas and Thomas K. Landauer, Richard Harshman, Indexing by Latent Semantic Analysis, 1990
4. Maciej Ceglowski, Aaron Coburn, and John Cuadrado, National Institute for Technology and Liberal Education Middlebury College, Semantic Search of Unstructured Data using Contextual Network Graphs, 2003
5. Gibson W, Ace Book, ISBN: 0-441-56959-5, Neuromancer, 1984
6. Tufis Dan, Filip Florin Gh, Limba Româna în Societatea Informationala - Societatea Cunoasterii, 2002
7. Hristea Florentina, Editura Universitatii din Bucuresti, Introducere in procesarea limbajului natural cu aplicatii in Prolog, 2000
8. Chomsky Noam, Mouton and Co, Language Syntax, 1957
9. Claudiu Mihaila, Sumarizare Automata focalizata Temporar, 2008
10. Regina Barzilay and Michael Elhadad, Mathematics and Computer Science Dept, Ben Gunon University, Using Lexical Chains for Text Summarization, 2001
11. Xiaojun Wan, Jianguo Xiao, Institute of Computer Science and Technology, Peking University, Beijing, Single Document Keyphrase Extraction Using Neighborhood Knowledge, 2008
12. Bruce Krulwich and Chad Burkey, Center for Strategic Technology Research Andersen Consulting LLP, Learning user information interests through the extraction of semantically significant phrases, 1996
13. Ken Barker and Nadia Cornacchia, School of Information and Technology Engineering, University of Ottawa, Noun Phrase Heads to Extract Document Keyphrases, 2000
14. Alberto Munoz, Compound Key Word Generation from Document Databases Using A Hierarchical Clustering ART Model, Intell, 1997
15. Rada Mihalcea and Paul Tarau, Department of Computer Science, University of North Texas, TextRank: Bringing Order into Texts, 2004
16. Andrew H. Morris, George M. Kasper, Dennis A. Adams, Advances in Automated Text Summarization, The effects and Limitations of Automated Text Condensing on Reading Comprehension Performance, 1989

17. Gerard Salton, Amit Singhal, Mandar Mitra Chirs Buckley, Advances in Automated Text Summarization, Automatic Text Structuring and Summarization, 1989
18. Andrew Goldberg, CS838-1 - Advanced NLP, Automatic Summarization, 2007
19. Eduard Hovy and Chin-Yew Lin, Advances in Text Summarization, Automatic Text Summarization in Summarist, 1999
20. George Petre, SpamConference 2010, Facebook - Another Breach in the Wall, 2010
21. Edgar GONZÀLEZ, Maria FUENTES, TALP Research Center, Spain, A New Lexical Chain Algorithm Used for Automatic Summarization, 1999
22. Catalin Cosoi, Madalin Vlad, Valentin Sgarciu, DAAAM 2008, proceedings, pg.0309-0310, indexat ISI Web of Knowledge, Introducing Syntactic Features into a Bayesian Classifier, 2008
23. Kyoung-Min Kim, Jin-Hyuk Hong, Sung-Bae Cho, Information Processing and Management 43 (2007) 225–236, A semantic Bayesian network approach to retrieving information with intelligent conversational agents, 2007
24. Sylvain Auroux, Histoire Epistemiologie Langage, La notion de linguistique generale, 1988
25. Noam Chomsky, Lectures on Government and Binding, 1981
26. W. Gerrod Parrott, Emotions in Social Psychology: Key Readings, 2000
27. Nitin Indurkhya, Fred J. Damerau, Handbook of Natural Language Processing, 2000
28. Bing Liu, Sentiment Analysis and Opinion Mining, Sentiment Analysis and Opinion Mining, 2012
29. L. Ku, Y. Liang and H. Chen, Proceedings of AAAI-2006, Opinion extraction, summarization and tracking in news and blog corpora, 2006
30. Li Zhuang, Feng Jing, Xiao-Yan Zhu, Proceedings of the 15th ACM international conference on Information and knowledge management, Movie review mining and summarization, 2006
31. Hatzivassiloglou, V. and K. McKeown., Proceedings of Annual Meeting of the Association for Computational Linguistics, Predicting the semantic orientation of adjectives, 2007
32. Wilson et al, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysiz, 2007
33. Kanayama and Nasukawa, Proc. of EMNLP'06, pp. 355-363, Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis, 2006
34. Qiu et al, Stud Health Technol Inform., Understanding the psychological motives behind microblogging, 2010
35. Inderjeet Mani, Mark T. Maybury, Cambridge, MA: The MIT Press,, Advances in Automatic Text Summarization, 1999
36. Bo Pang, Lillian Lee, Proceedings of ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, 2004
37. Gerard Salton, Amit Singha, Proceedings of ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Information Processing & Management, 1997
38. GChinatsu Aonet Mary Ellen Okurowskit James Gorlinskyt Bjornar Larsent, SRA Internatlon, A Scalable Summarization System Using Robust NLP, 2003

39. E Nistor, E Roman, Revue Roumaine de Linguistique, Attempts at Automatic Detection of Some Semantic Deviations, 1980
40. Luhn H. P., IBM Journal of Research, Automatic Creation of Literature Abstracts, 1959
41. Cremmins, Edward T., ISI Press, The Art of Abstracting, 1982
42. Inderjeet Mani, Eric Bloedorm, AAAI '98/IAAI '98 Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence Pages 820-826 , Machine learning of generic and user-focused summarization, 1982
43. Barker, K. and Cornacchia, N,Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pages 40–52, Using noun phrase heads to extract document keyphrases, 2000
44. Daniel Marcu, The MIT Press - ISBN 0-262-13372-5, The Theory and Practice of Discourse Parsing and Summarization, 2000
45. Mark Twain, The Chapters from the North American Review, Mark Twain's Own Autobiography, 1898
46. AC Cosoi, MS Vlad, V Sgarciu, US Patent 8,131,655, 2012, Spam filtering using feature relevance assignment in neural networks, 2007
47. AC Cosoi, MS Vlad, V Sgarciu, MIT Spam Conference, Asigning Relevancies to Individual Features for Large Patterns in Artmap Networks, 2007
48. AC Cosoi, Proc. MIT Spam Conf, A False Positive Safe Neural Network; The Followers of the Anatrim Waves, 2008
49. AC Cosoi, MS Vlad, V Sgarciu, AQTR 2008, vol.III, Proceedings, pg.230-233, indexat? ISI Web of Knowledge, On neural networks and the future of spam, 2008
50. AC COSOI CM COSOI V SGARCIU B DUMITRU, Proceedings of DAAAM 2009, Viena, Austria, 25-28 nov, pp.103-104, ISSN 1726-9679, ISBN 978-3-901509-70-4, indexat? ISI Web of Knowledge, The Romanian Social Fractal, 2009
51. AC Cosoi, Virus Bulletin, The medium or the message? Dealing with image spam, 2006
52. Maria Corduneanu Carmen Maria Cosoi Catalin Alexandru Cosoi Madalin Vlad Valentin Sgarciu, INFORMACIJSKA DRU?BA ? IS 2009, A HYBRID NEURAL NETWORK MODEL FOR SPAM DETECTION, 2009
53. C.Cosoi; M.St.Vlad; V.Sgarciu, Annals of DAAAM for 2007 & Proceedings of the 18th International DAAAM Symposium, ISBN 3-901509-58-5,vol. 18, pp 617-618, october 2007, indexat? ISI Web of Knowledge, Using a Neural Network to Determine a Cat's Emotional State by its Meow, 2007
54. Catalin Cosoi, USPTO 8695100, Systems and methods for electronic fraud prevention, 2014
55. Catalin Cosoi, USPTO 8572184, Systems and methods for dynamically integrating heterogeneous anti-spam filters, 2013
56. Catalin Cosoi, USPTO 8335383, Image filtering systems and methods, 2012
57. Catalin Cosoi, Madalin Vlad, Valentin Sgarciu, USPTO 8131655, Spam filtering using feature relevance assignment in neural networks, 2012
58. Musat Claudiu, Catalin Cosoi, USPTO 8010614, Systems and methods for generating signatures for electronic communication classification, 2011
59. Catalin Cosoi, USPTO 7751620, Image spam filtering systems and methods, 2010

60. Nasui D.V; Cosoi A.C; Sgarciu V; Rancea I,Proceedings of DAAAM 2009, Viena, Austria, 25-28 nov, pp.1153-1154, ISSN 1726-9679, ISBN 978-3-901509-70-4, indexat? ISI Web of Knowledge, Using Wireless Monitoring for Market Research Interviewers, 2009

61. Gams M; Cosoi A.C; Corduneanu M; Kolbe M; Vlad M.S; Sgarciu V, Proceedings of DAAAM 2009, Viena, Austria, 25-28 nov, pp.1415-1416, ISSN 1726-9679, ISBN 978-3-901509-70-4, indexat? ISI Web of Knowledge, A New Method for Fingerprint Clasification, 2009
62. C.Cosoi, V.Sgârciu, M.St.Vlad, Annals of DAAAM for 2007 & Proceedings of the 18th International DAAAM Symposium, ISBN 3-901509-58-5,vol. 18, pp 615-616, october 2007, indexat ISI Web of Knowledge, Asigning Relevancies to Individual Features for Large Patterns in Artmap Networks, 2007
63. Cosoi A.C.; Sgarciu V.; Vlad M.S, DAAAM 2008, proceedings, pg.0307-0308, indexat? ISI Web of Knowledge, Build Your Antiphishing Technology in Just 5 Minutes, 2008
64. Cosoi AC, Vlad MS, Sgarciu V, DAAAM 2008, proceedings, pg.0311-0312, indexat? ISI Web of Knowledge, Heuristic Image Recognition, 2008
65. Cosoi C.A; Cosoi C.M, Sgarciu V; Dumitru B, Vlad M.S, Proceedings of DAAAM 2009, Viena, Austria, 25-28 nov, pp.105-106, ISSN 1726-9679, ISBN 978-3-901509-70-4, indexat ISI Web of Knowledge., Spam / Twitter, 2009
66. Cosoi, A. C.; Vlad M. S. & Sgarciu, V, DAAAM International Scientific Book 2009, VOL.8, pp. 377-384 CHAPTER 39, ISBN 978-3-901509-69-8, ISSN 1726-9687; aparitie 2010, EBSCO bibliographic database, Heuristic Image Recognition, 2009
67. C.Cosoi; M.S.Vlad; V.Sgarciu , preprints, vol.3, CSCS 16, Bucharest, 22-25 May 2007, pg.33-37, ISBN 978-973-718-741-3 sau ISBN 978-973-718-744-4, Image Retrieval System Inspired from Antispam Filters, 2007
68. A.C.Cosoi; B.Dumitru; V.Sgarciu; M.St.Vlad; M.Corduneanu, IAFA 2009, vol.3, pp 36-40, ISSN: 2066-4451, The Romanian Blogosphere as a Social Fractal, 2009