Tema 3 IA

Sumarizarea documentelor

TUDOR Costin-Cristian
GRUPA 342B2, INGINERIA SISTEMELOR

Cuprins

Introducere

Pentru început, am implementat citirea articolelor, împărțirea acestora pe categorii de știri și asocierea cu rezumatul propus. De asemenea, am creat, folosindu-mă de biblioteca *nltk*, funcțiile necesare pentru preprocesarea textului, împărțirea în propoziții și n-grame, precum și o funcție de evaluare a performanțelor, care calculează metricile *BLEU1, ROUGE1, BLEU2, ROUGE2, BLEU4* si *ROUGE4*. Am împărțit setul de date așa cum a fost propus în cerință (75% pentru antrenare și 25% pentru testare) și am pus la dispoziție parametrul *“shuffle”*, care poate fi setat pe *True* pentru a alege la întâmplare documentele care vor fi folosite pentru antrenare/testare.

Apoi, am implementat metodele Naive Bayes și TF-IDF pentru unigrame, bigrame și 4-grame. Pentru metoda Naive Bayes, este posibilă setarea parametrilor stop words și lematizare, iar pentru TF-IDF se poate opta pentru calcularea scorului doar pentru substantive, setarea unei ponderi pentru similaritatea cu titlul și ponderarea propozițiilor în funcție de poziția acestora în text. Metoda Naive Bayes va genera rezumatul concatenând toate propozițiile clasificate ca parte a acestuia, pe când metoda TF-IDF va returna primele k propoziții cu cel mai mare scor din articol, unde:

$$\boxed{k = \frac{\text{nr. propoziții din articol} \cdot \text{nr. propoziții în rezumatele din setul de date de antrenare}}{\text{nr. propoziții în articolele din setul de date de antrenare}}}$$

Analiza rezultatelor

Pentru a realiza analiza rezultatelor, am calculat metricile *BLEU* si *ROUGE* pentru fiecare metodă și categorie de știri în parte, apoi am generat două tabele (unul pentru Naive Bayes și unul pentru TF-IDF) în care am salvat performanțele pentru fiecare modificare în parte sub forma unui tabel HTML pe care l-am convertit într-un tabel dintr-o bază de date MySQL.

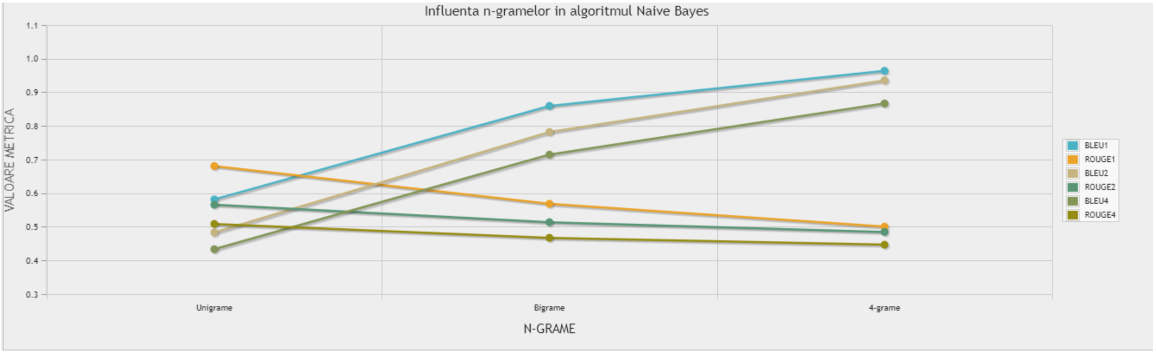
Pe baza acestor tabele, am scris interogări SQL pentru a determina influența pe care o are fiecare modificare asupra performanțelor algoritmilor și să creez grafice care să ilustreze cât mai bine această influență.

Metoda Naive Bayes

Pentru fiecare categorie de știri, am calculat metricile *BLEU* și *ROUGE* pentru unigrame, bigrame și 4-grame, în fiecare dintre cele 3 cazuri: stopwords incluse în text, stopwords excluse din text, stopwords excluse din text + lematizare.

Influența n-gramelor în metoda Naive Bayes

Pentru a observa mai ușor influența n-gramelor în metoda Naive Bayes, am calculat mediile metricilor pentru toate categoriile și am obținut următorul grafic, care îmi indica valorile metricilor BLEU și ROUGE în funcție de n-grame:



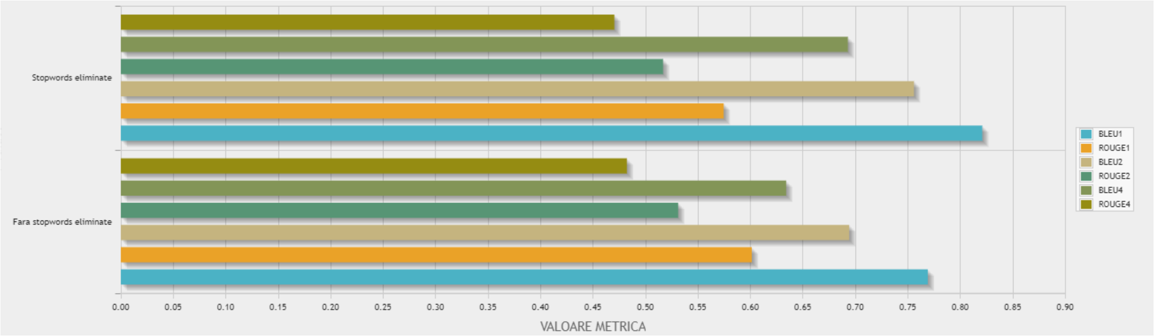
Pe baza acestui grafic se poate observa faptul că, odată cu creșterea n-gramelor, cresc și valorile pentru metrica *BLEU* (care măsoară precizia), însă scad puțin valorile pentru metrica *ROUGE* (care măsoară recall-ul).

Prin precizie, ne referim la cât de multe cuvinte/n-grame din rezumatul generat apar în rezumatul luat ca referință, iar prin recall, la cât de multe cuvinte/n-grame din rezumatul luat ca referință au aparut în rezumatul generat.

Deoarece creșterea metricii *BLEU* este mai abruptă decât scăderea metricii *ROUGE*, consider că am obținut cele mai bune rezultate pentru 4-grame.

Influența stop-words în metoda Naive Bayes

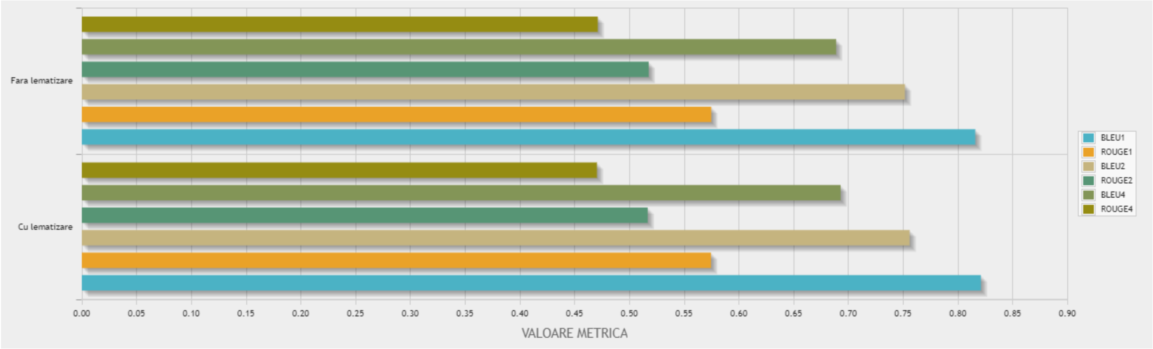
Pentru a realiza această comparație, am ignorat valorile pentru stopwords + lematizare și am calculat din nou mediile metricilor, obținând următorul grafic:



Și în acest caz, prin eliminarea de stopwords cresc valorile pentru metrica *BLEU*, însă scad foarte puțin cele pentru metrica *ROUGE*. Astfel, consider că eliminarea de stopwords aduce o îmbunătățire metodei.

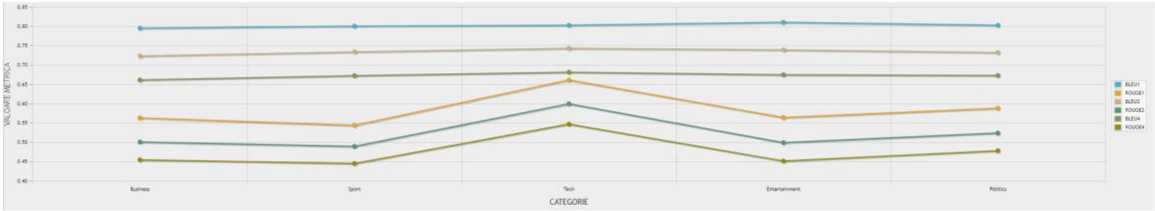
Influența lematizării în metoda Naive Bayes

Am comparat valorile obținute cu/fără lematizare (cu stopwords eliminate) în acest grafic:



Folosirea lematizării duce la o creștere a valorilor metricilor, însă această creștere este foarte mică (sub 1%).

Comparație între categoriile de știri în metoda Naive Bayes



Conform graficului generat, nu sunt diferențe mari între categoriile de știri, cu excepția creșterii semnificative a valorilor metricilor *ROUGE1*, *ROUGE2* și *ROUGE4* pentru categoria tech. Explicația logică la care m-am gândit pentru a explica această creștere este aceea că în categoria tech există mai multe cuvinte specifice care se regăsesc în majoritatea articolelor, iar modelul este mai bine antrenat pe aceste cuvinte.

Metoda TF-IDF

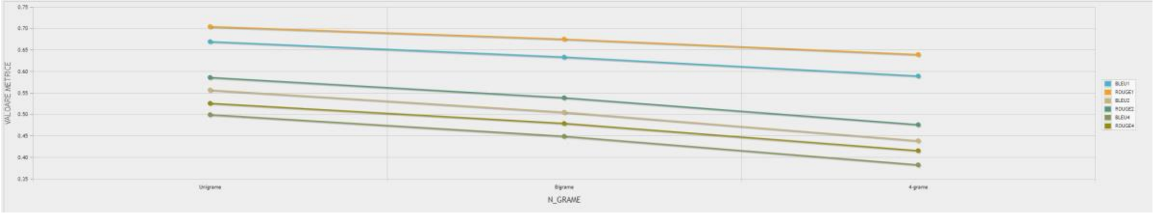
După cum am precizat mai sus, pentru a decide primele câte propoziții (cu cel mai mare scor) trebuie să apară în rezumat, am numărat câte propoziții există în toate rezumatele din setul de antrenare și am împărțit la numărul total de propoziții din articole (pentru a obține o constantă între 0 și 1, care îmi indică câte dintre propozițiile dintr-un articol sunt în rezumat). Am înmulțit apoi această constantă cu numărul de propoziții din articolul pe care trebuia să îl sumarizez.

Pentru fiecare categorie de știri, am calculat metricile *BLEU* si *ROUGE* pentru unigrame, bigrame și 4-grame, în fiecare dintre cazurile:

[1] fără nicio modificare;
[2] cu scorul calculat doar pentru substantive;
[3] cu scorul calculat pentru substantive + pondere de similaritate cu titlu;
[4] cu scorul calculat pentru substantive + pondere de similaritate cu titlu + ponderea locației în document.

Influența n-gramelor în metoda TF-IDF

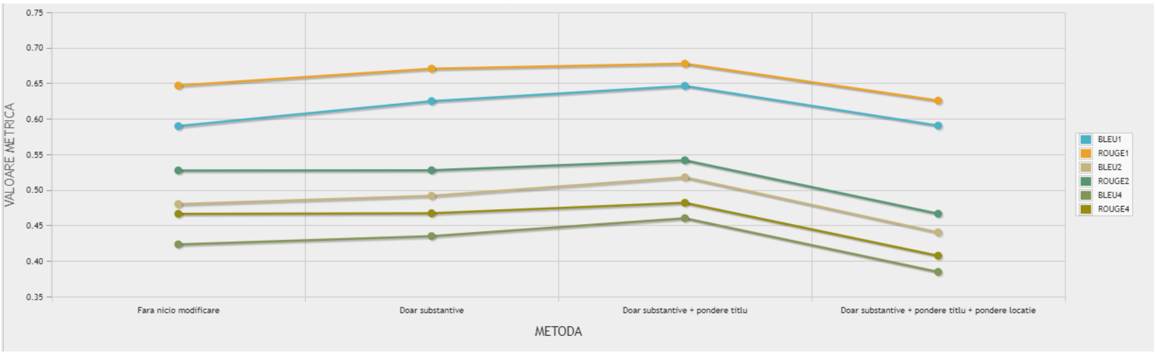
La fel ca pentru metoda Naive Bayes, am generat un grafic care îmi indică valorile metricilor *BLEU* și *ROUGE* în funcție de n-grame:



De aceasta dată, am obținut cele mai bune rezultate pentru unigrame, întrucât valorile tuturor metricilor tind să scadă odată cu creșterea lui n.

Comparație între variațiile metodei (TF-IDF)

Am comparat toate cele 4 variații ale metodei TF-IDF în următorul grafic:



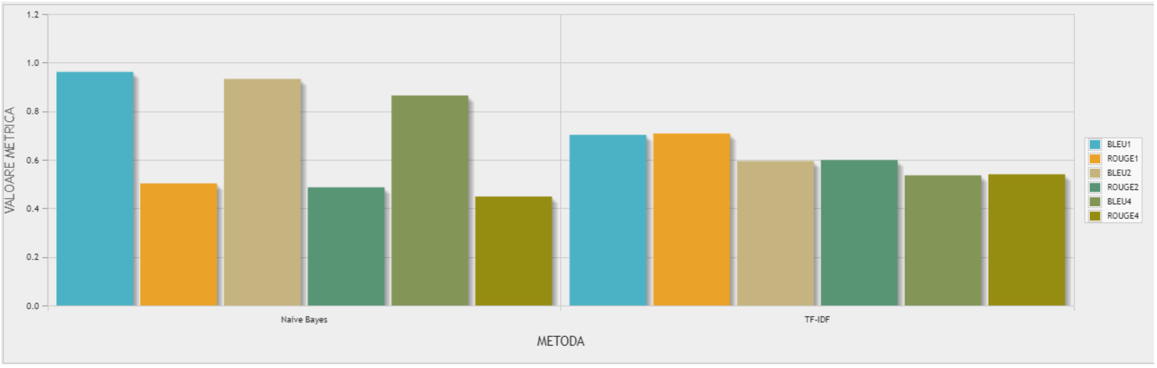
La adăugarea pasului 1 (de calculare a scorului doar pentru substantive), se obține o precizie mai bună, iar recall-ul rămâne aproximativ la fel.

Pentru ponderea de similaritate cu titlu am ales valoarea 0.3, după ce am testat mai multe valori. După adăugarea acestui pas, metrica *BLEU* (precizie) crește semnificativ, și, de asemenea, există o mică creștere și pentru metrica *ROUGE* (recall).

La adăugarea pasului 3, însă, toate metricile scad semnificativ, ajungând la valori chiar mai mici decât în varianta fără nicio modificare.

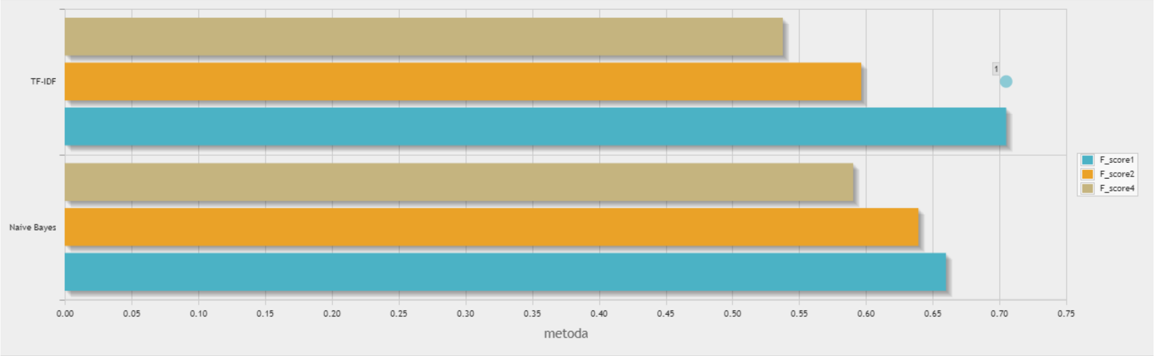
Comparație Naive Bayes și TF-IDF

Pentru a compara cele două metode, am ales să mă folosesc de rezultatele pentru cea mai bună variație. Astfel, pentru Naive Bayes am folosit 4-grame, cu stopwords eliminate și lematizare, iar pentru TF-IDF unigrame, cu scor calculat doar pentru substantive și pondere de similaritate cu titlul.



După cum se poate observa, variația metodei Naive Bayes obține valori foarte mari pentru precizie, însă un recall scăzut, pe când metoda TF-IDF obține valori (bune) foarte apropiate ale preciziei și recall-ului. Astfel, pentru a le compara, am calculat metrica F-score pentru unigrame, bigrame și 4-grame.

$$F - score@N = 2 \cdot \frac{BLEU@N \cdot ROUGE@N}{BLEU@N + ROUGE@N}$$



Calculând F-score pentru unigrame, am obținut o valoare mai mare cu metoda TF-IDF, însă pentru bigrame și 4-grame, valorile au fost mai bune cu metoda Naive Bayes.

Analiza calitativă

Am extras la întâmplare câte un exemplu din fiecare categorie de știri, cu rezumatele generate de cele două metode comparate mai sus, alături de rezumatul referință:

Categoria tech

Referința

One in 10 adult Americans - equivalent to 22 million people - owns an MP3 player, according to a survey. A study by the Pew Internet and American Life Project found that MP3 players are the gadget of choice among affluent young Americans. The survey did not interview teenagers but it is likely that millions of under-18s also have MP3 players. MP3 players are still the gadget of choice for younger adults. Of the 22 million Americans who own MP3 players, 59% are men compared to 41% of women. "IPods and MP3 players are becoming a mainstream technology for consumers" said Lee Rainie, director of the Pew Internet and American Life Project. The American love affair with digital music players has been made possible as more and more homes get broadband.

Generat cu Naive Bayes

One in 10 adult Americans - equivalent to 22 million people - owns an MP3 player, according to a survey. A study by the Pew Internet and American Life Project found that MP3 players are the gadget of choice among affluent young Americans. The survey did not interview teenagers but it is likely that millions of under-18s also have MP3 players. The American love affair with digital music players has been made possible as more and more homes get broadband. Of the 22 million Americans who own MP3 players, 59% are men compared to 41% of women. MP3 players are still the gadget of choice for younger adults. "IPods and MP3 players are becoming a mainstream technology for consumers" said Lee Rainie, director of the Pew Internet and American Life Project.

Generat cu TF-IDF

MP3 players are still the gadget of choice for younger adults. The survey did not interview teenagers but it is likely that millions of under-18s also have MP3 players. The American love affair with digital music players has been made possible as more and more homes get broadband. Of the 22 million Americans who own MP3 players, 59% are men compared to 41% of women. Almost a quarter of those with broadband at home have players, compared to 9% of those who have dial-up access. One in 10 adult Americans - equivalent to 22 million people - owns an MP3 player, according to a survey. A study by the Pew Internet and American Life Project found that MP3 players are the gadget of choice among affluent young Americans. Sixteen percent of parents living with children under 18 have digital players compared to 9% of those who don't.

Categoria sport

Referința

Simon Taylor has been named in the Scotland squad for Saturday's Six Nations clash with Italy. Edinburgh were not in action during the first two weeks of this season's Six Nations. Both Scotland and Italy have lost their opening two Six Nations games and, just like last season's encounter which Italy won, this weekend's game could turn out to be a battle to avoid the wooden spoon. Taylor suffered knee ligament damage playing against Ireland in Dublin in the 2004 Six Nations championship. Taylor has turned out for English side Saracens in recent weeks during a short-term loan to improve his fitness.

Generat cu Naive Bayes

Simon Taylor has been named in the Scotland squad for Saturday's Six Nations clash with Italy. The 25-year-old number eight made a scoring return for Edinburgh at the weekend - his first game in a year for the capital side. Taylor suffered knee ligament damage playing against Ireland in Dublin in the 2004 Six Nations championship. "Simon is one of Scotland's truly world class players so it is a huge bonus," said team-mate Chris Paterson. "He brings a whole new dimension to us, especially in defence and his ability to slow the opposition ball down could be key against Italy if he is involved. " Taylor has turned out for English side Saracens in recent weeks during a short-term loan to improve his fitness. Edinburgh were not in action during the first two weeks of this season's Six Nations. So Taylor played the last 20 minutes of Saracens' win over Northampton and then 40 minutes in a friendly against South African Super 12 side The Cats. Scotland coach Matt Williams is due to name his match-day 22 on Thursday. Both Scotland and Italy have lost their opening two Six Nations games and, just like last season's encounter which Italy won, this weekend's game could turn out to be a battle to avoid the wooden spoon. M Blair (Edinburgh), A Craig (Glasgow), C Cusiter (Borders), S Danielli (Borders), M Di Rollo (Edinburgh), A Henderson (Glasgow), B Hinshelwood (Worcester), R Lamont (Glasgow), S Lamont (Glasgow), D Parks (Glasgow), C Paterson (Edinburgh), G Ross (Leeds), H Southwell (Edinburgh), S Webster (Edinburgh). - R Beattie (Northampton), G Bulloch (Glasgow, capt), B Douglas (Borders), J Dunbar (Leeds), I Fullarton (Saracens), S Grimes (Newcastle), N Hines (Edinburgh), A Hogg (Edinburgh), G Kerr (Leeds), N Lloyd (Saracens), S Murray (Edinburgh), J Petrie (Glasgow), R Russell (London Irish), C Smith (Edinburgh), T Smith (Northampton), S Taylor (Edinburgh), J White (Sale).

Generat cu TF-IDF

The 25-year-old number eight made a scoring return for Edinburgh at the weekend - his first game in a year for the capital side. Taylor has turned out for English side Saracens in recent weeks during a short-term loan to improve his fitness. Edinburgh were not in action during the first two weeks of this season's Six Nations. Simon Taylor has been named in the Scotland squad for Saturday's Six Nations clash with Italy. Taylor suffered knee ligament damage playing against Ireland in Dublin in the 2004 Six Nations championship.

Categoria business

Referința

Continued obstruction of foreign investment could get in the way not only of privatisation plans, but also of Mr Khatami's hope of modestly reducing the government's reliance on oil revenues. In contrast, oil revenues were expected to fall to $14. 1bn from $16bn in the year to March 2005. In an address to the Majlis, Mr Khatami predicted economic growth of 7. 1% in 2005-6, up from 6. 7% in the current year. Mr Khatami's second term as president ends on 1 August, making this his last budget. "Current government expenditure should come from tax revenues," Mr Khatami said. Late last year, they backed a law which would give parliament a veto over foreign investment. Mr Khatami has already been blocked by parliament from reducing the subsidies on many products including bread and petrol, reducing his room to manoeuvre.

Generat cu Naive Bayes

Iran's president, Mohammad Khatami, has unveiled a budget designed to expand public spending by 30% but loosen the Islamic republic's dependence on oil. The budget for the fiscal year starting on 21 March calls for the sell-off of 20% of the state's corporate holdings. Mr Khatami's second term as president ends on 1 August, making this his last budget. But opposition from members of parliament who have attacked previous privatisations could block his plans. Elections in May 2004 ousted many of Mr Khatami's supporters in parliament in favour of more hard-line religious conservatives. Late last year, they backed a law which would give parliament a veto over foreign investment. The ruling was a response to the involvement in telecoms and airport projects by Turkish companies, which hardliners accused of doing business with Israel. It came not long after the Expediency Council - Iran's ultimate decision-maker - blessed Mr Khatami's policy of selling stakes in sectors protected by the constitution such as energy, transport, telecoms and banking. Continued obstruction of foreign investment could get in the way not only of privatisation plans, but also of Mr Khatami's hope of modestly reducing the government's reliance on oil revenues. In an address to the Majlis, Mr Khatami predicted economic growth of 7. He said he wanted to increase the 2005-6 budget to 1,546 trillion rials ($175. 6bn) from the previous year's 1,070 trillion. Within that figure, taxation would rise to $14. 3bn, a rise of over 40% from what is expected from the current year. In contrast, oil revenues were expected to fall to $14. "Current government expenditure should come from tax revenues," Mr Khatami said. "Oil revenues should be used for productive investment. " Mr Khatami has already been blocked by parliament from reducing the subsidies on many products including bread and petrol, reducing his room to manoeuvre.

Generat cu TF-IDF

Iran's president, Mohammad Khatami, has unveiled a budget designed to expand public spending by 30% but loosen the Islamic republic's dependence on oil. The budget for the fiscal year starting on 21 March calls for the sell-off of 20% of the state's corporate holdings. But opposition from members of parliament who have attacked previous privatisations could block his plans. Mr Khatami's second term as president ends on 1 August, making this his last budget. The ruling was a response to the involvement in telecoms and airport projects by Turkish companies, which hardliners accused of doing business with Israel. Mr Khatami has already been blocked by parliament from reducing the subsidies on many products including bread and petrol, reducing his room to manoeuvre. It came not long after the Expediency Council - Iran's ultimate decision-maker - blessed Mr Khatami's policy of selling stakes in sectors protected by the constitution such as energy, transport, telecoms and banking. Late last year, they backed a law which would give parliament a veto over foreign investment.

Categoria politics

Referința

BBC political editor Andrew Marr said that Mr Brown's article was "a warning shot" to Mr Blair not to try and cut him out of the manifesto writing process. Mr Blair argued that under New Labour the country had changed for the better and that was "in part" because of Mr Brown's management of the economy. Mr Blair said a decision had yet to be taken over how the election would be run but the chancellor's role would be "central". The prime minister was asked about Mr Brown's article and about his election role when he appeared on BBC Radio 4's Today programme. The premier insisted Mr Brown will have a key role in Labour's campaign, and praised his handling of the economy. Mr Blair said he was taking "nothing for granted" ahead of the vote - warning that the Tory strategy was to win power via the back door by hinting they were aiming to cut Labour's majority instead of hoping for an outright win.

Generat cu Naive Bayes

The premier insisted Mr Brown will have a key role in Labour's campaign, and praised his handling of the economy. BBC political editor Andrew Marr said that Mr Brown's article was "a warning shot" to Mr Blair not to try and cut him out of the manifesto writing process. The prime minister was asked about Mr Brown's article and about his election role when he appeared on BBC Radio 4's Today programme. Mr Blair said a decision had yet to be taken over how the election would be run but the chancellor's role would be "central". Mr Blair argued that under New Labour the country had changed for the better and that was "in part" because of Mr Brown's management of the economy. Mr Blair said he was taking "nothing for granted" ahead of the vote - warning that the Tory strategy was to win power via the back door by hinting they were aiming to cut Labour's majority instead of hoping for an outright win.

Generat cu TF-IDF

The prime minister was asked about Mr Brown's article and about his election role when he appeared on BBC Radio 4's Today programme. BBC political editor Andrew Marr said that Mr Brown's article was "a warning shot" to Mr Blair not to try and cut him out of the manifesto writing process. The premier insisted Mr Brown will have a key role in Labour's campaign, and praised his handling of the economy. "As our manifesto and our programme for the coming decade should make clear, Labour's ambition is not simply tackling idleness but delivering full employment; not just attacking ignorance, disease and squalor but promoting lifelong education, good health and sustainable communities. " The chancellor has previously planned the party's election strategy but this time the role will be filled by Alan Milburn - a key ally of Tony Blair... but entirely deliberate," was Mr Marr's assessment.

Categoria entertainment

Referința

In the UK box office chart, Meet the Fockers pushed Closer off the top spot while police action movie Assault On Precinct 13, starring rapper Ja Rule, made Â£750,000 in its first weekend. Comedy sequel Meet the Fockers, in which he stars with Ben Stiller, Dustin Hoffman and Barbra Streisand, shot to the top of the UK chart at the weekend. At the same time, US audiences were won over by his new thriller Hide and Seek. London Underground thriller Creep was another new entry at six while quirky comedy Sideways, which got five Oscar nominations last week, entered in eighth place. Robert De Niro has completed a transatlantic box office double by topping the UK and US film charts with two different films at the same time.

Generat cu Naive Bayes

Robert De Niro has completed a transatlantic box office double by topping the UK and US film charts with two different films at the same time. Comedy sequel Meet the Fockers, in which he stars with Ben Stiller, Dustin Hoffman and Barbra Streisand, shot to the top of the UK chart at the weekend. 2m in three days - eight times more than the number two, Closer. At the same time, US audiences were won over by his new thriller Hide and Seek. In Meet the Fockers, he picks up the role of an uptight father and ex-CIA agent from 2000 hit comedy, Meet the Parents. It is a big leap to his role in Hide and Seek, a supernatural horror in which he plays a widower whose daughter's imaginary friend turns nasty. In the UK box office chart, Meet the Fockers pushed Closer off the top spot while police action movie Assault On Precinct 13, starring rapper Ja Rule, made Â£750,000 in its first weekend. London Underground thriller Creep was another new entry at six while quirky comedy Sideways, which got five Oscar nominations last week, entered in eighth place. The Oscar nominations do not seem to have had an impact on fans' choices at cinemas. Leading contenders The Aviator, Million Dollar Baby and Ray all suffered substantial drops in takings compared with the previous weekend.

Generat cu TF-IDF

Robert De Niro has completed a transatlantic box office double by topping the UK and US film charts with two different films at the same time. In the UK box office chart, Meet the Fockers pushed Closer off the top spot while police action movie Assault On Precinct 13, starring rapper Ja Rule, made Â£750,000 in its first weekend. The Oscar nominations do not seem to have had an impact on fans' choices at cinemas. At the same time, US audiences were won over by his new thriller Hide and Seek. Assault on Precinct 13 was in third.

Concluzii analiza calitativă

Se poate observa faptul că rezumatele generate prin metoda Naive Bayes au de cele mai multe ori o lungime mult mai mare decât referința, lucru care explică precizia foarte ridicată și recall-ul scăzut, întrucât există o probabilitate mai mare să le includă pe cele din referință (și, de cele mai multe ori, le include chiar pe toate).

În cazul rezumatelor generate prin TF-IDF, lungimea este potrivită datorită calculului pentru numărul de propoziții realizat în această metodă. Totuși, ordinea propozițiilor în rezumat este dată de scorul acestora și nu întotdeauna această ordine va avea sens pentru cititor. Pentru a rezolva această problemă, ar putea fi selectate propozițiile cu cel mai mare scor, ca până acum, însă păstrând ordinea din articol.

Deși după analiza cantitativă părea că rezumatele obținute prin metoda Naive Bayes vor fi puțin mai bune, acestea includ mult prea multă informație inutilă pe lângă informațiile de interes, așa că voi alege metoda TF-IDF dintre cele două.