Klasifikacija Online Vesti

Dragan Markovic
RN 2/13

Novembar 2016.

Opis Problema

- Dat je skup online vesti, treba odrediti vrstu teme clanka. Moguce klasifikacije su:
  - Poslovna vest (b)
  - Zdravlje i zivot (m)
  - Zabava (e)
  - Tehnologija (t)

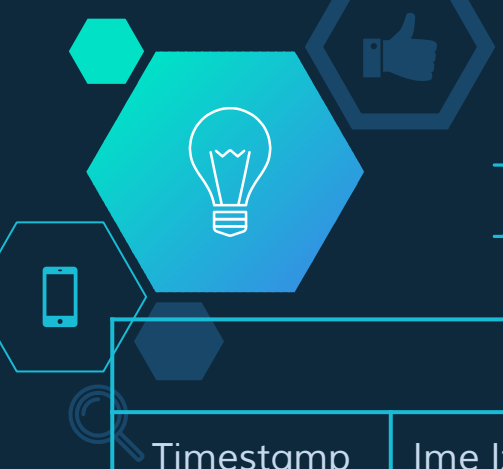
Moguci Attributi

- Podaci se sastoji od 420k redova u csv fajlu, gde svaki red sadrzi:
  - Naslov vesti – “EBay And Icahn Keep Trading"
  - Ime izdavaca – “Capital GR”
  - Url izdavaca – “english.capital.gr”
  - Url clanka – “http://english.capital.gr...”
  - Timestamp – 1394470922077
  - Klasifikacija – b

Algo 1: Naivni Bajes

- P(L | A) = P(L) * P(A | L) / P(A)
- Potencijalni Atributi:
  - Ime izdavaca, url izdavaca, timestamp, url clanka, i
    - Pristup 1: Naslov clanka kao jedan atribut.
    - Pristup 2: Svaka rec u naslovu kao zaseban atribut (bag of words).

L = Labela, A = Atribut



Naivni Bajes: Rezultati

Podskup Atributa				Preciznost (N = 100k)	
Timestamp	Ime Izdavaca	Url Izdavaca	Url Clanka	Ceo Naslov	Bag of Words
1	1	1	1	58%	83%
0	1	1	1	59%	83%
0	1	0	1	59%	84%
0	1	0	0	60%	82%
0	0	0	0	40%	82%

“Najteze” reci i izdavaci

Reci u Naslovu Najvece Tezine (N = 100k)			
Rec	L1	L2	L1 : L2
china	b	e	576.1
stocks	b	e	310.8
miley	e	b	489.3
ebola	m	e	1819.4
microsoft	t	e	1322.4

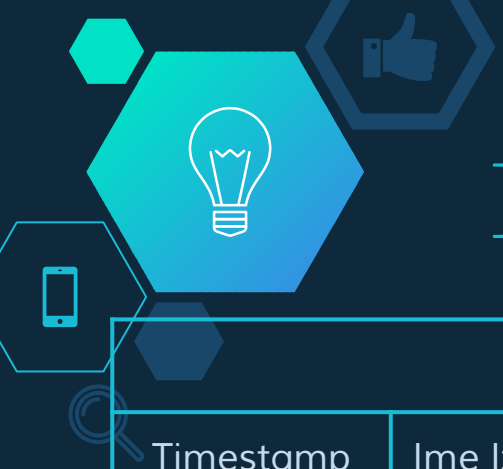
Izdavaci Najvece Tezine (N = 100k)			
Izdavac	L1	L2	L1 : L2
MarketWatch	b	e	94.1
FXStreet.com	b	t	92.7
SheKnows.com	e	t	60.9
Medical Daily	m	b	51.5
Tech Times	t	e	46.9

Algo 2: Max. Entropija

- Zasniva se na principu maksimizacije entropije.
- Za razliku od Bajesa ne pretpostavlja nezavisnost podataka.
- Sporiji od Bajesa. O(N^2) za razliku od O(N).

$$P(L) = \text{Argmax}\ (-\sum p^{*}(A)p(L|A)\log p(L|A))$$

L = Labela, A = Atribut, p*(A) – broj pojavljivanja A-ova.



MaxEnt: Rezultati

Podskup Atributa				Preciznost (N = 100k)	
Timestamp	Ime Izdavaca	Url Izdavaca	Url Clanka	Ceo Naslov	Bag of Words
1	1	1	1	61%	80%
0	1	1	1	62%	79%
0	1	0	1	62%	79%
0	1	0	0	61%	78%
0	0	0	0	26%	77%

Algo 3: SVM

- SVM (Support Vector Machine) su modeli klasifikacije, koji mogu da nauce ne-linearne zavisnosti.
- Funkcije Aktivacije:
  - Linearna: $y = kx$
  - Polinomna: $y = a_3x^3+a_2x^2+a_1x+a_0$
  - Logisticka: $y = 1/(1-e^x)$
- Bag of words pristup izbora atributa.

SVM vs MaxEnt vs NB

Rezultati (SVM vs MaxEnt vs NB)			
Model	Funkcija Aktivacije	N	Preciznost
SVM	Polinomna	100	87%
SVM	Linearna	100	86%
SVM	Logisticka	100	88%
Naive Bayes	-	100,000	84%
MaxEnt	-	100,000	80%

Pitanja?

Fajlovi

- source.py – sors kod
  - NLTK (nltk.org) i scikit learn (scikit-learn.org) paketi su korisceni.
- analiza.xls – excel fajl sa korisnim podacima (rezultati, itd.).
  - Sadrzi stvari kojih nema u prezentaciji.
- newsCorpora.csv – podaci za treniranje.