PRESENTACIÓ CAS KAGGLE

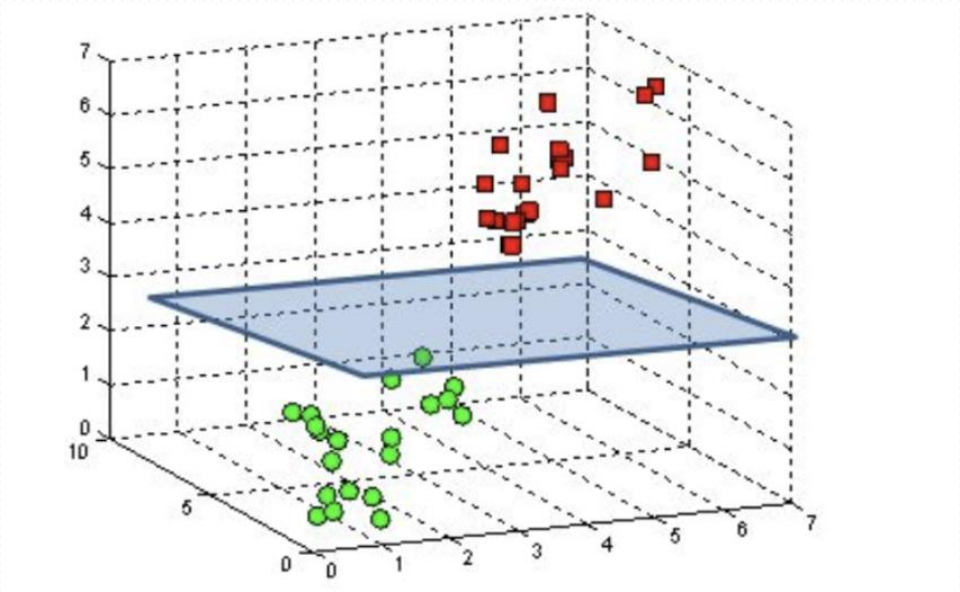
Autor: Òscar Pocurull Rodríguez. NIU: 1496348

Dataset: Classification of websites

URL: http://www.kaggle.com/hetulmehta/website-classification

INDEX:

- 1. INTRODUCCIÓ
- 2. PREPROCESSING
- 3. SELECCIÓ DE MODELS
- 4. RESULTATS
- 5. CONCLUSIÓ



1. INTRODUCCIÓ

- En aquesta practica analitzarem un dataset que classifica les webs segons la categoria a la qual pertanyen.

	Unnamed: 0	website_url	cleaned_website_text	Category
0	0	https://www.booking.com/index.html?aid=1743217	official site good hotel accommodation big sav...	Travel
1	1	https://travelsites.com/expedia/	expedia hotel book sites like use vacation wor...	Travel
2	2	https://travelsites.com/tripadvisor/	tripadvisor hotel book sites like previously d...	Travel
3	3	https://www.momondo.in/?ispredir=true	cheap flights search compare flights momondo f...	Travel
4	4	https://www.ebookers.com/?AFFCID=EBOOKERS-UK.n...	bot create free account create free account si...	Travel
...	...	...	...	...
1403	1403	http://www.oldwomen.org/	old nude women porn mature granny sex horny ol...	Adult
1404	1404	http://www.webcamslave.com	bdsm cams bdsm chat bondage cams free bdsm vid...	Adult
1405	1405	http://www.buyeuroporn.com/	porno dvd online european porn dvd cheap adult...	Adult
1406	1406	http://www.analdreamhouse.com/30/03/agecheck/i...	anal dream house anal dream house anal dream h...	Adult
1407	1407	http://www.world-sex-news.com/	world sex news daily sex news adult news eroti...	Adult

1. INTRODUCCIÓ

- L'objectiu es aprendre quina es la categoria ('Category') a la qual pertanyen les diferents pàgines web del nostre data set.

2. PREPROCESSING

- - Eliminem atribut ‘Unnamed: 0’
- + Afegim atribut ‘Category_ID’
- Eliminem les webs repetides amb drop_duplicates(‘cleaned_website_text’) on passem de 1408 a 1384 webs.
- No te valors nulls

2. PREPROCESSING

- Tf-idf: Convertim el text en dades numèriques.
- Paràmetres tfidfVectorizer-> sublinear_tf=True, min_df: 6, ngram_range(1,2) (unigrams i bigrams), stop_words='english'.
- l'utilitzarem per trobar les paraules (unigrams i bigrams) més correlacionades de cada categoria amb la funció chi2() de sklearn.feature_selection.

3. SELECCIÓ DE MODELS

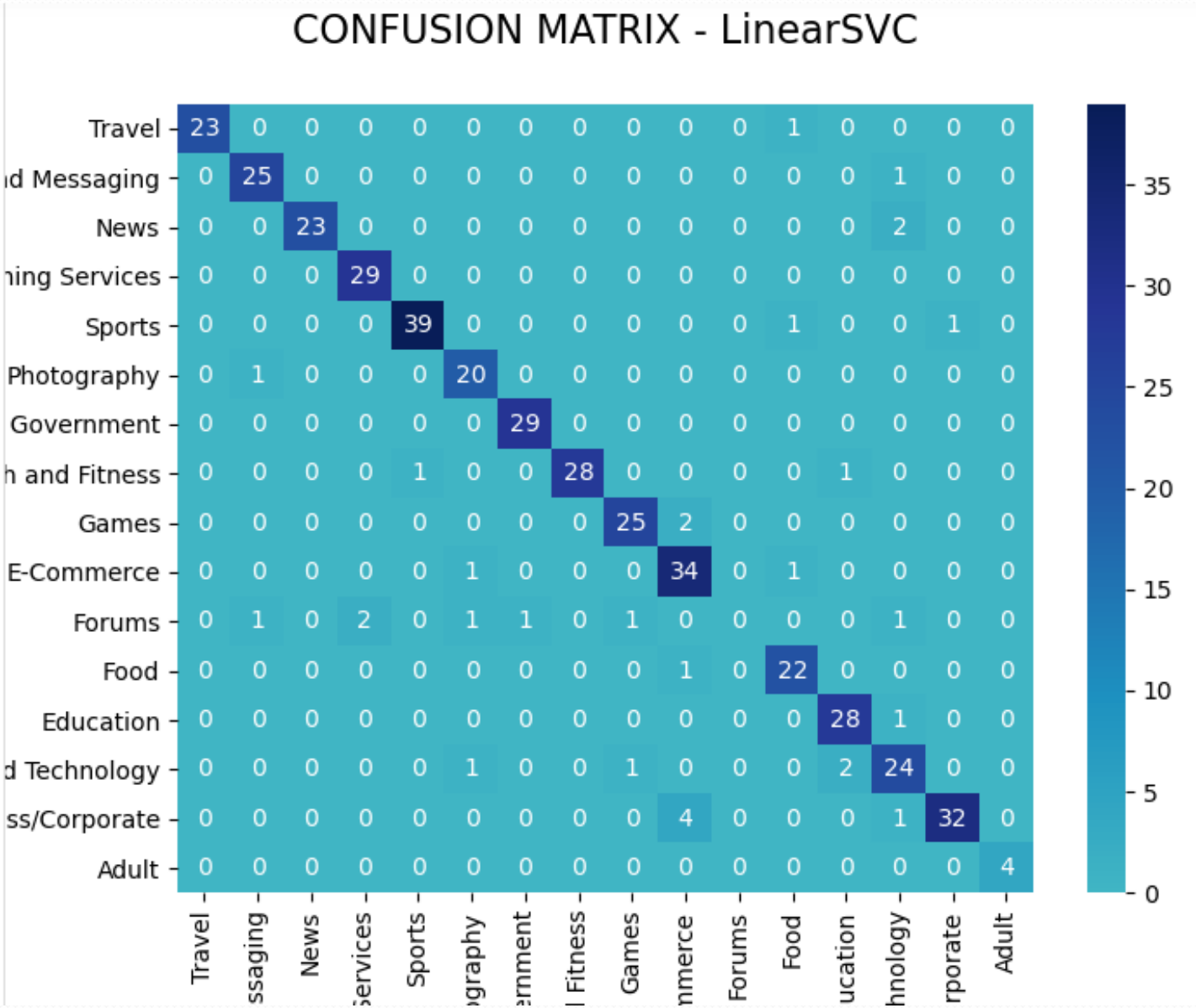
- Models: 70% train, 30% test i cross-validation k-fold, k=5.

Model	Hiperparametres	Mètrica	Temps
Random Forest	100 Trees, max_dept: 5	68%	5227 ms
Random Forest	100 Trees, max_dept: inf	83%	17310 ms
LinearSVC	penalty: l2, multi_class: ovr, max_iter: 1000	91%	2304 ms
MultinomialNB	alpha: 1.0	85%	812 ms
GaussianNB	var_smoothing: 1e-09	73%	6173 ms
Logistic Regression	penalty: l2, C: 1, max_iter: 100	89%	17900 ms

4. RESULTATS

- LinearSVC: model amb millor accuracy.
- LinearSVC calibrat = 94% accuracy.
- LinearSVC usant Pipeline (no cv) = 89% acc. (89 < 91).

- SVC amb kernel: rgb = 87%. Time = 257 seg.
- SVC amb kernel: linear = 90% acc. Time = 242 seg.
- SVC amb kernel: linear, C:100 = 83% acc.
- SVC amb kernel: poly = 43% acc. Time = 252 seg.


CONFUSION MATRIX - LinearSVC
Travel 23 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
d Messaging 0 25 0 0 0 0 0 0 0 0 0 0 0 1 0 0
News 0 0 23 0 0 0 0 0 0 0 0 0 0 2 0 0
ning Services 0 0 0 29 0 0 0 0 0 0 0 0 0 0 0 0
Sports 0 0 0 0 39 0 0 0 0 0 0 1 0 0 1 0
Photography 0 1 0 0 0 20 0 0 0 0 0 0 0 0 0 0
Government 0 0 0 0 0 0 29 0 0 0 0 0 0 0 0 0
h and Fitness 0 0 0 0 1 0 0 28 0 0 0 0 1 0 0 0
Games 0 0 0 0 0 0 0 0 25 2 0 0 0 0 0 0
E-Commerce 0 0 0 0 0 1 0 0 0 34 0 1 0 0 0 0
Forums 0 1 0 2 0 1 1 0 1 0 0 0 0 1 0 0
Food 0 0 0 0 0 0 0 0 0 1 0 22 0 0 0 0
Education 0 0 0 0 0 0 0 0 0 0 0 0 28 1 0 0
d Technology 0 0 0 0 0 1 0 0 1 0 0 0 2 24 0 0
ss/Corporate 0 0 0 0 0 0 0 0 0 4 0 0 0 1 32 0
Adult 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4
Travel
ssaging
News
ervices
Sports
ography
ernment
l Fitness
Games
mmerce
Forums
Food
ucation
hnology
rporate
Adult
35
30
25
20
15
10
5
0

4. RESULTATS

- Aplicant només unigrames al tf-idf:

Model	Hiperparametres	Mètrica	Temps
Random Forest	100 Trees, max_dept: inf	84%	13776 ms
LinearSVC	penalty: l2, multi_class: ovr, max_iter: 1000	91%	1511 ms
MultinomialNB	alpha: 1.0	85%	420 ms
GaussianNB	var_smoothing: 1e-09	64%	3365 ms
Logistic Regression	penalty: l2, C: 1, max_iter: 100	89%	13563 ms

- Aplicant només bigrames al tf-idf:

Model	Hiperparametres	Mètrica	Temps
Random Forest	100 Trees, max_dept: inf	65%	17830 ms
LinearSVC	penalty: l2, multi_class: ovr, max_iter: 1000	76%	555 ms
MultinomialNB	alpha: 1.0	71%	195 ms
GaussianNB	var_smoothing: 1e-09	74%	1700 ms
Logistic Regression	penalty: l2, C: 1, max_iter: 100	73%	4926 ms

5. CONCLUSIONS

- El millor model es el LinearSVC: max. Accuracy (91%) i ràpid al executar-lo (< 2 seg).
- SVC i NuSVC (implementats amb **libsvm**) tenen pitjor rendiment que el LinearSVC (amb **liblinear**) per a un gran numero de mostres (1384).
- Seria interesant optimitzar més les dades de text per trobar el balanç de millor temps i rendiment possible.