Text Mining and Search Project

Amazon Reviews Dataset
Text Classification

Marco Ferrario 795203
Giorgio Ottolina 838017

CONTENTS

DATASET

AMAZON REVIEWS DATASET

Topics: Electronics, Kindle, CDs and Movies/TV
Format: JSON
Size: > 5 000 000 reviews

{
"**asin**": "B00000JBLH",
"**helpful**": [3, 4],
"**overall**": 5.0,
"**reviewText**": "I bought my first HP12C in about 1984 or so,
and it served me faithfully until 2002 when I lost it while travelling.
I searched for another one to replace it, but found one difficult
to come by in my area.
…
I didn't even have to replace the batteries in well over a decade of use!
HP 12C, I'm coming home!",
"**reviewTime**": "09 3, 2004",
"**reviewerID**": "A32T2H8150OJLU",
"**reviewerName**": "ARH",
"**summary**": "A solid performer, and long time friend",
"**unixReviewTime**": 1094169600
}

Classe

Testo

SUBSET CONSIDERATO

1 ***Recensioni valutate più di 5 volte***

Righe: da 5466932 a 874364

"helpful": [3 , 4]

> 5

2 ***Recensioni con almeno 5 parole***

Righe: da 874354 a 873718

```
cond = df['word_count'] > 4
df = df[cond]
```

3 ***Recensioni solo inglesi***

Righe: da 873718 a 872307

Rilevate con “langdetect”

PREPROCESSING

- **DATASET**
- **PREPROCESSING**
- **TEXT REPRESENTATION**
- **CLASSIFICATION**
- **EVALUATION AND CONCLUSION**

PREPROCESSING

STEP 1 **TOKENIZATION**

STEP 2 **NORMALIZATION**

STEP 3 **REMOVE NUMBERS AND PUNCTUATION**

STEP 4 **STOP WORDS REMOVAL**

STEP 5 **LEMMATIZATION**

STEP 6 **STEMMING**

PREPROCESSING

TOKENIZATION

So, I decided to buy up and purchased an HP 49G. What a mistake! I know that many people view the HP 49G (now 49G+)

[So,, i, decided, to, buy, up, and, purchased, an, hp, 49g., what, a, mistake!, i, know, that, many, people, view, the, hp, 49g, (, now, 49g+,),]

PREPROCESSING

NORMALIZATION

So, I decided to buy up and purchased an HP 49G.
What a mistake! I know that many people view the HP 49G (now 49G+)

so, i decided to buy up and purchased an hp 49g.
what a mistake! i know that many people view the hp 49g (now 49g+)

PREPROCESSING

REMOVE NUMBERS AND PUNCTUACTION

so i decided to buy up and purchased an hp 49g what a mistake! i know that many people view the hp 49g (now 49g+)

so i decided to buy up and purchased an hp g what a mistake i know that many people view the hp g now g

PREPROCESSING

STOP WORDS REMOVAL

~~so~~ ~~i~~ decided ~~to~~ buy ~~up~~ ~~and~~ purchased ~~an~~ hp g ~~what~~ ~~a~~ mistake ~~i~~ know ~~that~~ many people view ~~the~~ hp g ~~now~~ g

decided buy purchased hp g mistake know many people view hp g g

PREPROCESSING

LEMMATIZATION/STEMMING

decided buy purchased hp g mistake know many people view hp g g

decid buy purchas hp g mistak know mani peopl view hp g g

PREPROCESSING

- **DATASET**
- **PREPROCESSING**
- **TEXT REPRESENTATION**
- **CLASSIFICATION MODELS**
- **EVALUATION AND CONCLUSION**

TEXT REPRESENTATION

COUNT VECTORIZER

- Considerati Uni-gram e Bi-gram → «decided» «decided buy» «purchased»
- Feature selection → *select Kbest:* 10000 feature con *chi2*
- 5 Classification models

Terms

Documents →

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0	0	1	0	1

TEXT REPRESENTATION

TF-IDF

- Considerati Uni-gram e Bi-gram → «decided» «decided buy» «purchased»
- Feature selection → *select Kbest:* 10000 feature con *chi2*
- 5 Classification models

Terms — Tf-Idf value

Documents →

```
(0,      10255)  0.0727336331113147
(0,      44451)  0.35101474415480843
(0,      88374)  0.13282060609760257
(0,      32885)  0.16335768847169346
(0,      54975)  0.09907168946999645
(0,      104657) 0.16505658094946435
                   ...
(0,      87359)  0.13961763759124002
(0,      82798)  0.08586926237109241
(0,      22729)  0.09028466248740455
(0,      12007)  0.06171032370388875
(0,      79021)  0.08285922464177037
(0,      44437)  0.2327801290815596
```

CLASSIFICATION MODELS

- **DATASET**
- **PREPROCESSING**
- **TEXT REPRESENTATION**
- **CLASSIFICATION MODELS**
- **EVALUATION AND CONCLUSION**

CLASSIFICATION MODELS

Model		CountVec	Tf-Idf
LinearSVC	Time:	24min 57s	8min 33s
	Accuracy:	0.6056 %	0.6158 %
Log. Regression	Time:	30min 13s	1min 36s
	Accuracy:	0.6090 %	0.6094 %
Random Forest	Time:	39.5 s	48.7 s
	Accuracy:	0.4489 % (*)	0.4489 % (*)
Multinomial NB	Time:	745 ms	622 ms
	Accuracy:	0.4996 %	0.5131 %
Gradient Boost	Time:	2min	2min 33s
	Accuracy:	0.4957 %	0.4992 %

(*) tutte le recensioni classificate come 5 stelle: F-Score nulla

EVALUATION AND CONCLUSION

- **DATASET**
- **PREPROCESSING**
- **TEXT REPRESENTATION**
- **CLASSIFICATION MODELS**
- **EVALUATION AND CONCLUSION**

EVALUATION AND CONCLUSION

1

RISULTATI OTTENUTI

Efficiency → TF-IDF

Effectiveness → LinearSVC
Logistic Regression

Accuracy: 60 %

2

POSSIBILI MIGLIORAMENTI

Altri metodi di Feature Extraction - Selection/Synthesis – Weighting (es. Word2vec, mutual dependence, Matrix decomposition…)