



Progettazione e sviluppo di un classificatore per l'individuazione di commenti non informativi

Relatori:

Prof.ssa Nicole Novielli

Prof. Pierpaolo Basile

Laureando:

Giuseppe Colavito



Problema



```
//loop from 1 to N
```



```
// Synchronize changes of the underlying date value with the temporalAccessorValue
```



```
//loop from 1 to N
```

```
for (int i = 1; i < N; i++) {  
    System.out.println(i);  
}
```



```
public TemporalAccessorPicker() {  
    setConverter(new InternalConverter());  
  
    // Synchronize changes of the underlying date value with the temporalAccessorValue  
    BindingsHelper.bindBidirectional(valueProperty(), temporalAccessorValue,  
        TemporalAccessorPicker::addCurrentTime,  
        TemporalAccessorPicker::getDate);  
}
```



```
//loop from 1 to N
```

```
for (int i = 1; i < N; i++) {  
    System.out.println(i);  
}
```



```
public TemporalAccessorPicker() {  
    setConverter(new InternalConverter());  
  
    // Synchronize changes of the underlying date value with the temporalAccessorValue  
    BindingsHelper.bindBidirectional(valueProperty(), temporalAccessorValue,  
        TemporalAccessorPicker::addCurrentTime,  
        TemporalAccessorPicker::getDate);  
}
```





Decluttering challenge

- **Obiettivo:** Costruire un classificatore per il riconoscimento di commenti *non informativi*
- **Task binario:** Commenti non informativi etichettati come "*Non-information = Yes*"

The screenshot shows the Kaggle website interface. On the left is a navigation sidebar with links for Home, Compete, Data, Notebooks, Discuss, Courses, and a 'More' dropdown. Below this is a 'Recently Viewed' section listing 'Declutter Challenge 20...', 'StackOverflow w/ BERT', 'Cross-Platform w/ Bert', 'BertNotebook_V2', and 'BertNotebook'. The main content area features a banner for the 'Declutter Challenge 2020' with a background image of code. The banner text reads: 'Build an automated tool that can identify unnecessary software documentation at the class or file level.' Below the banner is a navigation bar with tabs: Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, Host, My Submissions, and Submit Predictions. The 'Overview' tab is selected, showing a page with an 'Edit' button. The page content includes a 'Description' section with the text: '** This is the 2nd version of the challenge featuring an expanded dataset**'. Below this is the 'Decluttering Challenge (DeClutter)' title, followed by a paragraph: 'In the scope of DocGen2, the Second Software Documentation Generation Challenge, hosted by ICSME 2020.' The 'Task Description' section begins with: 'The goal of the DeClutter challenge is to build an automated tool that can identify unnecessary software documentation at the class or file level. For example, a comment saying //loop from 1 to N just before code implementing'.

Dataset



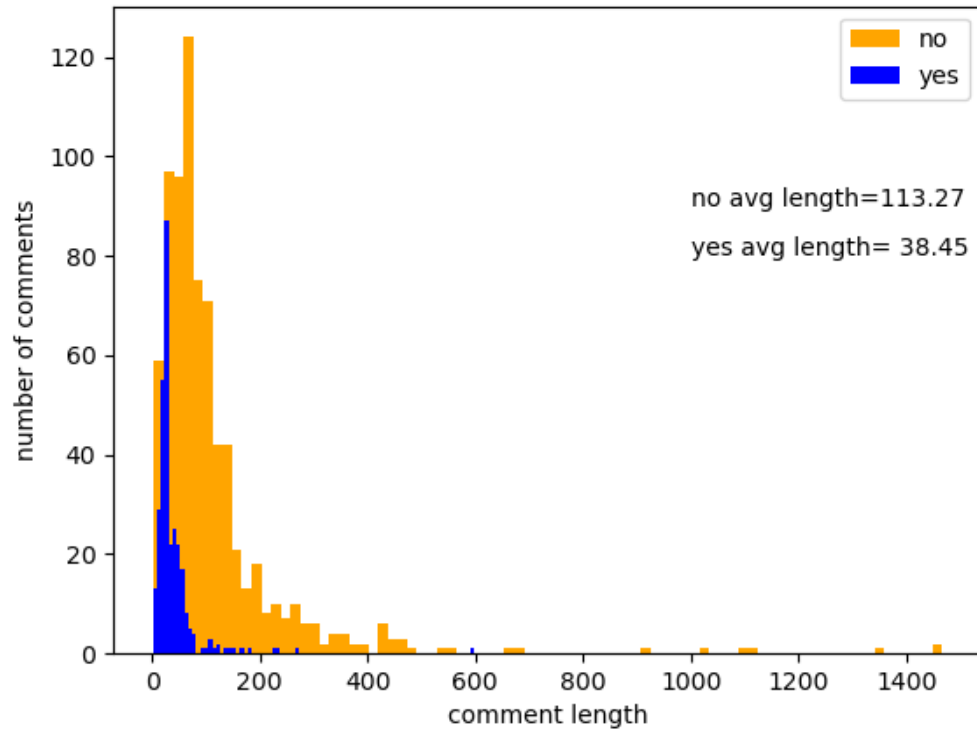
- Repository *Jabref*



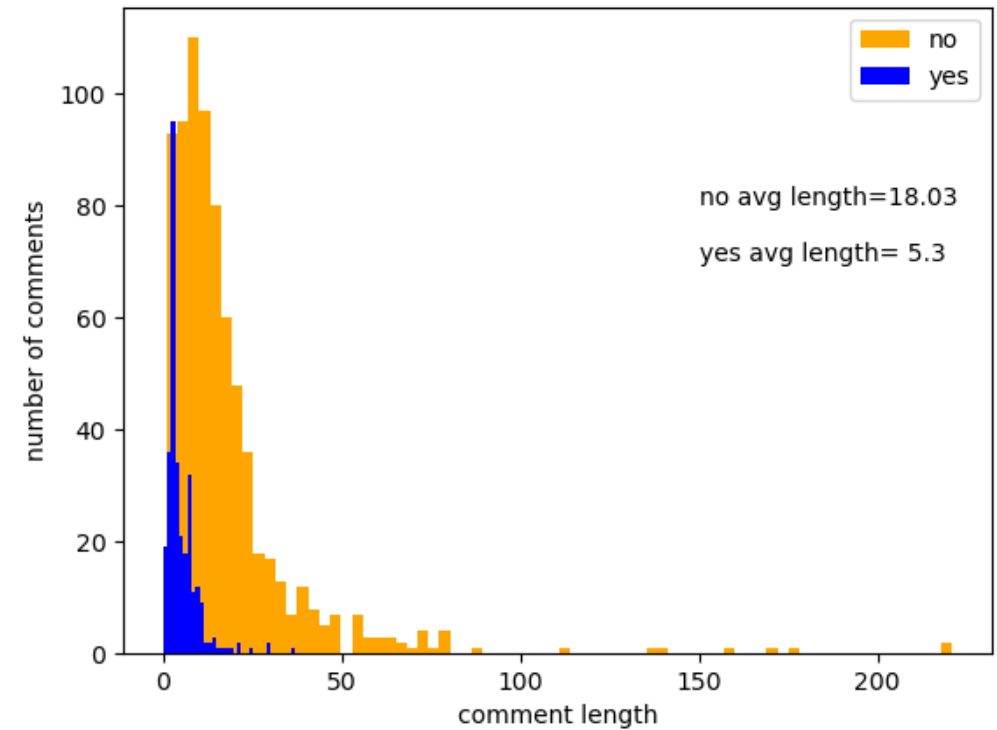
- Linguaggio Java

	Non-information		Totale
	Yes	No	
Training set	304 (29%)	743 (71%)	1047
Test set	261

Lunghezza del commento



Analisi lunghezza (numero lettere) commenti



Analisi lunghezza (numero di parole) commenti

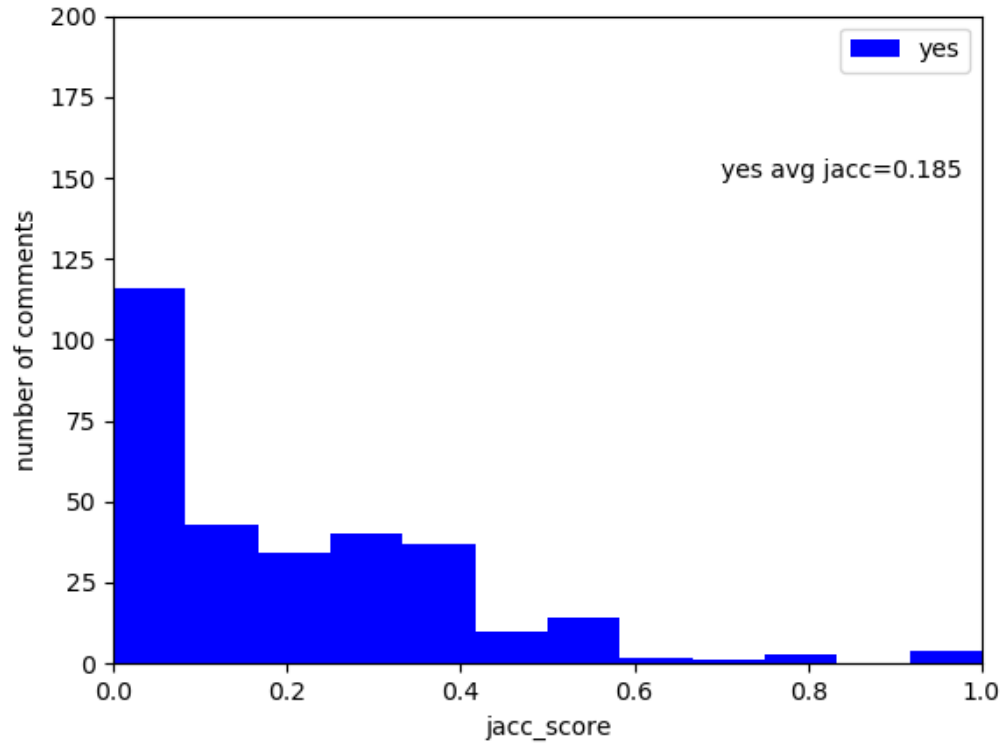
Indice di Jaccard



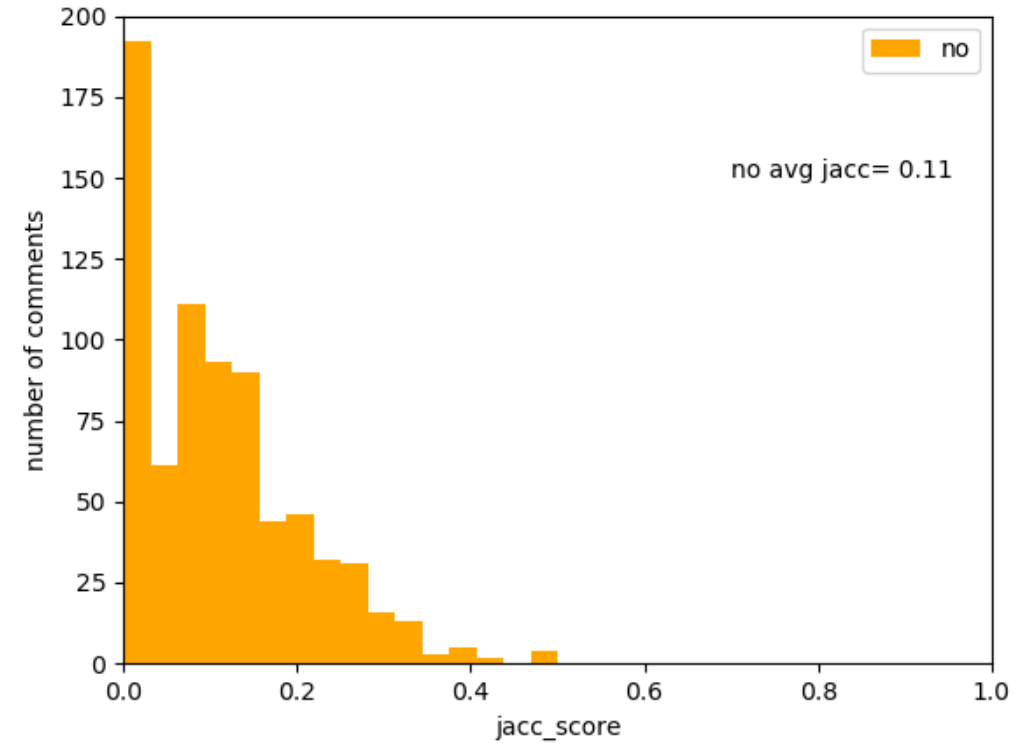
```
// update all tab titles  
updateAllTabTitles();
```

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Indice di Jaccard



Analisi indice di Jaccard su commenti “Non-information=Yes”



Analisi indice di Jaccard su commenti “Non-information=No”

Posizione



1 `public String getSchema() throws SQLException {`

`// Auto-generated method stub`

`return null;`

`}`

2 `@Override`

`public void abort(Executor executor) throws SQLException {`

`// Auto-generated method stub`

`}`

Posizione



3

```
return Optional.of(suffix); // return the first one we found, anyway.
```

```
}
```

4

```
} else { //Dir must be a folder, not a file
```

```
if (!Files.isDirectory(directory)) {
```

```
directory = directory.getParent();
```

```
}
```

Posizione



Position	Non-Information = Yes	Non-Information = No
method declaration	83 (28%)	200 (28%)
if	25 (8%)	56 (8%)
method call	64 (22%)	146 (20%)
assignment	55 (19%)	155 (22%)
class declaration	9 (3%)	111 (16%)
return	43 (14%)	19 (2%)
cycle	12 (4%)	16 (2%)

Tags



Tag	Non-Information = Yes	Non-Information = No
param	5 (36%)	6 (6%)
return	4 (29%)	7 (7%)
see	0	2 (2%)
link	3 (21%)	74 (74%)
Override	0	1 (1%)
implNote	1 (7%)	1 (1%)
code	0	7 (7%)
inheritDoc	1 (7%)	0

Tipo



Java offre tre modi alternativi per commentare il codice sorgente:

- Line

```
boolean maskText = (textInputControl instanceof PasswordField); // (maskText("A") != "A");
```

- Block

```
MOVE_TAB_ARROW(MaterialDesignIcon.ARROW_UP_BOLD), /*css: arrow-up-bold */
```

- Javadoc

```
/**
 * This class contains some code taken from {@link
 * com.sun.javafx.scene.control.behavior.TextInputControlBehavior},
 * which is not accessible and thus we have no other choice.
 * TODO: remove this ugly workaround as soon as control behavior is made public
 * reported at https://github.com/javafxports/openjdk-jfx/issues/583
 */
```

Tipo



	Non-information = Yes	Non-information = No
Javadoc	66 (22%)	305 (41%)
Line	219 (72%)	424 (57%)
Block	19 (6%)	14 (2%)

Valutazione



Metriche utilizzate:

Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient

Baseline:

- Classificatore che predice sempre la classe di maggioranza
- Classificatore Bag-of-words

Validazione:

- K-Fold cross validation (K=10)



Modelli

Features extratestuali

- Lunghezza
- Indice di Jaccard
- Posizione
- Tags
- Tipo

Features testuali ed extratestuali

- Lunghezza
- Indice di Jaccard
- Posizione
- Tags
- Tipo
- *Tf-idf*
- *Word-count*



Risultati

sole features extratestuali

classifier	accuracy	precision	recall	f1-score	matthews corrcoeff
Baseline Dummy	0.71	0.35	0.5	0.41	0.0
Baseline Tf-idf	0.82	0.79	0.75	0.76	0.53
BernoulliNB	0.7	0.69	0.73	0.68	0.41
LinearSVC	0.81	0.78	0.74	0.76	0.52
SVC (poly degree=2)	0.74	0.75	0.58	0.57	0.27
MLPClassifier	0.83	0.79	0.77	0.78	0.57
RandomForestClassifier	0.82	0.78	0.76	0.77	0.54
AdaBoostClassifier	0.84	0.82	0.78	0.8	0.6
BaggingClassifier	0.82	0.78	0.76	0.77	0.54
ExtraTreesClassifier	0.81	0.77	0.75	0.76	0.52
GradientBoostingClassifier	0.83	0.8	0.77	0.78	0.57
LogisticRegression	0.81	0.78	0.75	0.76	0.53
DecisionTreeClassifier	0.78	0.73	0.73	0.73	0.45
SGDClassifier	0.8	0.77	0.74	0.74	0.5
SoftVoting	0.83	0.8	0.77	0.78	0.57
HardVoting	0.84	0.81	0.78	0.79	0.59

Baseline Miglior performance



Risultati

combinazione di tf-idf e features extratestuali


classifier	accuracy	precision	recall	f1-score	matthews corrcoeff
Baseline Dummy	0.71	0.35	0.5	0.41	0.0
Baseline Tf-idf	0.82	0.79	0.75	0.76	0.53
BernoulliNB	0.73	0.72	0.77	0.72	0.49
LinearSVC	0.83	0.8	0.77	0.79	0.58
SVC (poly degree=2)	0.74	0.71	0.58	0.57	0.25
MLPClassifier	0.8	0.76	0.75	0.75	0.51
RandomForestClassifier	0.84	0.83	0.77	0.79	0.6
AdaBoostClassifier	0.84	0.81	0.8	0.81	0.62
BaggingClassifier	0.84	0.82	0.78	0.79	0.6
ExtraTreesClassifier	0.83	0.81	0.74	0.76	0.55
GradientBoostingClassifier	0.85	0.83	0.79	0.8	0.61
LogisticRegression	0.82	0.79	0.76	0.77	0.54
DecisionTreeClassifier	0.81	0.77	0.78	0.77	0.55
SGDClassifier	0.83	0.8	0.79	0.79	0.59
SoftVoting	0.85	0.83	0.8	0.81	0.62
HardVoting	0.85	0.83	0.79	0.81	0.62

Baseline Miglior performance

Valutazione su test set (Kaggle)











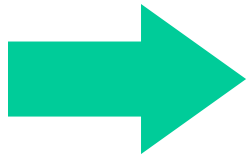
Modello	Voting	F1-Score
Bag-of-Words (word-count) e features extratestuali	hard	0.81
Bag-of-Words (tf-idf) e features extratestuali	hard	0.80
Features extratestuali	hard	0.79
Bag-of-Words (word-count) e features extratestuali	soft	0.77
Bag-of-Words (tf-idf) e features extratestuali	soft	0.77
Features extratestuali	soft	0.76
Bag-of-Words (tf-idf)	hard	0.74
Bag-of-Words (tf-idf)	soft	0.73

 Miglior performance

Classifica pubblica (Kaggle)



#	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	codeup-yang			0.86923	14	2d
2	ZZZ			0.85384	5	3d
3	fire			0.83076	4	3d
4	i'm here			0.81538	30	3d
Your Best Entry ↑ Your submission scored 0.80000, which is not an improvement of your best score. Keep trying!						
5	BTW			0.76153	9	1mo
6	NPW			0.75384	4	1mo
7	[Deleted]			0.73846	5	1mo
8	Boda0124			0.73846	4	1mo



Sviluppi futuri



- Supporto per altri linguaggi di programmazione (attualmente solo Java)
- Supporto per altre lingue (attualmente solo inglese)
- Replicare lo studio su dataset più numerosi e bilanciati
- Integrazione del “declutter” in IDE di sviluppo



GRAZIE PER L'ATTENZIONE