Progetto di Data Mining
21/02/2020
Luca Ragazzi

Text Mining su articoli della BBC

Lo scopo del progetto si articola in 2 task:

- Classificare gli articoli della BBC rispetto al topic di appartenenza (sport, tecnologia, politica, business e intrattenimento) mediante l'utilizzo del software Weka;
- Per ogni topic, capire qual'è l'argomento più discusso (con rilevazione dei termini più correlati ad ogni topic);

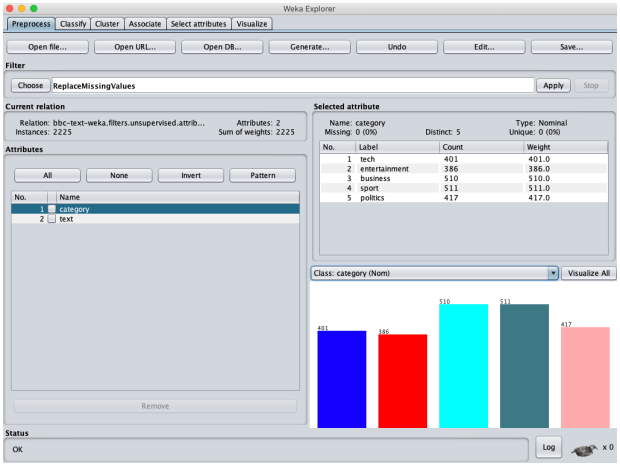
Data set utilizzato

Sono presenti dati su 2225 articoli della BBC relativi a 5 topic diversi, tra cui sport, tecnologia, politica, business e intrattenimento.

Il data set è composto da 2 colonne, una per il topic dell'articolo (*category*) e una per il testo (*text*).

Classificazione degli articoli della BBC con Weka

Caricamento dei dati



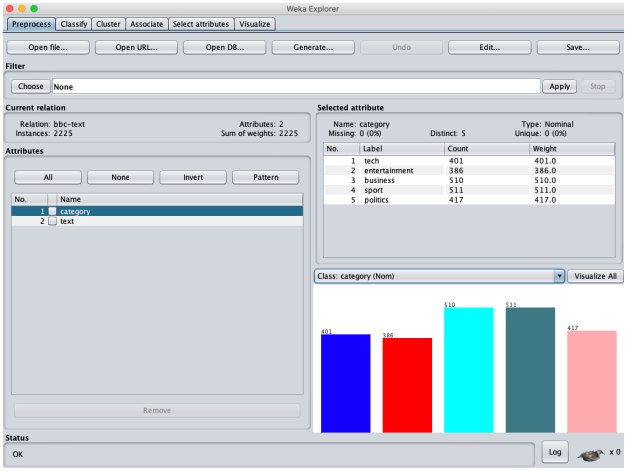
Il primo step è il caricamento dei dati in Weka:

- Aprire l'*Explorer* di Weka;
- Aprire il file *bbc-text.arff*;
- Impostare *category* come attributo classe;

Preprocessing dei dati

Bisogna effettuare il preprocessing dei dati:

- Applicare il filtro *ReplaceMissingValues* (unsupervised, attribute) per gestire attributi con un numero elevato di valori mancanti (sostituiti con la media o la moda dell'attributo corrispondente);

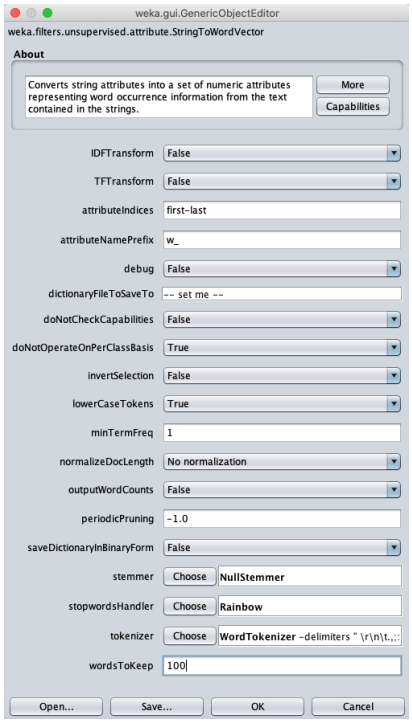


Estrazione delle feature

Per estrarre le feature dal testo, siccome si sta lavorando con un attributo testuale, occorre strutturare l'attributo destrutturato (*text*):

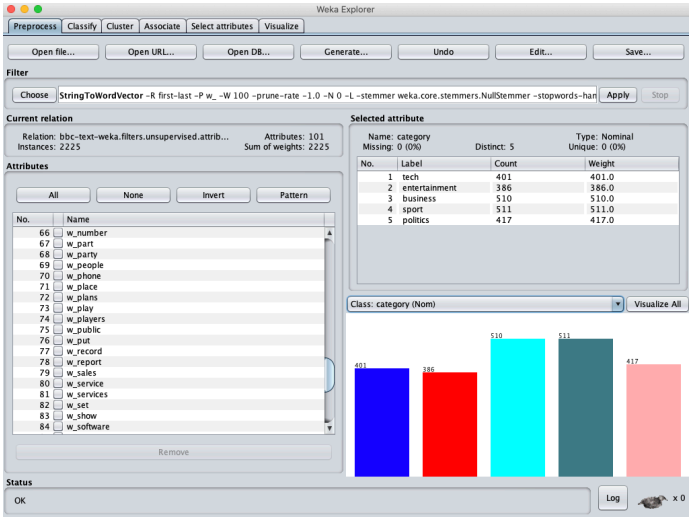
- Applicare il filtro *StringToWordVector* (unsupervised, attribute) per poter effettuare diverse operazioni, come l'estrazione delle singole parole dai testi (tokenization), la rimozione/ trasformazione delle parole per ridurne il numero (stemming), la selezione delle parole più rilevanti (con TF-IDF) e aggiunta di nuovi attributi in sostituzione di quelli originali. Impostare le seguenti opzioni:

  - *attributeNamePrexis* = "w_";



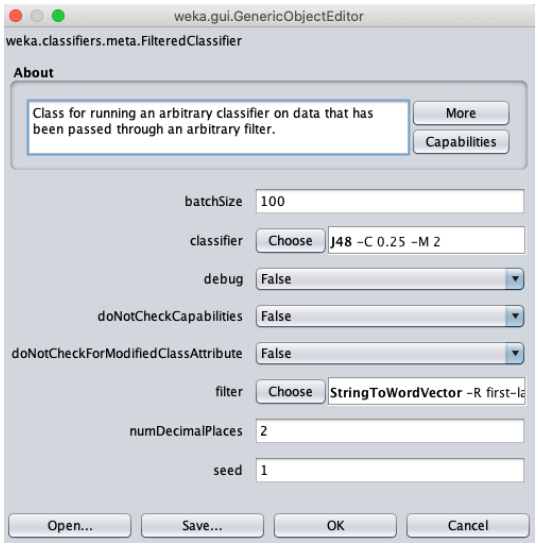
- *doNotOperateOnPerClassBasis* = True;
- *lowerCaseTokens* = True;
- *stopwordsHandler* = Rainbow;
- *wordsToKeep* = 100;

In questo modo viene rimosso l'attributo destrutturato (string) e sono creati 100 attributi "w_parola" che valgono 1 nelle istanze dove parola era presente, 0 altrimenti.



Generazione del modello di classificazione

L'ultimo step è la generazione del modello di classificazione. Per classificare nuovi documenti le stesse trasformazioni con *StringToWordVector* si applicano ad ogni documento da classificare per renderli compatibili con il modello, ma non insieme alla creazione del training set. A tale scopo si utilizza il *FilteredClassifier* (meta-algoritmo):



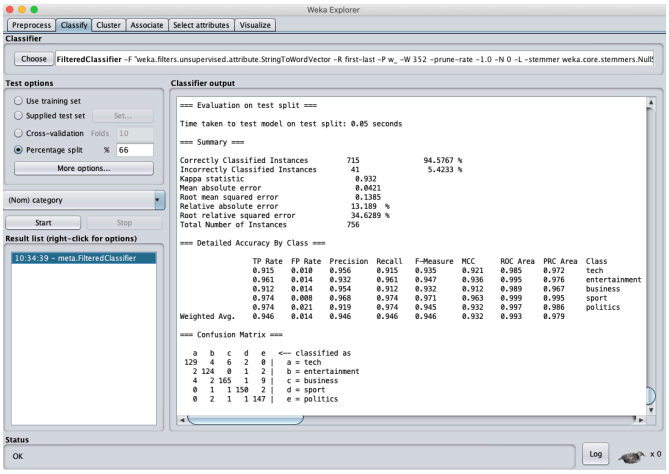
- Con *StringToWordVector* in *FilteredClassifier* prima il training set è usato per estrarre le feature e trasformato e solo successivamente il test set è trasformato con le stesse feature;
- Impostare *Percentage split* a 66% e classe *category*;
- In *FilteredClassifier* utilizzare sia l'algoritmo *J48* (trees) che l'algoritmo *DMNBtext* (bayes);

Impostando *doNotOperateOnPerClassBasis* a False verrano estratte almeno *wordsToKeep* per ciascuna classe.

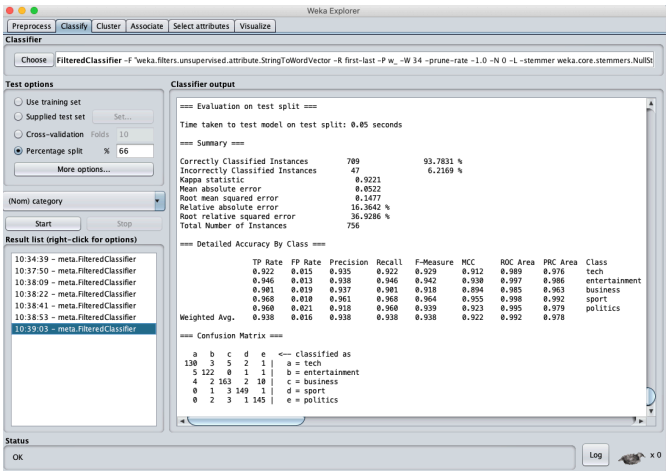
Obiettivo della classificazione è trovare, per ogni algoritmo, il valore minimo di *wordsToKeep* per continuare ad avere sia la Precision che la Recall di ogni classe oltre il 90%, utilizzando sia l'estrazione di feature per ciascuna classe (*doNotOperateOnPerClassBasis* = False) che l'altra tecnica. Inoltre, testare il modello sia con l'utilizzo dello Stemming che senza nel preprocessing dei dati.

Risultati ottenuti per l'algoritmo DMNBtext

Selezionando le feature senza considerare le classi (*doNotOperateOnPerClassBasis* = True) il numero minore di *wordsToKeep* che permette di mantenere una Precision e una Recall superiori al 90%, per ogni classe, è 352 e l'accuratezza della classificazione è 94,5767%.



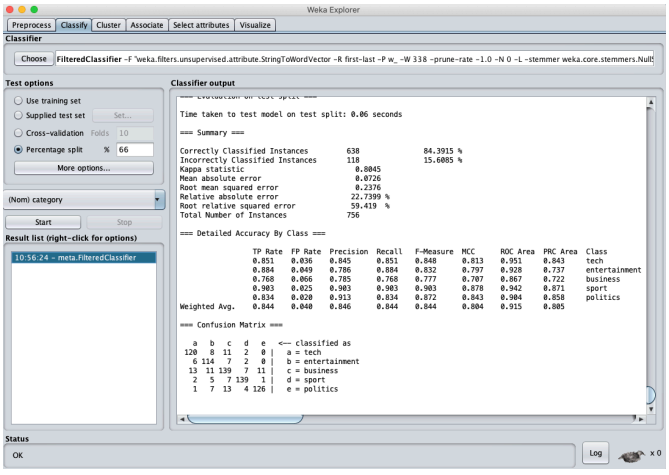
Nell'altra casistica (*doNotOperateOnPerClassBasis* = False) il numero minimo di *wordsToKeep* è 34 e l'accuratezza della classificazione è 93,7831%.



Risultati ottenuti con J48

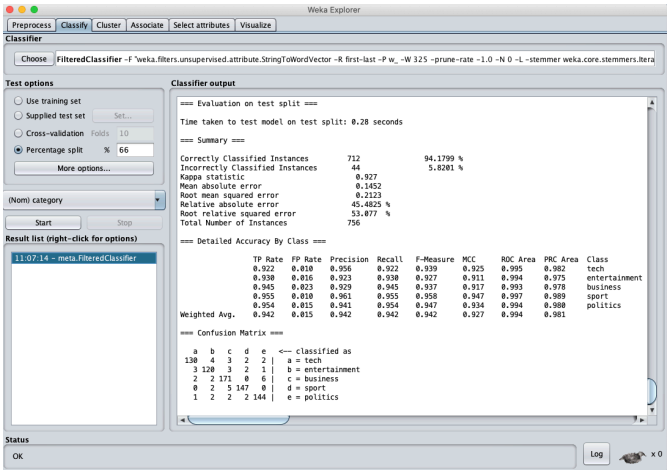
Selezionando le feature senza considerare le classi, il numero di wordsToKeep è 338 e l'accuratezza è 84,3915%, decisamente inferiore rispetto l'algoritmo precedente.

Questo risultato ha portato a non considerare l'estrazione di feature per classe, ma a testare un ulteriore algoritmo di classificazione, RandomForest.

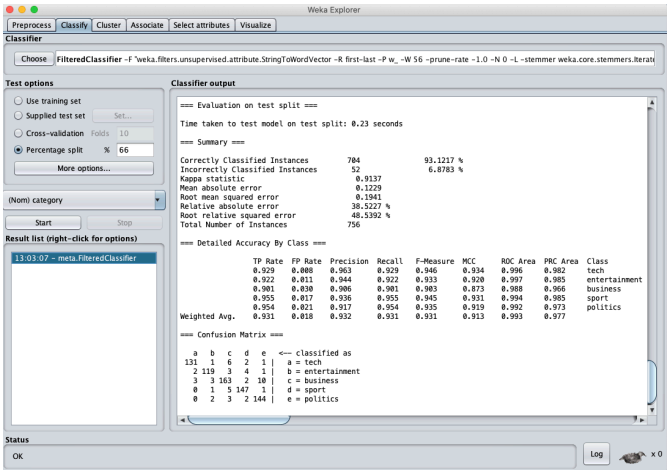


Risultati ottenuti con RandomForest

Per continuare ad avere una Precision e Recall oltre il 90% il numero minimo di *wordsToKeep* è 325 con accuratezza del 94,1799%. Inoltre, è stato effettuato lo Stemming con *IteratedLovinsStemmer*, notando che dava risultati migliori.

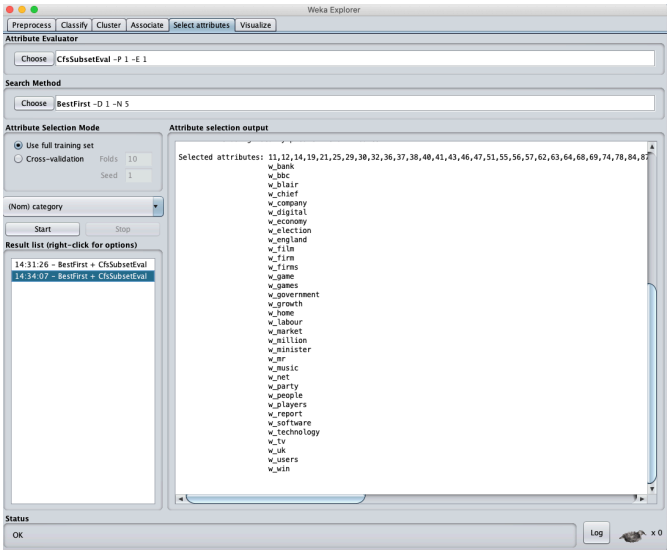


L'ultimo test, con *doNotOperateOnPerClassBasis* a False, ha ottenuto un'accuratezza del 93,1217% con *wordsToKeep* a 56, sempre con l'utilizzo di *IteratedLovinsStemmer* nel preprocessing.



Visualizzazione degli attributi rilevanti per la classificazione

Dalla scheda *Select attributes* si esegue la selezione di un sottoinsieme di attributi più informativi (quelli che più di altri aiutano a comprendere il topic di un articolo). La selezione prevede due operazioni, una per il metodo di valutazione e una per il metodo di ricerca. In questa analisi è stato utilizzato *CfsSubsetEval* come metodo di valutazione e *BestFirst* come metodo di ricerca. Nel Preproccesing dei dati è stato selezionato *doNotOperateOnPerClassBasis* a True e *wordsToKeep* a 100 (per poter visualizzare in un'unica schermata tutti gli attributi selezionati).



Estrazione della conoscenza sugli articoli della BBC con R

Con il software R si vuole individuare quale è l'argomento più discusso all'interno degli articoli delle diverse categorie.

A tale proposito è stata applicata la Latent Semantic Analysis con lo scopo di portare sullo stesso spazio multi-dimensionale sia i termini che i documenti per cercare correlazioni semantiche tra essi. Sono stati utilizzati test statistici (con test chi-quadro) per verificare la correlazione tra i termini e i topic dei documenti.

Applicando il fold-in di nuove query (sempre più specifiche) nello spazio LSA è stato possibile capire quali fossero gli argomenti più discussi all'interno dei diversi topic.

Nel seguito sono riportati risultati ottenuti per le categorie di "sport" e di "politica".

Categoria SPORT

I termini risultati più rilevanti sono "*team cup coach ireland*". Il topic più ricorrente per gli articoli sportivi è correlato a questi termini, ovvero si parla delle interviste fatte a diversi allenatori (*coach*) relative ai match della loro squadra (*team*) in diverse competizione di coppa (*cup*) contro l'Irlanda (*ireland*) o nel futuro scontro contro essa.

Categoria POLITICA

Il topic più ricorrente per gli articoli sulla politica è correlato ai termini "*labour party blair tory tories*", ovvero negli articoli si parla degli scontri tra il partito laburista britannico (*labour party*) con il politico Tony Blair (*blair*) e il partito conservatore Tory (*tory*) con i suoi sostenitori (*tories*).