Relazione

Questo progetto si divide in cinque file (Preprocessing, Processing, prova, Data\_analisys e util )e una directory (Datasets).

# Dataset

Partendo dalla directory troviamo al suo interno quattro file rappresentanti i nostri dataset, tre disponibili dal sito Kaggle da cui abbiamo preso la sfida, uno di nostra creazione per semplificare il processo di predizione. Tutti i file sono informato tsv (tab-separeted-values) e quindi di facile lettura e comprensione:

* **Labled Train Data**: il nome è al quanto esplicativo, infatti questo file è il dataset che viene usato per il train del modello. Si compone di 25000 righe e tre colonne:
  + Id: la colonna che associa ad ogni recensione uno specifico id, irrilevante per quanto riguarda il progetto in se.
  + Review: il cuore del nostro train set, ovvero la X, questa colonna contiene tutte le 25000 recensioni di film prese dal sito IMBD contenenti alcuni pezzi di linguaggio HTML.
  + Sentiment: ciò che rende la classificazione supervised, infatti questa è la nostra Y nel momento dell’allenamento del modello. Il suo valore è binario ed è 1 se la recensione è positiva, 0 altrimenti.
* **Unlabled Train Data**: questo file si compone di 50000 righe e il suo formato è simile a quello descritto in precendeza. L’unica differenza sta nel fatto che non è presente la colonna “Sentiment”, ovvero la Y. Abbiamo comunque usufruito di questo dataset nel momento della trasformazione delle recensioni da stringhe a vettori numerici.
* **Test Data**: anche qui il nome dice tutto. Il file si compone di 25000 righe e due colonne, una contenete le recensioni l’altra l’id.
* **Test Data Labled**: questo è il dataset da noi creato. Il motivo che sta alla base di questo passo è la mancanza della colonna “Sentiment” nel test dataset fornito da Kaggle. Infatti il sito fornisce la possibilità di inviare la propria predizione per ottenere il risultato, ma questa viene limitata per sei volte nell’arco di 24 ore. Cercando su google ho scoperto che nell’id del Test Dataset è presente il voto in decimi correlato alla recensione, e dato che la sfida fornisce il criterio di valutazione ( un voto minore di 5 indica una recensione negativa mentre uno maggiore di 7 una positiva), abbiamo creato una piccola funzione ausiliaria che ci ha fornito la colonna “Sentiment” facilitando il processo di stima dell’errore.

# Preprocessing

Qui è dove i dati vengono trattai per permettere una migliore classificazione.

Le funzioni principali sono tre e vengono chiamate nell’ordine nelle quali adesso le illustreremo.

**Sentence polishing**

Il compito di questa funzione è quello di ripulire le recensioni da tutta una serie di agenti “inquinanti” che possono traviare il classificatore.

La funzione prende come input una lista di stringe, in questo caso tutte le recensioni e vi applica cinque tipi di filtraggio tramite l’uso della list comprehension e di funzioni provenienti dalla libreria [Genism](https://radimrehurek.com/gensim/):

* **Rimozione dei tag di HTML**: come accennato prima le recensioni vengono prese direttamente dal sito HTML e per cio portano con se alcuni pezzi di linguaggio HTML. Qui abbiamo usato la funzione strip\_tags che ha esattamente il compito di eliminare i tag HTML.
* **Rimozione delle stop words**: le stop words sono parole presenti in tutte le lingue la cui ripetizione nelle frasi è elevata ma non sono di vitale importanza per il training di un modello, infatti possono essere anche causa di stime molto basse nella predizione poichè la loro frequenza oscura parole di maggior rilevanza. Alcuni esempi di stop words possono essere trovati [qui](http://www.ranks.nl/stopwords).
* **Rimozione della punteggiatura**: come per le stop words, anche la punteggiatura può influire negativamente sul training del modello, per questo abbiamo deciso di rimuoverla.
* **Rimozione dei caratteri non alfa numerici**: essendo le recensioni scritte da persone possono essere presenti dei caratteri non alfanumerici. La frequenza di questi caratteri è molto bassa, quindi non rappresentano un vero e proprio ostacolo per i modelli ma sono comunque stati tolti per questioni di ottimizzazione di memoria.
* **Rimozione di frasi vuote**: è possibile che dopo i precedenti filtraggi alcuni elementi della lista siano vuoti, cio genererebbe degli errori al tempo di esecuzione e quindi li abbiamo eliminati.

Dopo questa serie di rimozioni la lista ha perso circa un terzo dei suoi elementi, garantendo un miglioramento in termini di memoria utilizzata, tempo di esecuzione e precisione nella predizione.

**String2VecTFIDF**

Una volta puliti i datasets abbiamo bisogno di trasformare le parole in vettori, per questo compito ci siamo avvalsi della funzione [TfidfVecorizer](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html) fornita dalla libreria sklearn.

La teoria alla base di questo algoritmo di vettorializzazione è relativamente semplice.

Quando ho un vasto insieme di stringhe alcune parole, per esempio le stop words, compaiono molto frequentemente. Utilizzando un normale algoritmo basato sulla bag of words, i pesi associati a queste parole maschererebbero la vera importanza di altre parole meno frequenti ma di più significato. Per questo il TF-IDF (term-frequency-inverse-document-frequency) utilizza due termini:

* Il **term frequency**: un intero che rappresenta il numero di volte che uno specifico termine compare in un documento (in questo caso in una recensione)
* **L’inverse document frequency**: semplicemente text{idf}(t) = log{\frac{1 + n_d}{1+\text{df}(d,t)}} + 1 , dove rappresenta il numero delle recensioni e df(d,t) è il numero di recensioni che contengono il termine t.

Questi due termini vengono moltiplicati e il loro prodotto è poi normalizzato secondo la norma euclidea.

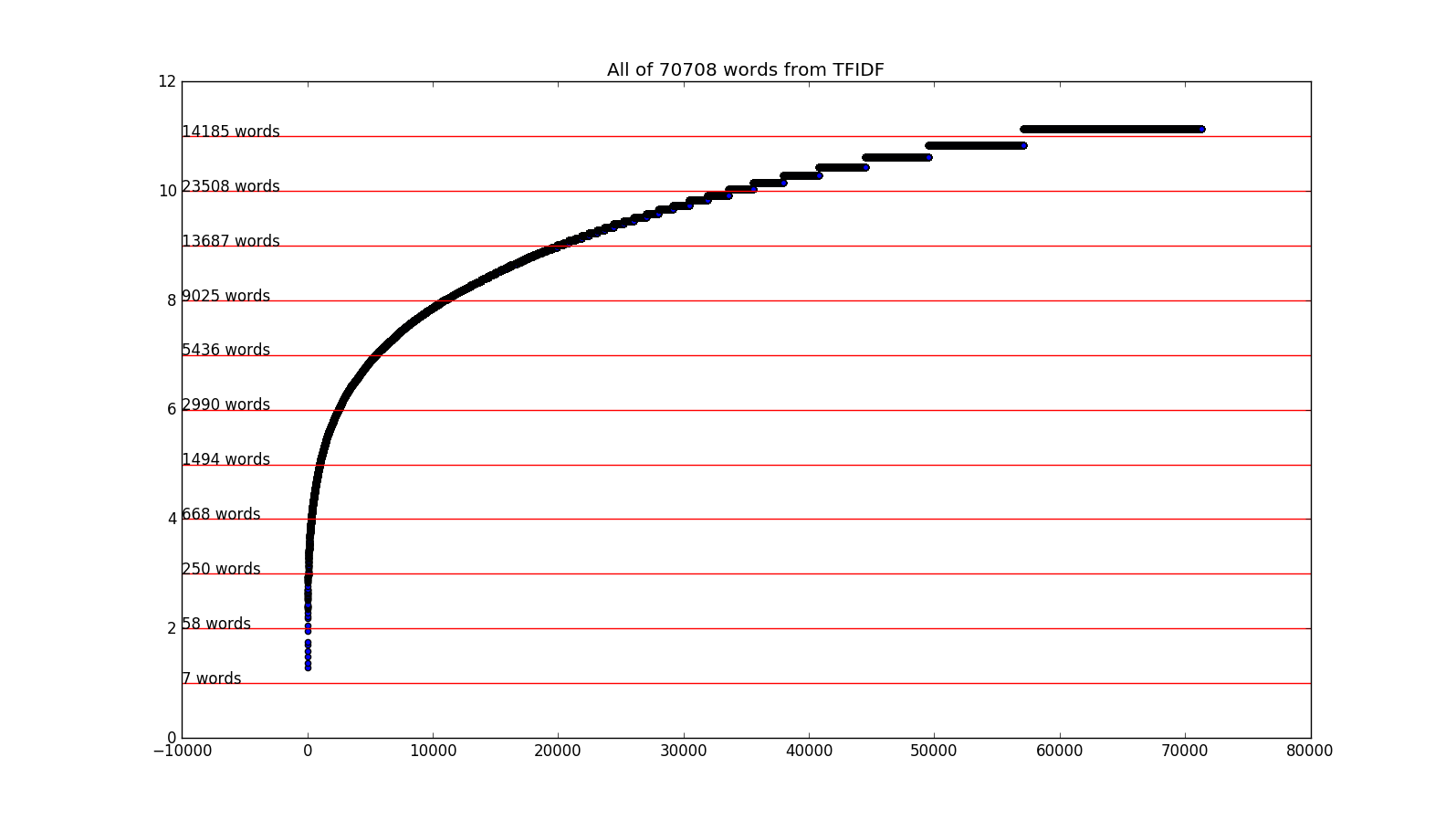
Nel nostro caso sono presenti anche dei termini all’interno della funzione che adesso andremo a spiegare:

* Min\_df: alcune parole hanno una frequenza cosi bassa all’interno della recensione che è possibile ingorarle, per questo abbiamo settato questo parametro a 2, cosi da ignorare tutte le parole che si ripetono meno di tre volte.
* Max\_df: semmai la nostra precedente pulizia delle stop words possa aver tralasciato qualcosa, entra in gioco questo parametro settato a 0.96. In pratica ordino alla funzione di ignorare le parole che si ripetono per più del 96% all’interno della recensione. Un esempio è dato da una recensione molto corta la quale contiene solamente ‘XD’.
* Sublinear\_tf: è una semplice sostituzione del parametro Term frequency, che diventa 1+log(tf).
* Max\_features: essendo il dataset composto da oltre 70000 parole abbiamo deciso di prendere solo le 20000 più rilevanti per evitare di portare il tempo di esecuzione oltre i 10 minuti.
* Strip\_accents: rimpiazza le lettere accentate con normali caratteri ascii, questo è stato possibile poichè la lingua inglese non presenta profonde differenze tra parole accentate e non, a differenza di quella italiana.

Dopo l’inizializzazione si passa al fittaggio del dataset e qui è stato di fondamentale importanza l’uso dell’Unlabled Train Data che ci ha permesso di aggiungere al dizionario creato dalla funzione più del 50% dei termini, aumentando l’accuratezza delle predizioni del 3% in media.

Finito il fittaggio siamo passati all’effettiva trasformazione del dataset dal domino delle stringhe a quello numerico tramite il metodo “transform” applicato al Train Labled Data e al Test Data.

In fine la funzione ritorna il Train e il Test datasets vettorializzati insieme all’oggetto TfidfVectorizer stesso (solamente per motivi di debug). Tramite la funzione “plot\_vector” nel file Data\_analisys siamo riusciti a plottare il risultato della vettorializzazione:

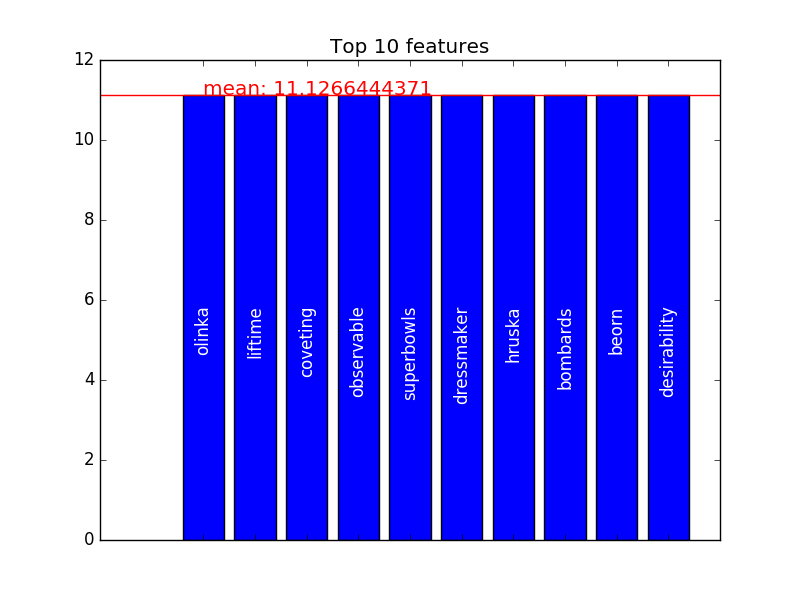


Sull’asse delle ascisse sono ripostati il numero delle parole del dizionario creato mentre sull’asse delle orinate il peso associato. Per una migliore comprensione del grafico abbiamo diviso l’asse y con una serie di rette in rosso e un numero che rappresenta le parole che hanno ricevuto un punteggio nel range di valori unitario. E’ interessante notare come più del 50% delle parole sia compreso nei range 9-10, 10-11 e >11, da qui la scelta di impostare il massimo numero delle features a 20000.

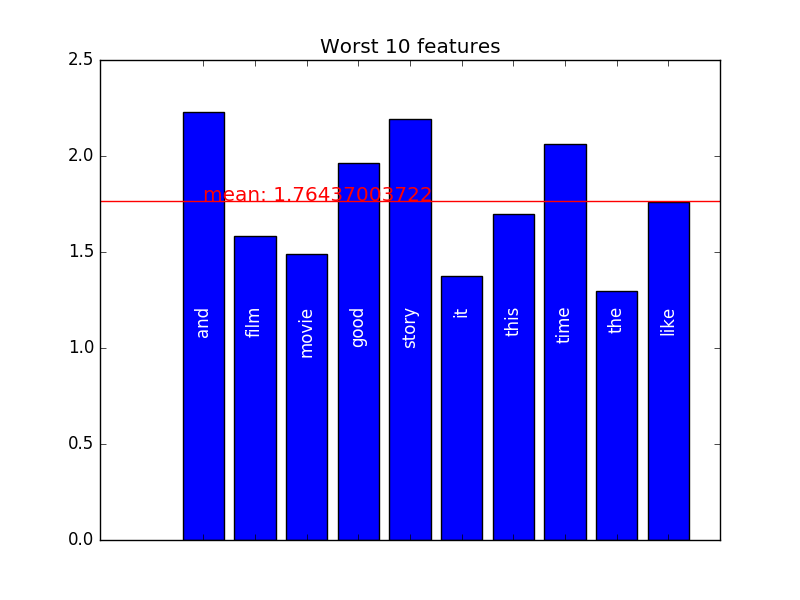
**Dimensionality\_reductionKB**

Dopo aver trasformato le stringhe in vettori, ci siamo posti l’obbiettivo di ridurre ulteriormente il numero di features. Il problema con la funzione precedente è che questa assegnava alle parole un peso in base alla frequenza all’interno del dataset, ma non tenva conto del significato in relazione alla predizione, anche perché la funzione non prende in input le Y.

Infatti abbiamo plottato le migliori 10 features e le peggiori 20 per il TFIDF tramite la funzione plot\_top\_n\_words nel file Data\_analisys e abbiamo ottenuto i seguenti risultati:



Come possiamo vedere le parole associati ai pesi maggiori non hanno particolare comunanza semantica ne interesse per valutare una recensione.

Mentre nelle peggiori 20 sono presenti sia stop words che parole molto interessanti quali “good” e “like”.

Per questo abbiamo usato la funzione [SelectKBest](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html) unito alla stima [Chi2](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2) della libreria sklearn.

Adesso passiamo a spiegare la teoria e il funzionamento dietro queste due funzioni.

Chi2

Prima di entrare nel particolare ricordiamo a caratteri generali che il chi2 viene utilizzato per misurare la dipendenza tra variabili stocastiche, in questo caso tra una feature e le Y, cosi facendo eliminiamo le features che risultano indipendenti dalle Y e che quindi non hanno importanza nella predizione.

Abbiamo scelto la funzione di stima chi2 per due ragioni fondamentali:

* Il nostro dataset è rappresentato da una matrice sparsa, ovvero da una matrice con una maggioranza di zeri. L’opposto di una matrice sparsa è una matrice densa, in cui vi è un’alta probabilità di trovare una diversa combinazione di features per ogni riga. La classificazione, o più in generale il train su di una matrice densa rende il fitting estremamente lento e impreciso, per questo abbiamo optato per la funzione chi2 che effettua la stima delle features senza rendere la matrice densa.
* L’altra ragione, scontata, ma di fondamentale importanza, risiede nelle Y. Infatti queste sono categoriche (e binarie), se cosi non fosse avremmo a che fare con un problema di regressione intrattabile con il test del chi2.

Volendo entrare di più nel dettaglio il chi2 si basa sull’ipotesi nulla, ovvero date due variabili categoriche questa sono assunte essere indipendenti tra loro finche non viene dimostrato il contrario.

Per smentire l’ipotesi nulla il procedimento chi2 è il seguente:

* Innanzitutto si calcola il cosi detto “Expected Count” [EC] per ogni cella del dataset:
  + RT= sarà la somma di tutte le features di una riga
  + CT= è la somma di tutte le righe di una features
  + SS= è la somma su tutte le righe e tutte le colonne
  + La formula applicata è EC=(RT\*CT)/SS
* Quindi l’EC rappresenta il valore che la cella avrebbe se le due variabili fossero davvero indipendenti. Il procedimento chi2 ora passa a rispondere alla seguente domanda: “La differenza tra L’EC e i valori delle celle è cosi importante da confermare una interdipendenza delle due variabili?”.
* Introducendo le seguenti variabili:
  + : valore originale della cella i-esima
  + : valore dell’EC della cella i-esima
  + = numero di righe
  + = numero di colonne
  + , denominato degree of freedom
  + n=numero di colonne\* numero di righe
* L’equazione del chi2 è
* Adesso bisogna calcolare il valore di P e per far ciò abbiamo bisogno del parametro DF e del livello di significatività [LS], il quale ci da una misura del livello di certezza che volgiamo ottenere dal risultato. Valori bassi di LS indicano che ci sia una bassa probabilità che i risultati siano casuali, solitamente questa parametro viene impostato a 0.05, ovvero il 5% di possibilità che l’evento sia riproducibile in un random sampling.
* L’ultimo passaggio sta nel cercare il valore di nelle tabelle del chi2, utilizzando il DF. Se il VC è minore di 0.05 possiamo sicuramente smentire l’ipotesi nulla e quindi esiste una correlazione tra le variabili, altrimenti non possiamo rigettare l’assunzione iniziale.