IDEA:

My project talks about satisfaction of employees in famous companies such as google, amazon , Microsoft, Netflix… I took the problem from a paper.

In the dataset there are five variables I decided to use **company, summary, pros, cons, overall ratings** with the name of the company, a brief comment on the company, the pros and cons and a rating from 1 to 5 for the firm

1. I want to perform a sentimental analysis on the summaries to get the general feeling of employees in the different companies (should I consider it as a classification problem with the ratings considered as labels or should I create a new column “polarity” with the labels “bad”, “neutral”, “good” and doing a classification problem with these labels?)

Considering different classifiers Naïve Bayes, Support Vector Machine and I want to study other methods

1. the pros and cons to understand the main topics and field in which a company should improve or not

Gensim for topic modeling come LDA (Latent Dirichlet Allocation) o scikit-learn for NMF (Non-Negative Matrix Factorization)

TF-IDF (Term Frequency-Inverse Document Frequency)

I know that it is based on how good the report it is written and also on the approach to the problem, but in my university it’s important to have high grade so I do not want to lose time on a not worthing project so do you think that is it possible to say if this could be graded with the maximum grade?

* le colonne importanti sono:

company, job-title, summary, pros, cons, overall ratings (gli altri ratings decidi di non considerarli)

(le company sono google, microsoft, netfix, ….)

* prima di tutto devi dividere il dataset in training e test quindi prendi il 20% dei dati di ogni azienda a random e li metti in un dataset test (considera il 20 % di ogni azienda e non il 20 % totali)
* il problema è quello di predirre la OVERALL RATINGS delle aziende
* Poi provi tutti i metodi di classificazione che conosci e calcoli i diversi valori della metrica che calcola l’errore e li confronti
* Altra idea è quella di fare i barplot della classificazione e anche i grafici con i puntini divisi per colori dopo aver fatto una riduzione della dimensionalità del dataset in 2D
* Puoi studiare i risultati sia in base all’azienda a cui ti stai riferendo e anche all’interno dell’azienda se li dividi in base al job-title
* Puoi anche studiare l’imbalance del dataset( ie in un ratings ci sono molti più valori degli altri)
* Un’idea potrebbe essere quella di avere due approcci diversi al problema:

visto che stiamo parlando di una sentiment analysis potremmo risolvere tutto il problema come una classificazione in ratings da 1 a 5 oppure come un sentiment analysis problem ovvero mettendo altre due colonne ie polarity e subjectivity

Cose che si potrebbero aggiungere in un secondo studio dato che hai le date in cui sono state scritte le reviews potresti studiare come cambiano le reviews nel tempo

Per caclolare l’errore puoi utilizzare

Mean Absolute Error (MAE) MAE = sum(over i) |pi − ri|/ n

and Root Mean Squared Error (RMSE) RMSE =sqrt( sum(over i) (pi − ri)^2/n)

studio I pros e i contro di ogni azienda e capisco quali sono i main topic per i pros e i contro di ogni azienda più discussi così l’azienda caisce dove deve migliorare

con la sentiment analysis del summary tra bene, male o neutro capisco invece quanto l’azienda deve migliorare in generale e se ho costruito un buon predittore allora posso dire che l’analisi è affidabile e posso seguire quello che mi dicono le percentuali ovvero se l’azienda in generale deve migliorare o va bene così

COSA FA DOCUMENT EMBEDDINGS ( DA LINK <https://medium.com/@evertongomede/doc2vec-understanding-document-embeddings-for-natural-language-processing-ba244e55eba3>)

Doc2Vec, also known as Paragraph Vector, is an extension of Word2Vec, a popular word embedding technique. While Word2Vec generates word embeddings that represent individual words as dense numerical vectors, Doc2Vec extends this idea to create embeddings for entire documents or paragraphs. The core idea behind Doc2Vec is to encode the semantic meaning and context of a document into a fixed-length vector, allowing us to perform various NLP tasks efficiently.

**Architecture**

Doc2Vec employs two main architectures: Distributed Memory (DM) and Distributed Bag of Words (DBOW). Let’s discuss each of them:

1. **Distributed Memory (DM)**: In the DM architecture, the model takes both the context of words and a unique document ID as input to predict the target word. It combines the document vector and the word vectors from the context to make the prediction. During training, the document vector becomes the representation of the entire document.
2. **Distributed Bag of Words (DBOW)**: In the DBOW architecture, the model ignores the word order and uses only the document ID as input to predict the target word. Here, the document vector is inferred without considering the context. DBOW is computationally less expensive than DM and is more suited for large corpora

Training

The training process for Doc2Vec involves updating the word vectors and document vectors iteratively to maximize the likelihood of predicting the correct words in a context. Similar to Word2Vec, Doc2Vec employs techniques like negative sampling or hierarchical softmax to speed up the training process.

**Document Classification**: By converting documents into fixed-length vectors, Doc2Vec enables efficient document classification tasks, such as spam detection, sentiment analysis, and topic categorization

STEP DEL PROGETTO:

1. STUDIO DELLA COLONNA SUMMARIES E OVERALL-RATINGS

* DOC2VEC PER CAPIRE QUALE MODELLO UTILIZZARE
* NAIVE BAYES
* SUPPORT VECTOR MACHINE
* REGRESSION MODEL
* CONFRONTO TRA I DUE MODELLI E STUDIO ACCURACY DI QUESTI MODELLI E QUALE è IL MIGLIORE
* BALANCED DATASET PRENDENDO LO STESSO NUMERO DI REVIEWS DA OGNI AZIENDA E VEDO COSA SUCCEDE RIFACENDO TUTTO QUELLO CHE HO FATTO PRIMA
* CONCLUSIONI: QUALE AZIENDA HA IL NUMERO PIU’ ALTO DI 1 E DI 5 e cerchi di capire quali sono gli elemeni che spingono una determinata azienda a avere il numero di 1 o 5 più elevati

1. STUDIO DELLE COLONNE PROS E CONS

* CERCO ARGOMENTO PRINCIPALE IN OGNI COMPANY E RISPETTO AL LABEL 1,2,3,4,5
* CONCLUSIONE:CAPISCO PER OGNI AZIENDA SU COSA DEVE LAVORARE E SU COSA INVECE è FORTE in base anche a quello che era stato concluso precedentemente

Quello che vuoi fare con la parte due è confermare quello trovato nella parte 1 ovvero nella parte 1 capisci quali sono le parole più influenti per i numeri 1 e 5

Però fai un confronto solo con i risultati del migliore dei modelli così hai un motivo per averne provati di più.

Devi dividere le parole più frquenti per aziende e non puoi metterli tutti insieme

È utile scrivere che hai considerato due parole insieme perché così vedevi che era più facile capire quali erano i problemi dato le reviews osno di poche parole tipo massimo 20 parole è più normale per un dipendente usare parole semplici e conscise piuttosto ceh paroloni che vengono usati in testi più lunghi

Conclusione sulla pipeline:

pulizia dataset

clean dataset per essere sicuri che i vectorizer di naive e svm prendano le aprole esattamente come le voglio io

uso naive bayes con default value of the alpha

uso svm e trovo che è performante quasi uguale

uso nuovo metodo di vettorizzazione doc2vec e regression ma è il meno performante

mi concentro sul bayes che è il più performante e il più facile dal punto di vista interpretativo

tuning di bayes e trovo che il migliore alpha è ‘0.3 in termini di accuracy e f1 score

provo a utilizzare come metodo di vettorizzazione per il naive bayes i 2-grams perché è più facile recuperare le parole chiave

a questo punto posso recuperare le parole chiave che caratterizzano ogni azienda e i livelli 1 e 5 ovvero quelli più positivi o negativi

nline customer reviews are very insightful and have a direct impact on revenue for

hotel owners. However, since these reviews are written by thousands of users, there’s

a possibility that some reviews present grammar mistakes, resulting in a loss of data

quality as a drawback. With that in mind, this research focused on analyzing whether

spell-checking applied in the pre-processing phase of data, could be impacting topic

modeling analysis on hotel reviews. Topic modeling was explored using BERTopic as

the main algorithm and LDA as a baseline. As expected, BERTopic’s generated topics

revealed a better efficiency than LDA regarding their coherence and overall quality.

Surprisingly, the research revealed that spell-checking usage not only doesn’t seem to

bring more value to the analysis but also disclosed some negative aspects (words that

were incorrectly tagged as misspelled) at a cost of higher computational demand.