Sviluppo di un recommendation system con modelli di machine learning

Giugno 2020

Enrico Ghidoni

279007@studenti.unimore.it

Introduzione

Questo report ha lo scopo di illustrare l'implementazione di un recommendation system realizzato interamente su Google Cloud Platform per il suggerimento di varietà di vino sulla base dei gusti espressi da un utente. Il sistema di ricerca all'interno di un dataset contenente recensioni[1] di diversi vini è basato su modelli di machine learning, che implementano tecniche di analisi del testo per raggiungere lo scopo. L'obiettivo del progetto è quello di allontanarsi da un sistema di ricerca manuale ed esplorare le possibilità offerte da un'architettura moderna, basata interamente su servizi cloud e machine learning, che rappresentano lo stato dell'arte nello sviluppo di interazioni con gli utilizzatori del sistema.

Analisi del dataset

Il dataset di riferimento per l'applicazione fornisce circa 130 mila recensioni di vini, estratte da un sito web di amatori che condividono opinioni. La prima idea per l'utilizzo del dataset per la costruzione del recommendation system è stata quella di consigliare agli utenti il nome della bottiglia che più si avvicinasse alle informazioni raccolte dal chatbot. Valutando rapidamente i dati, è tuttavia evidente che tale strada risulta impossibile da percorrere in quanto sono presenti all'incirca 119 mila nomi di bottiglie differenti all'interno del dataset, di conseguenza ogni bottiglia potrebbe contare su poco più di una recensione per distinguerla all'interno della collezione. E' necessario un numero di recensioni sostanziale per ottenere un buon risultato da un qualsiasi modello di machine learning, pertanto si è deciso di considerare come output del recommendation system la varietà di vino anziché il nome della bottiglia. Le varietà di vino presenti sono infatti nettamente inferiori (circa seicento). Come diretta conseguenza si ha una numero medio di recensioni per varietà accettabile, nonostante la distribuzione dei valori effettivi sia fortemente sbilanciata. Sono state quindi scartate le varietà con un numero di recensioni inferiore alla media, ottenendo un totale finale di 62 varietà di vino.
Siccome lo scopo del chatbot è quello di consigliare l'utente sulla base dei suoi gusti personali, la feature presa come input del modello di machine learning è il testo della recensione (denominato *description*). Si è quindi deciso di trattare il problema come una sorta di motore di ricerca, in cui l'utente fornisce una descrizione del vino desiderato e il modello effettua una classificazione in base alle recensioni presenti nel dataset, per ottenere la varietà di vino più rilevante. A tale scopo sono state adottate tecniche di Natural Language Processing (NLP) per discretizzare i testi delle recensioni, il risultato di questo processo rappresenta il dataset di

[1] (2017, November 27). Wine Reviews | Kaggle. https://www.kaggle.com/zynicide/wine-reviews

addestramento del modello. Al fine di valorizzare i termini specifici presenti nelle descrizioni sono dapprima state rimosse le cosiddette *stop-words* (ovvero parole come articoli, che non portano informazioni utili), è stata poi applicata una trasformazione per ottenere lo score di *Term Frequency - Inverse Document Frequency*[2], che attribuisce a ciascuna parola presente nel vocabolario (ovvero le parole contenute nelle recensioni) un peso in base alla presenza all'interno di un documento (in questo caso le recensioni stesse) e un valore di importanza basato sulla "rarità" della parola. La stessa tecnica di *preprocessing* viene applicata alle descrizioni inserite dagli utenti, in modo tale da ottenere un vettore contenente valori simili a quelli utilizzati per l'addestramento del modello. Il modello utilizzato è di tipo Support Vector Classification (con kernel *rbf*, non lineare)[3] e i risultati ottenuti sono discreti, il valore di *precision* ha una media di circa *0.7* nonostante sia variabile a seconda della varietà di vino.

[2] Inverse document frequency - Stanford NLP.
https://nlp.stanford.edu/IR-book/html/htmledition/inverse-document-frequency-1.html
[3] 1.4. Support Vector Machines — scikit-learn 0.23.1 .... http://scikit-learn.org/stable/modules/svm.html

	precision	recall	f1-score	support
Aglianico	0.77	0.54	0.63	91
Albariño	0.61	0.50	0.55	112
Barbera	0.79	0.60	0.68	180
Blaufränkisch	0.73	0.46	0.57	71
Bordeaux-style Red Blend	0.68	0.78	0.73	1735
Bordeaux-style White Blend	0.72	0.53	0.61	297
Cabernet Franc	0.64	0.40	0.49	353
Cabernet Sauvignon	0.61	0.72	0.66	2399
Carmenère	0.63	0.49	0.55	137
Champagne Blend	0.73	0.62	0.67	356
Chardonnay	0.72	0.87	0.79	2985
Chenin Blanc	0.70	0.49	0.57	146
Corvina, Rondinella, Molinara	0.75	0.70	0.72	164
Gamay	0.78	0.61	0.69	270
Garganega	0.71	0.67	0.69	70
Garnacha	0.60	0.42	0.49	79
Gewürztraminer	0.69	0.58	0.63	258
Glera	0.81	0.88	0.84	171
Grenache	0.61	0.26	0.37	179
Grüner Veltliner	0.75	0.74	0.75	342
Malbec	0.58	0.51	0.54	639
Melon	0.82	0.48	0.60	65
Meritage	0.54	0.11	0.18	64
Merlot	0.58	0.45	0.51	772
Montepulciano	0.69	0.33	0.45	60
Moscato	0.80	0.54	0.65	90
Mourvèdre	0.73	0.25	0.38	63
Nebbiolo	0.72	0.84	0.78	717
Nero d'Avola	0.81	0.56	0.67	85
Petit Verdot	0.75	0.30	0.42	61
Petite Sirah	0.59	0.39	0.47	180
Pinot Blanc	0.59	0.25	0.35	109
Pinot Grigio	0.66	0.59	0.62	248
Pinot Gris	0.64	0.55	0.59	350
Pinot Noir	0.71	0.83	0.77	3233
Port	0.74	0.61	0.67	173
Portuguese Red	0.66	0.71	0.68	601
Portuguese White	0.70	0.54	0.61	292
Primitivo	0.71	0.36	0.48	55
Prosecco	0.93	0.79	0.85	67
Red Blend	0.68	0.65	0.67	2262
Rhône-style Red Blend	0.69	0.66	0.67	358
Rhône-style White Blend	0.82	0.68	0.74	103
Riesling	0.78	0.84	0.81	1271
Rosé	0.72	0.77	0.75	922
Sangiovese	0.65	0.56	0.60	721
Sangiovese Grosso	0.78	0.88	0.83	172
Sauvignon	0.86	0.69	0.76	86
Sauvignon Blanc	0.68	0.73	0.71	1233
Shiraz	0.70	0.58	0.63	193
Sparkling Blend	0.72	0.69	0.70	530
Syrah	0.62	0.58	0.60	1024
Tempranillo	0.51	0.45	0.48	455
Tempranillo Blend	0.35	0.17	0.23	130
Torrontés	0.63	0.73	0.68	56
Touriga Nacional	0.75	0.24	0.36	50
Verdejo	0.76	0.51	0.61	63
Vermentino	0.82	0.55	0.66	58
Viognier	0.61	0.38	0.47	221
White Blend	0.75	0.58	0.65	606
Zinfandel	0.64	0.64	0.64	681
Zweigelt	0.79	0.44	0.57	52
micro avg	0.69	0.69	0.69	29566
macro avg	0.70	0.56	0.61	29566
weighted avg	0.68	0.69	0.68	29566

Figura 1. Classification report del modello basato sulla descrizione

Non si è ritenuto necessario utilizzare ulteriori tecniche di manipolazione dei dati, tuttavia il vocabolario risultante dalla rimozione delle stop-words è composto da circa 40 mila termini e risulta, pertanto, inutilizzabile all'interno di un chatbot interattivo. Infatti, per aiutare gli utenti meno esperti o più pigri nella selezione delle varietà di vino si è adottato un secondo approccio.

Pur mantenendo la modalità sopra descritta di inserimento di una descrizione, è stata preso in considerazione una lista ridotta di descrizioni riguardanti i diversi aspetti di un vino[4]. In questo modo è possibile guidare l'utente tramite il chatbot nella selezione degli aggettivi di suo interesse, in una modalità più semplice per non intenditori. A tale scopo è stato sviluppato un secondo modello di machine learning. Dai testi delle recensioni è stato estratto un dataset complementare che contiene (in versione discretizzata tramite colonne indicatrici), per ogni termine dell'insieme delle descrizioni, un valore che ne indica la presenza o meno all'interno del testo stesso. Per questo modello l'input corrisponde ad un vettore contenente, come per il dataset complementare, la presenza o meno degli aggettivi indicati dall'utente, mentre l'output rimane la varietà di vino. I primi risultati ottenuti con questa tecnica non sono stati soddisfacenti a causa del grosso sbilanciamento del numero di campioni per varietà, è stata quindi effettuata un'operazione di *oversampling* delle varietà con un minor numero di recensioni in modo da bilanciare il dataset. Il modello utilizzato è un Random Forest Classifier composto di cento Decision Tree Classifier, i risultati di classificazione sono anche in questo caso sufficienti per l'applicazione considerata e compongono una buona base di partenza.

[4] (2013, September 10). Subway style Wine Descriptions Chart (Infographic) | Wine Folly. https://winefolly.com/tips/wine-descriptions-chart-infographic/

	precision	recall	f1-score	support
Aglianico	0.63	0.75	0.69	3064
Albariño	0.78	0.74	0.76	3083
Barbera	0.81	0.63	0.71	3086
Blaufränkisch	0.58	0.73	0.65	3071
Bordeaux-style Red Blend	0.40	0.24	0.30	3044
Bordeaux-style White Blend	0.67	0.49	0.56	3117
Cabernet Franc	0.76	0.51	0.61	3150
Cabernet Sauvignon	0.58	0.22	0.32	3103
Carmenère	0.44	0.64	0.52	3122
Champagne Blend	0.50	0.47	0.49	3004
Chardonnay	0.67	0.20	0.31	3024
Chenin Blanc	0.68	0.60	0.64	3048
Corvina, Rondinella, Molinara	0.60	0.59	0.60	3016
Gamay	0.58	0.48	0.52	3040
Garganega	0.58	0.74	0.65	2969
Garnacha	0.53	0.66	0.59	2981
Gewürztraminer	0.77	0.63	0.69	3065
Glera	0.58	0.67	0.62	3083
Grenache	0.75	0.62	0.68	3081
Grüner Veltliner	0.52	0.58	0.55	3068
Malbec	0.80	0.45	0.57	3097
Melon	0.62	0.69	0.65	3089
Meritage	0.56	0.65	0.60	3080
Merlot	0.57	0.36	0.44	3047
Montepulciano	0.68	0.83	0.75	3036
Moscato	0.57	0.63	0.60	3030
Mourvèdre	0.33	0.78	0.47	3013
Nebbiolo	0.82	0.65	0.72	3085
Nero d'Avola	0.74	0.73	0.74	3101
Petit Verdot	0.43	0.71	0.54	3129
Petite Sirah	0.66	0.55	0.60	3170
Pinot Blanc	0.55	0.62	0.58	3030
Pinot Grigio	0.67	0.49	0.56	3081
Pinot Gris	0.53	0.49	0.51	3095
Pinot Noir	0.47	0.13	0.21	3089
Port	0.25	0.61	0.35	3096
Portuguese Red	0.46	0.42	0.44	3203
Portuguese White	0.63	0.51	0.56	3049
Primitivo	0.54	0.70	0.61	3015
Prosecco	0.36	0.81	0.50	3031
Red Blend	0.76	0.26	0.39	3122
Rhône-style Red Blend	0.65	0.46	0.54	2990
Rhône-style White Blend	0.41	0.67	0.51	3074
Riesling	0.77	0.42	0.55	2937
Rosé	0.57	0.43	0.49	3050
Sangiovese	0.82	0.54	0.65	2990
Sangiovese Grosso	0.71	0.65	0.68	2996
Sauvignon	0.65	0.77	0.70	3085
Sauvignon Blanc	0.74	0.42	0.53	3034
Shiraz	0.48	0.57	0.52	3079
Sparkling Blend	0.80	0.50	0.61	3145
Syrah	0.73	0.36	0.48	3029
Tempranillo	0.70	0.49	0.58	3042
Tempranillo Blend	0.75	0.66	0.70	3066
Torrontés	0.68	0.85	0.76	3036
Touriga Nacional	0.58	0.81	0.67	3087
Verdejo	0.73	0.78	0.75	3141
Vermentino	0.49	0.76	0.60	3021
Viognier	0.65	0.57	0.61	3025
White Blend	0.83	0.42	0.56	3165
Zinfandel	0.67	0.44	0.53	3087
Zweigelt	0.36	0.73	0.48	3029
micro avg	0.57	0.57	0.57	190015
macro avg	0.62	0.57	0.57	190015
weighted avg	0.62	0.57	0.57	190015

Figura 2. Classification report del modello basato su lista di aggettivi

Serving delle predizioni sul cloud

Per ottenere le classificazioni dai due modelli sviluppati è stato utilizzato il servizio AI Platform di GCP, che offre un sistema di predizioni online. I modelli sono interrogati tramite API REST. AI Platform supporta nativamente i modelli costruiti tramite la libreria *scikit-learn*, permettendo un approccio immediato per quanto concerne il

deployment. La piattaforma supporta inoltre un meccanismo di *custom routine* che permette di definire un flusso di *preprocessing* e *prediction*.

Gestione delle conversazioni

Per la gestione delle interazioni testuali con gli utenti viene utilizzato il servizio Dialogflow di GCP, che mette a disposizione una piattaforma supportata da elementi di machine learning per la comprensione di linguaggio naturale. Per ottenere il match della varietà di vino, una volta ottenute le informazioni necessarie, l'agente Dialogflow inoltra una richiesta di *fulfillment* tramite API REST ad un servizio (descritto nella sezione seguente) che interroga uno dei due modelli di machine learning, che restituisce la varietà con probabilità di match più alta. In ultimo, Dialogflow riporta il match all'utente e chiude la conversazione.
Dialogflow offre nativamente integrazioni con diversi servizi di messaggistica, per semplificare l'architettura del sistema nella parte di frontend è stata utilizzata l'integrazione con il servizio Telegram. Questa integrazione libera lo sviluppo dalla gestione della comunicazione con i client, delegando interamente il carico di lavoro a Dialogflow.

Connessione tra Dialogflow e AI Platform

Il match con la varietà sulla base degli input forniti dall'utente e raccolti dall'agente Dialogflow viene effettuato da uno script python, in esecuzione sul servizio Cloud Functions di GCP. Sulla base della modalità scelta dall'utente, descrizione o lista di aggettivi, lo script si occupa del parsing della richiesta proveniente dall'operazione di fulfillment di Dialogflow e la inoltra al modello corretto su AI Platform. L'utilizzo di Cloud Functions riduce significativamente i tempi di sviluppo rispetto, ad esempio, al servizio App Engine per lo sviluppo di programmi che consumano API. Non richiede alcun tipo di setup o installazione di dipendenze e fornisce un ambiente pronto all'utilizzo, pur mantenendo la possibilità di sviluppare e servire applicazioni complesse e articolate.
Nonostante il deployment dei servizi AI Platform e Cloud Functions sia stato effettuato all'interno della stessa *region*, si riscontrano alcuni problemi di latenza. Dialogflow pone come vincolo di esecuzione di un fulfillment cinque secondi e, nel caso di utilizzo del modello basato sulla lista di aggettivi, a volte questo timeout non è rispettato. Il servizio Cloud Functions non supporta sistemi di *caching*, tuttavia sono descritti diversi comportamenti da implementare per migliorare le performance del servizio.

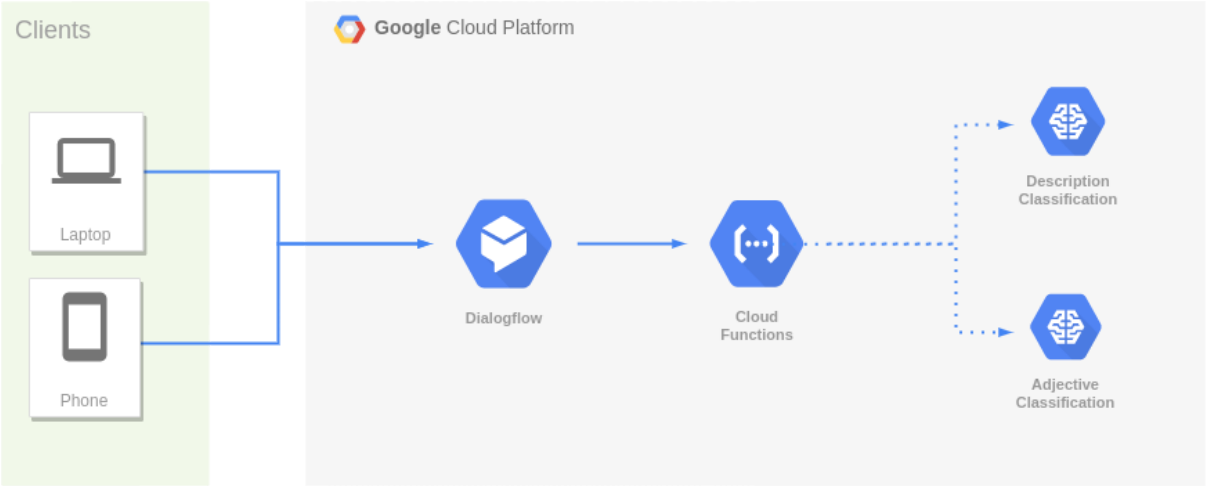

Figura 3. Architettura generale di GrapeMaster

Conclusioni e sviluppi futuri

L'applicazione realizzata sfrutta solo alcuni dei campi presenti all'interno del dataset, ma sono presenti svariate indicazioni come, ad esempio, il prezzo di una bottiglia (da cui è possibile ottenere il prezzo medio per varietà di vino) e la regione di provenienza, attraverso cui è possibile realizzare dei filtri per la ricerca. Un altro campo che è possibile sfruttare è l'username Twitter dell'autore della recensione, tramite cui si può costruire una valutazione della rilevanza in base, ad esempio, al numero di follower. Inoltre, questo progetto rappresenta a tutti gli effetti un *recommendation system* che sfrutta un meccanismo di *content-based filtering*. Raccogliendo nel tempo informazioni sulle preferenze di un utente è possibile implementare modelli diversi e più precisi, che si basano sulla similarità tra gli utenti (ad esempio utilizzando tecniche di *collaborative filtering*).

L'utilizzo di Google Cloud Platform ha sostanzialmente mantenuto la promessa di permettere agli sviluppatori di concentrarsi sulle funzionalità dell'applicazione che si intende costruire, anziché sulla gestione dell'infrastruttura. Aggiungere funzionalità, come quelle descritte nel precedente paragrafo, risulta dunque relativamente semplice.