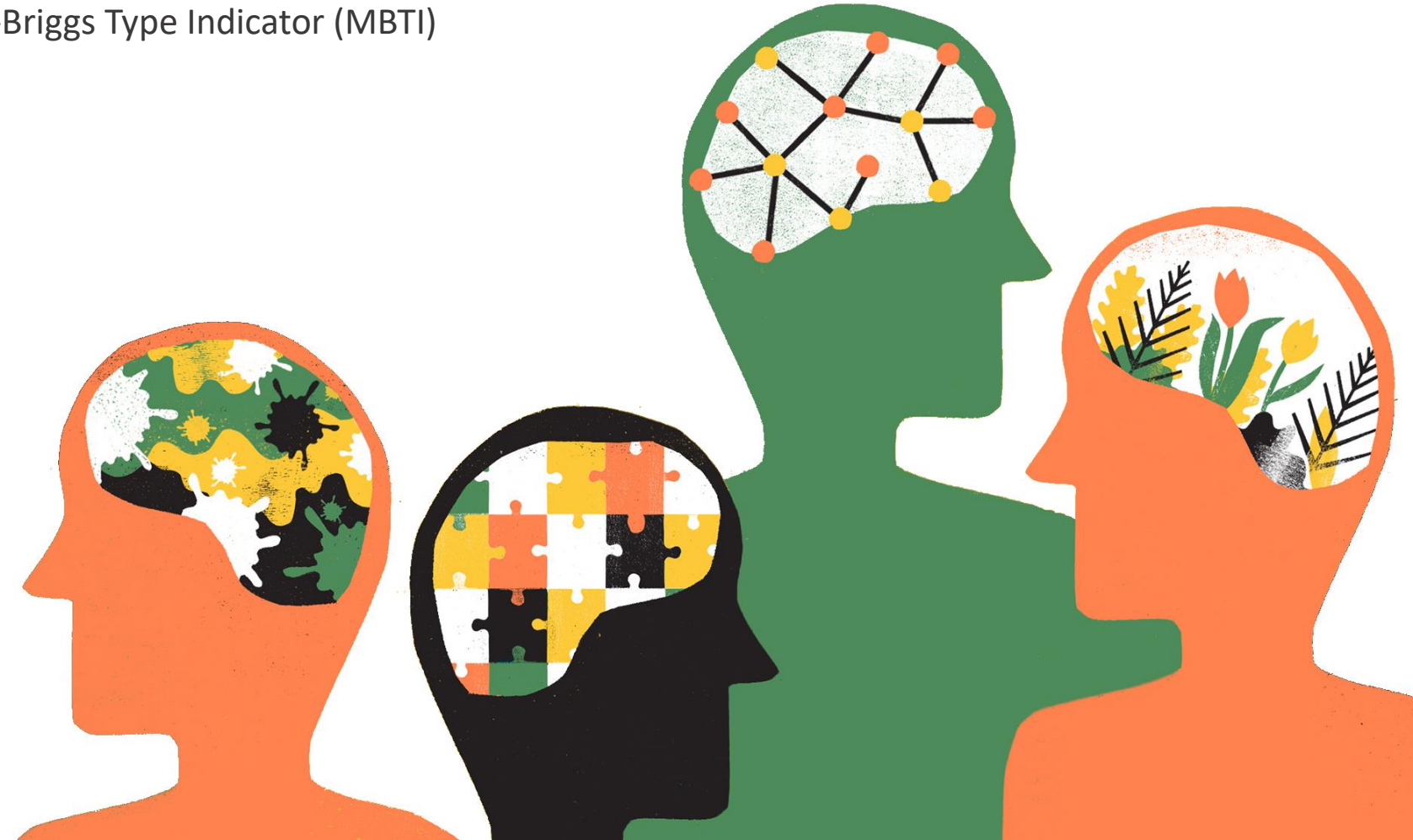
PERSONALITY TYPES PREDICTOR

Categorize Twitter profiles live into a Myers-Briggs Type Indicator (MBTI)



Leonetti Antonio - Mat. 699511

PERSONALITY TYPES PREDICTOR

Questo progetto nasce come replica ed estensione di un lavoro già esistente, discusso nel report disponibile nell'archivio di **web.stanford.edu**[1]. Il progetto utilizza il machine learning per creare il classificatore per il **Myers-Briggs Type Indicator** (**MBTI**) [2], una metrica che utilizza un risultato di 4 lettere (es. INFJ o ENFP) per riassumere le diverse caratteristiche della personalità in termini di come gli individui percepiscono il mondo e prendono decisioni.

Il lavoro presentato espande tutto quello che viene discusso nel report e presenta delle estensioni al progetto originale, come l'utilizzo dell'API di **Twitter** per acquisire il feed di un account per poi classificarlo in tempo reale e il confronto tra i due classificatori di tipi differenti citati nel report per verificare quale approccio sia il più vantaggioso.

Normalmente questo risultato deriva dall'utilizzo di questionari e test psicometrici somministrati a ciascuna persona, ma qui otterremo automaticamente un risultato attraverso una semplice interfaccia grafica web, alla quale basta fornire soltanto l'handle del proprio account Twitter.

Il tutto è realizzato principalmente in **Python**, ad eccezione dell'interfaccia web che viene comunque gestita da python tramite l'utilizzo di **Flask**.

INTRODUZIONE - MBTI

L'obiettivo dell'MBTI è quello di consentire agli intervistati di esplorare e comprendere la propria personalità, compresi i propri gusti, antipatie, punti di forza, punti deboli, possibili preferenze di carriera e compatibilità con altre persone. Nessun tipo di personalità è migliore o peggiore dell'altra. Il questionario stesso è centrato sull'analisi di quattro dicotomie:

Estroversione (E) – Introversione (I)

La dicotomia estroversione-introversione si utilizza come un metodo per descrivere come le persone rispondono e interagiscono con il mondo che li circonda. Gli estroversi sono "volti verso l'esterno" e tendono ad essere orientati all'azione mentre gli introversi tendono ad essere orientati al pensiero.

Sensibilità (S) – Intuizione (N)

Questa implica l'osservazione di come le persone raccolgono informazioni dal mondo che le circonda. Le persone che preferiscono la sensibilità tendono a prestare molta attenzione alla realtà, in particolare a ciò che possono imparare dai propri sensi, mentre coloro che preferiscono l'intuizione prestano maggiore attenzione alle possibilità e alla pianificazione del proprio futuro.

INTRODUZIONE - MBTI

Pensare (V) – Sentire (F)

Questa scala si concentra su come le persone prendono decisioni in base alle informazioni che hanno raccolto dalle loro funzioni di rilevamento o intuizione. Le persone che preferiscono pensare danno maggiore enfasi ai fatti e ai dati oggettivi. Coloro che preferiscono i sentimenti sono più propensi a considerare le persone e le emozioni quando arrivano a una conclusione.

Giudicare (J) – Percepire (P)

L'ultima dicotomia riguarda il modo in cui le persone tendono a trattare con il mondo esterno. Coloro che tendono a giudicare preferiscono pensiero strutturato e le decisioni ferme. Le persone che propendono per la percezione sono più aperte, flessibili e adattabili

DATASET

Il **dataset**[3] utilizzato è disponibile pubblicamente su **Kaggle**. Si tratta di un csv contenente oltre 8600 righe di dati, nel quale ognuna è composta da due colonne con i dati relativi ad una profilo:

- Il **tipo** (codice/tipo MBTI di 4 lettere di questa persona)
- Una sezione di ciascuna degli ultime **50 post** che hanno pubblicato ogni voce separata da 3 caratteri pipe "|||"

Questi dati sono stati raccolti tramite il forum **PersonalityCafe**, un forum online in cui agli utenti viene sottoposto prima il questionario per l'assegnazione del loro tipo MBTI e poi consente loro di chattare con altri utenti.

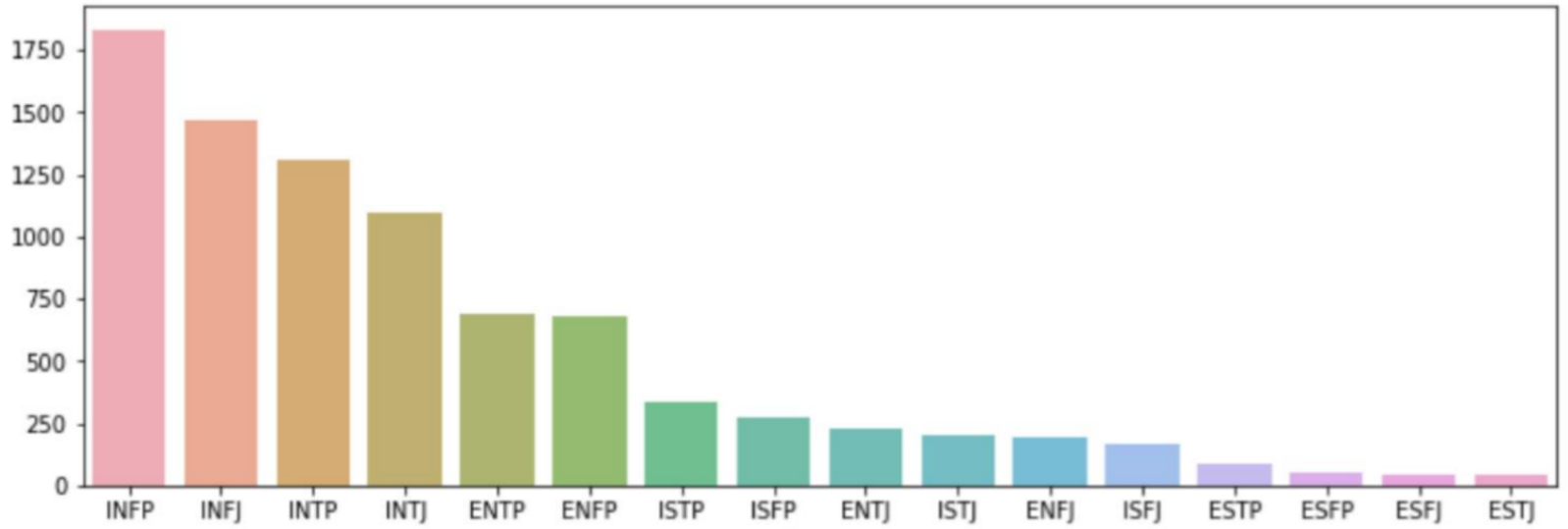
Poiché ci sono cinquanta post inclusi per ogni utente, il numero di post è circa 430'000.

```
>>> import pandas as pd
>>> df = pd.read_csv("dataset/mbti_1.csv")
>>> df.head()
   type                                               posts
0  INFJ  'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1  ENTP  'I'm finding the lack of me in these posts ver...
2  INTP  'Good one  _____   https://www.youtube.com/wat...
3  INTJ  'Dear INTP,   I enjoyed our conversation the o...
4  ENTJ  'You're fired.|||That's another silly misconce...
```

mbti_1.csv **-** *Anteprima delle prime 5 righe dei dati*

PROPORZIONALITÀ

L’unico problema in questo dataset è la rappresentazione non uniforme dei tipi, che non rispettano le proporzioni del mondo reale, che rende necessaria una modifica per riportare una proporzione uniforme nel test set. Pertanto, si è operata una selezione sul dataset selezionando il tipo di MBTI con il minor numero di post e seguendo il riferimento alle proporzioni trovate su **myersbriggs.org** [4] si è creato un test set che rifletta le proporzioni reali.



In questo modo si cercano di prevenire gli errori nei risultati, dovuti alla rappresentazione distorta delle classi nel set di test che sono indicati nel grafico sottostante.

STOP WORDS

Il dataset analizzato proviene da un forum online, diventa dunque necessario la rimozione di alcune parole che nell'analisi risulterebbero di poco valore. Ad esempio, nei post sono presenti le cosiddette stop words, link a siti esterni ed anche codici per rappresentare emoticons. Infine sono stati rimossi anche i riferimenti espliciti ai tipi stessi (ad es. 'INTJ', 'INFP', ecc.), in modo da prevenire al modello imparando a riconoscere i tipi di MBTI per nome. La rimozione è stata effettuata tramite l'utilizzo di **Natural Language Toolkit** (**NLTK**), una libreria Python open source per l'elaborazione del linguaggio naturale e con l'utilizzo delle **regex** per identificare links e emoticons/emoji.

LEMMATIZZAZIONE

Per la lemmatizzazione è stato utilizzato il **WordNetLemmatizer** sempre della libreria NLTK.
Lemmatizzare il testo significa ridurre più forme differenti della stessa parola ad un'unica parola radice. Questo ci permette di ricondurre più forme della stessa parola ad un unico significato e semplifica notevolmente il lavoro del classificatore.

TOKENIZZAZIONE

Utilizzando un tokenizzatore presente nella libreria **Keras**, sono state tokenizzate le **2500 parole** più comuni del testo lemmatizzato.

In questo modo il testo viene trasformato una serie di elenchi di interi selezionati dal nostro vocabolario, eliminando dal testo tutte le parole che non sono presenti nel dizionario. Così facendo si rimuovono quelle parole che non avrebbero avuto lo stesso impatto nel modello.

PADDING

Poiché i post tokenizzati sono di lunghezza molto variabile, procediamo rendendoli tutti della stessa lunghezza.

Viene allora aggiunto del **padding** per ogni post tokenizzato in modo tale che abbia esattamente **40 interi**.

Se sono presenti meno di 40 interi nel post tokenizzato, vengono aggiunti degli zero fino a quando non abbiamo raggiunto la dimensione necessaria, mentre vengono rimossi se superano la soglia.

EMBEDDING LAYER

Per l’embedding layer si usa una embedding matrix sotto forma di dizionario che mappa ogni parola lemmatizzata alla rappresentazione **GloVe** a 50 dimensioni di quella parola.

GloVe (**Global Vectors for Word Representation**) è un algoritmo per ottenere rappresentazioni vettoriali per le parole.

In questo caso utilizziamo un vettore disponibile sul **repository ufficiale**[5] pre-addestrato su dati presi da Twitter.

RECURRENT NEURAL NETWORK

Sempre seguendo le indicazioni del report e dato che il set di dati che viene utilizzato è composto da dati di testo sequenziali, confermiamo l’utilizzo di una rete neurale ricorrente (RNN).

Tra i vari tipi di reti neurali ricorrenti (RNN) per questo passaggio, come descritto nel report, si sceglie di utilizzare l'opzione LSTM che offre i migliori risultati. Dopo una serie di tentativi si dimostrano ottimali anche i valori dei parametri utilizzati per il livello LSTM.

Ci serviamo sempre della libreria Keras per importare i vari layer e modelli utilizzati.

DENSE LAYER E OTTIMIZZAZIONE

Infine, utilizziamo uno dense layer con la sigmoide come funzione di attivazione per produrre un valore compreso tra 0 e 1, che rappresenta la probabilità di classe prevista, poiché ci sono solo due classi.

Inoltre, seguendo il report, viene utilizzata binary crossentropy per la loss function (poiché ci sono solo due classi) e un ottimizzatore Adam dalla libreria degli optimizers per karas di tensorflow.

ESPERIMENTO

Una volta osservata la natura dell'indicatore Myers-Briggs ed anche la sproporzionalità delle classi nel test set, si sceglie di suddividere la classificazione delle 16 classi in quattro attività di classificazione, facendo riferimento alle quattro dicotomie dell’indicatore. Questo perché un tipo MBTI è composto da quattro classi binarie, dove ogni classe binaria rappresenta una dimensione della personalità come teorizzato dagli inventori del modello.

Pertanto si andranno ad addestrare quattro diversi classificatori binari, in modo tale che ciascuno sia specializzato in una delle dimensioni della personalità. La somma dei risultati andrà a rappresentare il tipo di indicatore.

Per confrontare i risultati con il report è stata svolta la classificazione dei post sul test set pre-elaborato e previsto la classe per ogni singolo post. Sono stati prodotti dunque l’accuracy e una matrice di confusione per ogni dimensione MBTI.

ESPERIMENTO

Allo stesso modo sono stati addestrati quattro classificatori Naive Bayes per le quattro dimensioni della personalità.

Questo confronto con un classificatore Bayesiano servirà a determinare se l'utilizzo della RNN riesca a produrre risultati

Superiori, come affermato nel report. Si utilizza un MultinomialNB come classificatore per il confronto dalla libreria **sklearn**.

«Due to the fact that our data set is composed of sequential text data, we decided to use a recurrent neural network in order to capture some of the information in the text data that would otherwise be ignored (e.g. as with a naive Bayes classifier).»

RISULTATI - RNN

Per la classificazione dei post sul test set precedentemente elaborato, abbiamo una previsione della classe per ogni singolo post. Ad entrambi è stata applicata la k-fold cross validation (CV) che consiste nella suddivisione dell'insieme di dati totale in k=5 parti di uguale numerosità e, a ogni passo, la kª parte dell'insieme di dati viene utilizzata come test set, mentre la restante parte costituisce il train set.

Qui seguono Il punteggio di accuratezza e una matrice di confusione per ogni dimensione MBTI:

I/E Accuracy: **59.1 %**

	Predicted I	Predicted E
Actual I	46374	37098
Actual E	31560	52904

N/S Accuracy: **64.0 %**

	Predicted N	Predicted S
Actual N	27739	15287
Actual S	15316	26738

T/F Accuracy: **59.9 %**

	Predicted F	Predicted T
Actual F	97762	76604
Actual T	63269	111286

P/J Accuracy: **55.8 %**

	Predicted P	Predicted J
Actual P	80067	66206
Actual J	63482	83854

Sebbene ciò sembri indicare una debole capacità complessiva del nostro modello di classificare correttamente tutte e 4 le dimensioni MBTI, va notato che, altri modelli che si concentrano sulla classificazione multiclasse di MBTI possono ottenere una maggiore precisione della classificazione perfetta, rischiando di sbagliare completamente la loro previsione. Il modello rappresenta un compromesso di questi due aspetti: otteniamo tassi inferiori di classificazione perfetta in cambio di tassi più elevati di classificazione approssimativamente corretta.

RISULTATI - NBC

Per la classificazione dei post sul test set precedentemente elaborato, abbiamo una previsione della classe per ogni singolo post, questa volta utilizzando i classificatori Bayesiani, che effettivamente, come affermato nel report, hanno prodotto dei risultati meno accurati.

Qui seguono Il punteggio di accuratezza e una matrice di confusione per ogni dimensione MBTI.

I/E Accuracy: **51.2%**

	Predicted I	Predicted E
Actual I	38451	45021
Actual E	36960	47504

N/S Accuracy: **51.8 %**

	Predicted N	Predicted S
Actual N	20362	22664
Actual S	18449	23605

F/T Accuracy: **52.2 %**

	Predicted F	Predicted T
Actual F	118074	56292
Actual T	110631	63924

P/J Accuracy: **50.4 %**

	Predicted P	Predicted J
Actual P	77032	69241
Actual J	76482	70854

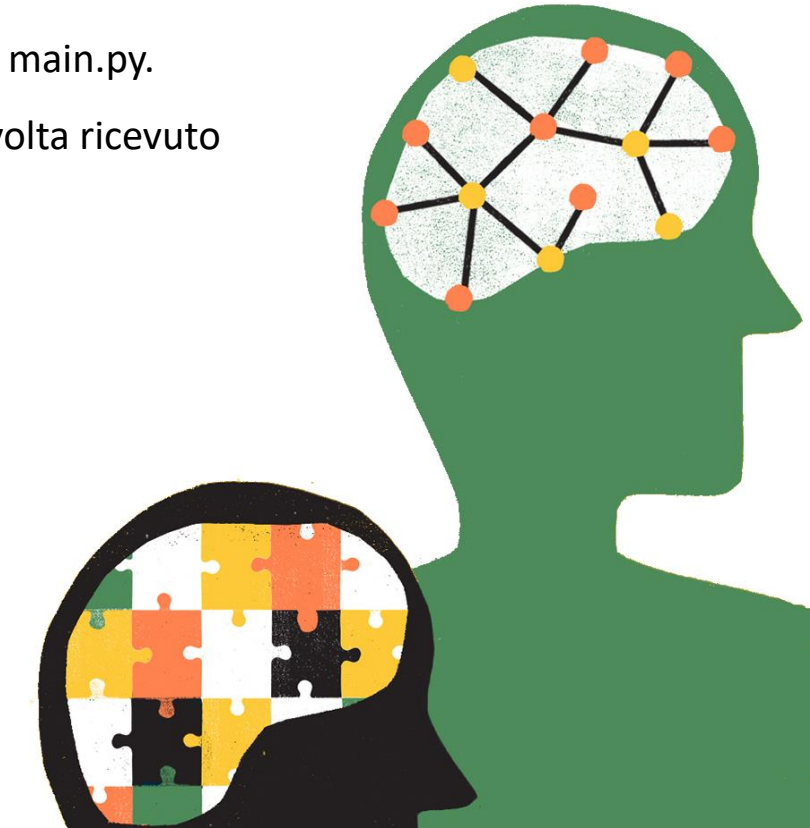
TWITTER

Tramite l'utilizzo dell'API fornita da Twitter dopo la creazione di un account da sviluppatore, il sistema riesce a recuperare in forma testuale gli ultimi 100 post dal profilo della persona analizzata. A questi viene eseguito lo stesso lavoro di rimozione, lemmatizzazione e tokenizzazione che è stato descritto in precedenza per il dataset. Su questi dati viene utilizzato il modello per determinare il tipo di MBTI.

INTERFACCIA

L'interfaccia è una semplice pagina web che può essere aperta in locale attraverso l'esecuzione di main.py.
Qui attraverso l'utilizzo di Flask avviene la comunicazione tra la pagina web e il sistema, che una volta ricevuto l'handle recupera i post e avvia la procedura.

Segue uno screen dell'interfaccia grafica a termine dell'esecuzione del processo.



TWITTER USER PERSONALITY PREDICTION

Lorem ipsum dolor sit amet consectetur adipisicing elit. Maxime mollitia, molestiae quas vel sint commodi repudiandae consequuntur voluptatum laborum numquam blanditiis harum quisquam eius sed odit fugiat iusto fuga praesentium optio, eaque rerum! Provident similique accusantium nemo autem.

OfficialDenzel

START

Your Myers-Briggs Personality Indicator is

INTP

Most of the time where do you choose to direct your energy?

EXTROVERSION

You usually direct your energy towards other people, objects, circumstances, and the world outside of you. You are more of an extrovert.

INTROVERSION

You usually direct your energy inwards to thinking and contemplating concepts, information, ideas, principles or beliefs. You are more of an introvert.

How do you take in and act on information?

SENSING

You want empirical information, hard facts and detailed data. You want to have a clear and practical understanding of every situation.

INTUITION

You would rather follow your gut on things, dive into the unknown, and are open to new things you have not tried before or know little about.

How do you decide on things and make choices?

THINKING

You analyse the information and facts available and calculate the possibilities using cold logic. Then you decide and act on your decision.

FEELING

Your decisions are based on what you believe and the things that you value and care for. You act by following the principles you hold on to.

How do you go about organising things around you and your life?

JUDGING

This means you want everything in your life to be as structured and organised as possible. You prefer things to be stable and predictable.

PERCEPTION

You would rather go with the flow, see how things unfold and turn out, flexibly responding to situations as they come.

RIFERIMENTI

[1] Report web.stanford.edu - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6839354.pdf

[2] Myers-Briggs Type Indicator (MBTI) - https://it.wikipedia.org/wiki/Indicatore_Myers-Briggs

[3] MBTI Dataset Kaggle - https://www.kaggle.com/datasnaek/mbti-type

[4] Proporzioni MBTI - https://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm

[5] GloVe - https://github.com/stanfordnlp/GloVe