Sistema di analisi dei report delle lezioni

Gruppo NLP a.a. 2019/2020

Università degli Studi di Roma Tor Vergata

23 gennaio 2020

La struttura dei dati

Codice	Data della lezione	Messaggio della lezione	Argomento di interesse	Argomento meno chiaro
2Be2	3/4/2019	Approccio positivo all'informatica	Programmazione	Nessuna
Q7Sj	3/4/2019	Interessare	La programmazione	Nessuna
...	...	...	...	...

interfaccia POSTagging

In questa schermata gli annotatori dovranno, per ogni token di una frase, assegnare un TAG ed un sentiment.
Come TAG e come guidelines si è scelto di seguire quelle fornite da **Universal Dependencies**:

https://universaldependencies.org/it/pos/

Il sentiment potrà essere: POS (positivo), NEG (negativo) oppure NEU (neutro). Si osserva che il sentiment risulta interessante solo per le ultime due colonne.

Interfaccia semantica

In questa schermata un annotatore, data una coppia di frasi, dovrà assegnare un grado di similarità che varia da 1 a 6.

- **6** Molto simili: si verifica se **S**-**V**-**O** sono uguali per entrambe le frasi e se le frasi sono formate dallo stesso numero di sotto-frasi.
- **5** Simili: si verifica se **S**-**V**-**O** presenti nelle due frasi sono sinonimi e se le frasi sono formate dallo stesso numero di sotto-frasi.
- **4** Abbastanza simili: si verifica se **S**-**V**-**O** sono uguali per entrambe le frasi, ma le frasi possono essere costituite da un numero di sotto-frasi differente.

- **3** Poco diverse: si verifica se **S**-**[V]**-**O** presenti nelle due frasi sono sinonimi, ma le frasi possono essere costituite da un numero di sotto-frasi differente.
- **2** Diverse: si verifica se **S**-**[V]**-**O** presenti nelle due frasi sono differenti, ma i termini presenti sono uno il contenuto/contenitore nell'/dell'altro.
- **1** Molto diverse: si verifica se **S**-**[V]**-**O** presenti nelle due frasi sono differenti e i termini presenti non hanno alcuna relazione di contenuto/contenitore.

Guideline in sintesi

Grado	Parole	Contenuto/contenitore	Numero frasi
6	stesse	-	stesso
5	sinonimi	-	stesso
4	stesse	-	diverso
3	sinonimi	-	diverso
2	diverse	si	-
1	diverse	no	-

Il - indica che quel valore è ininfluente per quel grado di similarità. L'ordine delle colonne implica il grado di rilevanza di tale campo, si va quindi dal campo più rilevante a quello meno rilevante.

Valutazione inter-annotator agreement

Una volta estratte a caso dal dataset 20 frasi e 20 coppie, si assegnano agli annotatori. Terminata l'annotazione, si esegue la valutazione del valore di inter-annotator agreement. Come misura di bontà si è scelto di usare:

Krippendorff, 1980

Superato lo scoglio dello 0.8, gli annotatori possono iniziare ad annotare un centinaio di frasi e di coppie in modo da formare un corpus adeguato.

Strumenti esistenti

Arrivati a questo punto si sono trovati strumenti adeguati per eseguire le dovute annotazioni in modo automatizzato.

1. Per il Part-Of-Speech Tagging si è individuato spaCy, avvalendoci del dependecy parsing: https://spacy.io/usage/linguistic-features
2. Per la sentiment analysis si è optato per Repustate: https://www.repustate.com/
3. Mentre per la similarità semantica si procede come segue. Data una frase, si estraggono **S**-**V**-**O** (ottenuti al punto 1) e questi costituiranno un vettore nello spazio. Per valutare la similarità tra due frasi, si applicherà la cosine similarity che sarà poi riscalata sui valori definiti in precedenza.

Clustering

Per ogni lezione, avremo le frasi pre-elaborate in modo automatico.
Ognuna di queste frasi costituirà un punto nello spazio.
Quello che si intende fare è definire, per ogni colonna, alcuni centroidi che meglio rappresentano lo spazio.
Visto che i centroidi molto probabilmente non sono rappresentati da frasi di senso compiuto, prenderemo il punto più vicino al centroide che rappresenterà la frase meglio caratterizzante di quel cluster.

Indicatori

L'utente finale, a questo punto, potrà definire alcuni parametri che saranno utilizzati in fase di clusterizzazione.
Potrebbe, per esempio, indicare un numero k di cluster in modo da vedere, per ogni lezione (o una specifica), quali sono gli argomenti emersi, gli argomenti di maggiore interesse oppure gli argomenti meno chiari. Il tutto verrebbe rappresentato con un grafico a torta diviso in k spicchi, ogni spicchio rappresenterà la popolazione del cluster inerente.

Altra possibilità, fornito un argomento della lezione con relativa data, vedere l'andamento della classe. Ovvero tale argomento costituirebbe il centroide e si andrebbe a vedere la distanza di tutte le risposte con il reale argomento e si ritornerebbe un istogramma, suddiviso per range di distanze, che rappresenterebbe l'affinità delle risposte.

Come ultimo indicatore si è pensato di, per ogni utente, vedere come è stato il suo percorso. Per far questo, si vedrebbe due fattori. Per il primo si va a verificare se l'argomento percepito è lo stesso dell'argomento emerso. Per il secondo si va a vedere se l'argomento meno chiaro ricade sempre nell'argomento emergente.

A questo punto si potrebbero pensare a dei possibili sviluppi, in cui si potrebbe tener conto anche dei voti finali. Una possibilità potrebbe essere ritornare un istogramma, che verrebbe suddiviso in due modi, la prima suddivisione sarebbe per voto; la seconda, fissato un voto, si avrebbero le misure di affinità definite nel paragrafo precedente.