Relazione del primo gruppo di esercizi

TLN_dicaro_1.1 and TLN_dicaro_1.2

Esercizio 1.1
Gruppo: Roberto Demaria

CONSEGNA:

- Date delle definizioni per quattro concetti (due concreti e due astratti), calcolare la similarità fra di esse.

SVOLGIMENTO:

I termini sono:

	Generico	Specifico
Concreto	Paper	Sharpener
Astratto	Courage	Apprehension

- Ho importato il file excel sotto forma di data frame per poterne estrarre le informazioni più agevolmente.
  Qui sotto si può vedere la testa del data frame creato grazie alla libreria pandas:

	Partecipante	Courage	Paper	Apprehension	Sharpener
1	2	Ability to face our own fears and do something...	Material derived from trees and used in sever...	fearful expectation or anticipation	Object used to shapen a pencil
2	3	the ability to face thing without fear	a type of material made from cellulose	A moode where one feel agitation	An object to sharpen a pencil
4	5	Inner strength thaht allow you to face particu...	Product obtained from wood cellulose. It is us...	State of disturbance	Tool used to sharpen pencils
5	6	Ability to control the fear	Flat material made from wood used for writing	Worry about the future	Little object which allow to sharpen a pencil
6	7	Ability to control fear and to be willing to d...	a short piece of writing on a particular subje...	act of understanding something, or the way tha...	tool for making something sharper

- Ho scelto di utilizzare due approcci al calcolo della similarità: uno basato su una funzione di libreria (SequenceMatcher) e uno costruendo io stesso una funzione per il calcolo della similarità.

- In questo secondo caso ho scelto un approccio bag of words e di filtrare le stopwords come fase di pre-processing, per concentrarsi sui termini più salienti.
- Ho costruito delle matrici di similarità fra le definizioni di ogni concetto.
- A partire da tali matrici ho calcolato la similarità generale per i due approcci facendo una media delle similarità escludendo la diagonale.

RISULTATI:

- Si è notato come, nel caso di termini concreti, la similarità sia più elevata di quanto non accada per i termini astratti. Questo è probabilmente dovuto alla possibilità di utilizzare degli attributi visivi per descrivere il termine.
- Nel caso dei termini astratti, invece, la mancanza di questi attributi concreti porta a definizioni meno simili fra di loro.
- I valori di similarità ottenuti con i due approcci sono quantitativamente diversi ma in entrambi i casi si può evincere questa tendenza:

	similar	my_similarity
Paper	0.395611	0.166860
Courage	0.398625	0.137788
Apprehension	0.316237	0.087922
Sharpener	0.520719	0.332071

Esercizio 1.2
Gruppo: Roberto Demaria

CONSEGNA:

- Dare una spiegazione dei risultati ottenuti nell'esercizio precedente

SVOLGIMENTO:

- Ho scelto di utilizzare due approcci: il primo approccio usando TF-IDF per vedere se vi fosse un nesso con la rilevanza statistica delle parole mentre il secondo approccio basato semplicemente sulla frequenza delle parole utilizzate nelle definizioni.

RISULTATI:

- Nel primo caso ho calcolato TF-IDF delle parole e li ho sommati per ogni definizione ottenendo una lista di 28 elementi (pari al numero di definizioni date). Dopodiché ho nuovamente sommato questi elementi ottenendo il seguente risultato:

	sum tf_idf	my_similarity
Paper	72.70662	0.166860
Courage	77.31542	0.137788
Apprehension	67.99832	0.087922
Sharpener	67.98177	0.332071

- Quello che si può evincere sembra essere una maggiore rilevanza statistica delle parole per i termini generici rispetto a quelli specifici e di quelli astratti rispetto a quelli concreti. Questo ci dice che i termini generici tendono ad essere meno omogenei nelle definizioni mentre i termini specifici utilizzano più spesso le stesse parole: in particolare sharpener che ha la maggiore similarità coincide anche con il minor valore di tf_idf rivelando che le definizioni sono molto simili avendo gli stessi termini più o meno distribuiti in tutte le definizioni.
- Nel secondo approccio ho fatto alcune analisi sulle frequenze delle parole andando a contare la frequenza totale di ogni parola e poi discriminando quelle con frequenza maggiore di 1 ottenendo diverse misure quantitative da confrontare con la similarità.
  Ho ottenuto i seguenti risultati:

Legenda:
S —> somma delle frequenze maggiori di 1
N —> numero di parole con frequenza maggiore di 1
R —> rapporto S/N

	S	N	R	my_similarity
Paper	97	20	4.85	0.166860
Courage	99	25	3.96	0.137788
Apprehension	76	20	3.80	0.087922
Sharpener	104	18	5.78	0.332071

- I termini concreti presentano un valore di R significativamente superiore rispetto a quelli astratti. I valori di R rispecchiano inoltre l'andamento della similarità.