Relazione esercitazione - Content2Form

Il problema affrontato in questa esercitazione è stato quello della ricerca onomasiologica, ovvero partendo da un insieme di definizioni l'obiettivo è stato proprio quello di individuare quale fosse il concetto migliore associabile alla lista di definizioni iniziale. Per indirizzare la ricerca è stato utilizzato il principio del "**genus**". Quello appena detto è stato fatto per 4 concetti: **Emotion, Person, Revenge e Brick**.

Il programma sviluppato per risolvere il problema detto pocanzi esegue i seguenti passi:

1) Per ogni concetto considera tutte le sue definizioni presenti nel dataset chiamato "definizioni.xlsx". Su queste definizioni è stata eseguita **una fase di pre-processing** in cui sono stati rimossi i segni di punteggiatura, **eliminate le stop words, eliminate eventuali stringhe vuote, eliminate eventuali circolarità nelle definizioni di ogni concetto e infine è stata applicata la lemmatizzazione** ad ogni parola rimanente per ogni frase.

2) Dopodichè tramite la funzione chiamata "**n_parole_piu_frequenti**" sono state individuate le n parole più frequenti nelle definizioni associate ad un certo concetto. In questo caso il valore migliore trovato per n è stato n=45.

3) L'assunzione che viene fatta è che probabilmente tra queste n parole più frequenti ci sarà/saranno il/i genus del concetto migliore che vogliamo ottenere per assegnarlo all'insieme di definizioni considerate o comunque potrebbe anche capitare che proprio una delle n parole più frequenti possa essere associata proprio al concetto migliore. Per queste ragioni quello che fa a questo punto l'algoritmo è questo:

   3.1) Per ogni genus trovato, viene utilizzata una mia implementazione dell'algoritmo di Lesk (chiamata "**The_Lesk_Algorithm_score**") che si preoccuperà di prendere in input in un certo momento, un certo synset di Wordnet associato al genus che si sta considerando in quel momento e l'insieme di tutte le definizioni associate al concetto che si vuole trovare in quel momento. Tale algoritmo restituirà uno score che sostanzialmente sarà un valore di **overlap tra la signature e il context**, in cui in questo caso:

   Context -> sarà l'insieme di tutte le parole presenti nelle definizioni associate al concetto corrente.

   Signature -> sarà invece, un insieme che conterrà: tutti i lemmi del synset corrente, la definizione del synset corrente e tutti gli esempi (qualora siano presenti) sempre del synset corrente. (Chiaramente sulla definizione e sugli esempi è stata sempre prima eseguita una fase di pre-processing esattamente uguale a quella citata al passo 1).

   3.2) Sempre per ogni synset associato al genus trovato vengono identificati subito dopo tutti i suoi iponimi e su ciascuno di questi vengono rieseguiti i passi citati a 3.1).

Ogni score restituito da "The_Lesk_Algorithm_score" sia al passo 3.1) che al passo 3.2) viene memorizzato in un dizionario. Tale dizionario alla fine dei due passi verrà ordinato in maniera decrescente e dopodichè verranno selezionati solamente i primi 5 synsets che saranno sostanzialmente quelli probabilmente più corretti per l'insieme di definizioni iniziali. Infine, i 5 synsets migliori per ogni insieme di frasi vengono stampati dal programma insieme alle proprie definizioni.

Risultati ottenuti:

```
I primi 5 synsets per il concetto Emotion sono i seguenti:

- Synset(feeling.n.01): the experiencing of affective and emotional states

- Synset(feeling.n.04): a physical sensation that you experience

- Synset(human.a.03): having human form or attributes as opposed to those of animals or divine beings

- Synset(impression.n.01): a vague idea in which some confidence is placed

- Synset(bad.s.03): feeling physical discomfort or pain (`tough' is occasionally used colloquially for `bad')
```

Il concetto Emotion possiamo dire che è stato individuato abbastanza bene dall'algoritmo in quanto sia il primo che il secondo synset restituito sembra essere abbastanza vicino al concetto di "Emozione".

```
I primi 5 synsets per il concetto Person sono i seguenti:

- Synset(human.a.03): having human form or attributes as opposed to those of animals or divine beings

- Synset(human.a.01): characteristic of humanity

- Synset(human.a.02): relating to a person

- Synset(homo.n.02): any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage

- Synset(world.n.08): all of the living human inhabitants of the earth
```

Per il concetto di Person possiamo notare come in questo caso, anche se l'algoritmo non sia riuscito comunque a trovare il concetto più appropriato in assoluto, ovvero person.n.01, comunque i synsets presenti nelle prime posizioni sembrano comunque essere in linea con quello che è il concetto di "Persona".

```
I primi 5 synsets per il concetto Revenge sono i seguenti:

- Synset(rusher.n.03): a person who rushes; someone in a hurry; someone who acts precipitously

- Synset(bad.a.01): having undesirable or negative qualities

- Synset(lover.n.01): a person who loves someone or is loved by someone

- Synset(damage.n.03): the act of damaging something or someone

- Synset(action.n.09): an act by a government body or supranational organization
```

Per quanto riguarda il concetto di Revenge invece, l'algoritmo sembra che abbia fatto un po' più fatica nel trovare i concetti giusti più vicini ad esso. Nonostante tutto però, il synset chiamato **damage.n.03**, presente in quarta posizione, sembra comunque essere abbastanza vicino al concetto di vendetta.

```
I primi 5 synsets per il concetto Brick sono i seguenti:

- Synset(material.a.03): directly relevant to a matter especially a law case

- Synset(use.v.06): habitually do something (use only in the past tense)

- Synset(brick.n.01): rectangular block of clay baked by the sun or in a kiln; used as a building or paving material

- Synset(coloring_material.n.01): any material used for its color

- Synset(thickening.n.01): any material used to thicken
```

Infine, per quanto riguarda il concetto Brick, i risultati mostrano che il synset più corretto in assoluto, ovvero **brick.n.01**, è stato individuato dall'algoritmo ed è stato posizionato in terza posizione.