Università degli Studi di Torino

Dipartimento di Informatica

Corsi di Laurea Magistrale in Informatica

Relazione esercitazioni TLN (prof. Di Caro)

Salvatore Coluccia
a.a. 2019/2020

Esercitazione 1

Questa esercitazione prevede il calcolo di similarità tra le definizioni date per 4 concetti mostrati a lezione.

L'obiettivo del laboratorio è calcolare in qualche modo l'agreement nelle definizioni degli studenti e capire se ci sono concetti per i quali si nota una netta differenza nelle definizioni.

L'algoritmo che ho utilizzato per calcolare queste percentuali è molto semplice e si divide in due macro-step:

1. Lemmatizzazione e rimozione delle stopword
2. Calcolo similarità tra tutte le definizioni di un dato concetto

Per la fase 1 ho utilizzato le funzioni di nltk che mette già a disposizione un elenco di stopwords inglesi che ho semplicemente rimosso dalle frasi che sono state precedentemente tokenizzate sempre utilizzando nltk. Sempre tramite la stessa libreria ho potuto procedere alla lemmatizzazione dei token trovati.

Per la fase 2 ho calcolato l'overlap medio dei lemmi delle definizioni di ogni concetto.

I risultati ottenuti sono i seguenti:

	ASTRATTO	CONCRETO
GENERICO	8.2% (Freedom)	24.6% (Building)
SPECIFICO	8.47% (Compassion)	15.8% (Molecule)

Dai risultati, per quanto chiaramente non possano essere dei dati scientificamente rilevanti dato il campione molto piccolo di esempi, si evince un agreement nettamente superiore per i concetti di tipo Concreto.
Intuitivamente era in effetti quello che mi aspettavo in quanto descrivere un concetto concreto risulta generalmente più semplice di un concetto più astratto.

Esercitazione 2

In questa esercitazione l'obiettivo è quello di identificare il concetto a partire da un insieme di definizioni.

Gli step che ho seguito per completare questo risultato sono i seguenti (ripetuti per ogni concetto):

1. Identificazione dei token, lemmatizzazione e rimozione stopwords da ogni definizione
2. Calcolo frequenza token considerando tutte le definizioni
3. Trovo contesto utilizzando examples,definitions, iperonimi e iponimi di ogni synset associato ad un termine (il synset associato ad ogni termine lo trovo applicando l'algoritmo di Lesk e usando i token trovati nelle definizioni come contesto)
4. Genero una lista con tutti gli iponimi di ogni iperonimo di ogni synset. Durante la creazione di questa lista considero iperonimi i termini che hanno una frequenza maggiore di una certa soglia (e quindi per questi conduco la ricerca sui loro iponimi)
5. Per ogni concetto ritorno l'iponimo che è più simile (in termini di overlap del suo contesto col contesto trovato al punto 3).

Questi sono i risultati ottenuti:

CORRETTO	SYNSET IDENTIFICATO
JUSTICE	human_right.n.01
PATIENCE	excitement.n.02
GREED	one.n.02

POLITICS	government.n.03
FOOD	plant.n.04
RADIATOR	dispersion.n.03
VEHICLE	airlift.n.01
SCREW	pin.n.09

Esercitazione 3

In questa esercitazione si vogliono sperimentare i principi della teoria di Hanks e quindi si cerca di calcolare delle frequenze di supersensi WordNet dato un corpus e scelto un verbo e una valenza.

Io ho scelto una valenza 2, usato il Brown Corpus e scelto il verbo build.

L'algoritmo esegue i seguenti macro step:

1. Estrae dal brown corpus le frasi che contengono il verbo TO BUILD
2. Da queste frasi estrae una lista di coppie che rappresentano i supersensi WordNet associati ai termini rispettivamente a sinistra e a destra del verbo in questione
3. Infine aggrega i risultati raggruppando per coppie uguali e visualizzandone la relativa frequenza

Questi sono i risultati ottenuti:

noun.possession___adj.all: 2
verb.possession___verb.change: 1
verb.stative___adj.all: 2
noun.group___noun.artifact: 1
noun.group___noun.attribute: 1
noun.group___noun.group: 2
adj.all___noun.artifact: 1
verb.emotion___noun.time: 2
adv.all___adj.pert: 1
noun.communication___noun.attribute: 1
verb.stative___verb.competition: 1
noun.group___adj.all: 1
verb.possession___noun.phenomenon: 1
noun.possession___noun.group: 1
verb.contact___noun.phenomenon: 1
noun.process___adj.all: 1
noun.person___adj.all: 1
noun.object___noun.relation: 1
noun.attribute___adj.all: 1
noun.act___noun.attribute: 2
noun.body___adj.all: 1
verb.cognition___noun.artifact: 2
noun.time___noun.artifact: 1
noun.artifact___noun.state: 1
adj.all___adj.all: 3
noun.substance___adj.all: 1
noun.state___noun.person: 1
noun.state___noun.artifact: 1
noun.state___noun.cognition: 1
verb.possession___adj.all: 1
verb.social___noun.person: 1
adv.all___adj.all: 1
adj.all___noun.group: 1
noun.state___noun.state: 1
verb.cognition___noun.quantity: 1
noun.person___noun.artifact: 2
noun.object___adj.all: 1
noun.person___noun.communication: 1
noun.state___noun.quantity: 1
noun.time___noun.relation: 1
noun.substance___adj.pert: 1
noun.substance___verb.social: 1
noun.artifact___adv.all: 1
noun.state___noun.attribute: 1
adv.all___verb.contact: 1
noun.person___verb.stative: 1
adj.all___adv.all: 1
noun.location___verb.stative: 1
noun.substance___noun.time: 1
noun.attribute___noun.process: 1
adv.all___noun.quantity: 1
noun.event___noun.artifact: 1
noun.group___noun.location: 1
noun.act___noun.artifact: 1
adj.all___verb.communication: 1
noun.substance___noun.artifact: 1
verb.change___noun.substance: 1
adv.all___noun.cognition: 1
adj.all___noun.person: 1
noun.state___adv.all: 1

	verb.perception___noun.quantity: 1 **noun.event___verb.cognition: 1** **noun.time___noun.cognition: 1** **adv.all___noun.phenomenon: 1** **noun.artifact___noun.artifact: 1**

Esercitazione 4

In questa esercitazione l'obiettivo è fornire un'implementazione di un algoritmo di segmentazione valido per un testo qualsiasi in modo da individuare i "cambi di contesto" in modo soddisfacente.

Il mio algoritmo esegue i seguenti step (sull'approccio text tiling):

1. Trovo le frasi e i relativi token
2. Lemming ed eliminazione stop words
3. Suddivido le frasi in un numero di finestre iniziali predefinito
4. Per ogni finestra calcolo la coesione media delle coppie di frasi in modo sequenziale utilizzando la wuPalmer similarity tra i synset wordnet associati ai token delle frasi e prendendo per ogni coppia di termini la max(similarity)
5. Se la coesione media è più bassa di una certa soglia allora divido la window in due sub-windows sulla base del break point individuato (la coppia di frasi meno simili). Inoltre diminuisco la soglia minima di un fattore pari a 0.7 per evitare "overfitting" e ridurre al minimo le window con solo una frase
6. Ripeto gli step 4 e 5 fino a quando non ci sono più split oppure arrivo al numero massimo di iterazioni

I risultati ottenuti possono essere considerati un primo risultato che sicuramente è migliorabile ma che comunque riesce a suddividere in modo un po' grossolano i vari contesti.

Le performance migliori sono state ottenute con un numero iniziale di finestre pari a 5 e con una soglia minima di coesione per una window pari a 0.6.

Il testo utilizzato per i test è stato creato a partire da diversi testi relativi a diverse news in modo da poter debuggare con meno ambiguità l'algoritmo.

Esercitazione 5

In questa esercitazione ho implementato un semplice esempio di applicazione della tecnica dell'Open Information Extraction su un testo preso da Wikipedia.

Per l'implementazione ho utilizzato la libreria Spacy eseguendo questi step:

1. Ho diviso il testo in frasi
2. Per ogni frase ho identificato il soggetto, il verbo e l'oggetto iterando sui token della frase e sfruttando l'informazione relativa al POS e alla dipendenza nell'albero delle dipendenze.
3. Ho infine installato e testato sullo stesso testo i servizi (sicuramente più professionali) di StanfordOpenIE sfruttando un wrapper Python disponibile online (https://github.com/philipperemy/Stanford-OpenIE-Python)

I risultati ottenuti permettono una prima divisione non sempre accurata della frase in triplette anche se non sono sicuramente paragonabili ad un approccio più complesso e strutturato che è messo a disposizione dalle librerie di Stanford OpenIE.

======➔ MIA IMPLEMENTAZIONE

Subject: empire world war | Verb: aimed dominate said begun | Obj: asia pacific

Subject: germany | Verb: conquered controlled formed | Obj: axis alliance

Subject: germany union | Verb: partitioned | Obj: territories

Subject: war | Verb: continued including running | Obj:

Subject: axis powers | Verb: launched opening trapped | Obj: invasion largest land theatre , trapped major

Subject: japan | Verb: attacked conquered | Obj: united states territories

======➔ CON STANFORD OPEN IE

|- {'subject': 'Empire', 'relation': 'was at', 'object': 'already war with Republic of China in 1937'}

|- {'subject': 'Germany', 'relation': 'conquered From', 'object': 'late 1939 to early 1941'}

|- {'subject': 'Germany', 'relation': 'formed Axis alliance with', 'object': 'Italy'}

|- {'subject': 'Germany', 'relation': 'conquered much From', 'object': 'late 1939 to early 1941'}

|- {'subject': 'Germany', 'relation': 'formed', 'object': 'Axis alliance'}

|- {'subject': 'their', 'relation': 'neighbours', 'object': 'Poland'}

|- {'subject': 'war', 'relation': 'continued', 'object': 'aerial Battle'}

|- {'subject': 'Britain', 'relation': 'of Battle is', 'object': 'Balkan Campaign'}

|- {'subject': 'war', 'relation': 'continued', 'object': 'aerial Battle of Britain'}

|- {'subject': 'war', 'relation': 'continued', 'object': 'Battle of Britain'}

|- {'subject': 'war', 'relation': 'continued Battle between', 'object': 'Axis powers'}

|- {'subject': 'war', 'relation': 'continued between', 'object': 'Axis powers'}

|- {'subject': 'war', 'relation': 'continued with', 'object': 'campaigns including North Africa'}

|- {'subject': 'war', 'relation': 'continued Battle with', 'object': 'campaigns'}

|- {'subject': 'war', 'relation': 'continued with', 'object': 'campaigns'}

|- {'subject': 'war', 'relation': 'continued Battle with', 'object': 'campaigns including North Africa'}

|- {'subject': 'war', 'relation': 'continued', 'object': 'Battle'}

|- {'subject': 'war', 'relation': 'continued', 'object': 'Balkan Campaign'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'land theatre of war in history'}

|- {'subject': 'Axis powers', 'relation': 'launched', 'object': 'invasion'}

|- {'subject': 'Axis powers', 'relation': 'launched invasion In', 'object': 'June 1941'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'land theatre in history'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'largest land theatre of war in history'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'largest land theatre of war'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'largest land theatre'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'largest land theatre of war'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'largest land theatre in history'}

|- {'subject': 'European Axis powers', 'relation': 'launched', 'object': 'invasion'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'largest land theatre of war in history'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'largest land theatre'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'largest land theatre in history'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'land theatre of war in history'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'land theatre of war'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'land theatre in history'}

|- {'subject': 'largest land theatre', 'relation': 'is in', 'object': 'history'}

|- {'subject': 'European Axis powers', 'relation': 'opening', 'object': 'land theatre'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'land theatre'}

|- {'subject': 'Axis powers', 'relation': 'opening', 'object': 'land theatre of war'}

|- {'subject': 'European Axis powers', 'relation': 'launched invasion In', 'object': 'June 1941'}

|- {'subject': 'Japan', 'relation': 'attacked territories In', 'object': 'December 1941'}

|- {'subject': 'Japan', 'relation': 'conquered', 'object': 'much of Western Pacific'}

|- {'subject': 'Japan', 'relation': 'attacked', 'object': 'European territories'}

|- {'subject': 'Japan', 'relation': 'attacked', 'object': 'territories'}

|- {'subject': 'Japan', 'relation': 'quickly conquered', 'object': 'much'}

|- {'subject': 'Japan', 'relation': 'attacked', 'object': 'United States'}

|- {'subject': 'Japan', 'relation': 'attacked territories in', 'object': 'Pacific Ocean'}

|- {'subject': 'Japan', 'relation': 'attacked United States in', 'object': 'Pacific Ocean'}

|- {'subject': 'Japan', 'relation': 'quickly conquered', 'object': 'much of Western Pacific'}

|- {'subject': 'Japan', 'relation': 'attacked United States In', 'object': 'December 1941'}

|- {'subject': 'Japan', 'relation': 'conquered', 'object': 'much'}