Università degli Studi di Torino

Corso di Laurea Magistrale in Informatica

Tecnologie del Linguaggio Naturale

Docente: Luigi Di Caro

**Relazione esercitazioni**

Alessandro Clocchiatti

Matricola 909105

Questa relazione riguarda le esercitazioni della terza parte del corso e comprende:

1. Definition similarity
2. Content-to-form
3. Semantic clustering (Hanks) TODO capire se è il nome giusto
4. Text segmentation
5. OIE system

**1. Esercitazione: Definition similarity**

**1.1 Consegna**

L’esercitazione prevede i seguenti passaggi:

1. Caricamento dei dati sulle definizioni (file *definizioni.xls* o documento Google presente su Moodle);
2. Preprocessing (su frequenza minima dei termini, stemming, etc. a vostra scelta);
3. Calcolo similarità tra definizioni (cardinalità dell’intersezione dei termini normalizzata su lunghezza minima tra le due, o varianti a scelta);
4. Aggregazione sulle due dimensioni (concretezza / specificità come da schema in basso);
5. Interpretazione dei risultati e scrittura di un piccolo report (da inserire nel vostro portfolio per l’esame).

|  |  |  |
| --- | --- | --- |
|  | Astratto | Concreto |
| Generico | % | % |
| Specifico | % | % |

**1.2 Svolgimento**

asdas

**2. Esercitazione: Definition similarity**

**2.1 Consegna**

L’esercitazione prevede i seguenti passaggi:

1. Caricamento dei dati content-to-form (presente su Moodle);
2. Preprocessing (si veda esercitazione precedente, a vostra scelta);
3. Utilizzo di WordNet come sense inventory, per inferire il concetto descritto dalle diverse definizioni;
4. Definire ed implementare un algoritmo (efficace ma anche efficiente) di esplorazione dei sensi di WordNet, usando concetti di similarità (tra gloss e definizioni, esempi d’uso, rappresentazioni vettoriali, etc.);

* Suggerimento A: sfruttare principi del genus-differentia;
* Suggerimento B: sfruttare tassonomia WordNet nell’esplorazione;
* Suggerimento C: pensare a meccanismi di backtracking.

**2.2 Svolgimento**

asdas

**3. Esercitazione: Semantic clustering (P. Hanks)**

**3.1 Consegna**

L’esercitazione prevede l’implementazione della teoria di P. Hanks:

1. Scegliere un verbo transitivo (minimo valenza = 2);
2. Recuperare da un corpus n istanze in cui esso viene usato;
3. Effettuare parsing e disambiguazione
4. Usare i super sensi di WordNet sugli argomenti (subj e obj) del verbo scelto;
5. Aggregare i risultati, calcolare le frequenze, stampare i cluster semantici ottenuti.
   1. Un cluster semantico è inteso come combinazione dei semantic types (ad esempio coppie di sem\_types se valenza = 2)

**3.2 Svolgimento**

<https://spacy.io/usage/linguistic-features>

<https://stackoverflow.com/questions/36610179/how-to-get-the-dependency-tree-with-spacy>

<https://spacy.io/usage/visualizers>

<https://spacy.io/api/token#attributes>

<https://spacy.io/api/annotation>

<https://stackoverflow.com/questions/39323325/can-i-find-subject-from-spacy-dependency-tree-using-nltk-in-python>

**4. Esercitazione: Text segmentation**

**4.1 Consegna**

L’esercitazione prevede l’implementazione, ispirandosi al Text Tiling, di un algoritmo di segmentazione del testo:

1. Usando informazioni come frequenze (globali, locali), co-occorrenze, risorse semantiche (WordNet, etc.), applicando step di preprocessing (as usual), etc.

* La scelta del testo è a discrezione dello studente.

**4.2 Svolgimento**

**SITOGRAFIA**

[1] <https://en.wikipedia.org/wiki/Pearson_correlation_coefficient>

[2] <https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>

[3] <http://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html#WordNetCorpusReader.synset_from_sense_key>

[4] <https://github.com/ptorrestr/py_babelnet>

[5] <https://babelnet.org/guide>