**Web Information Retrieval** **Spring 2018**

# WIR Project Proposal

**Members**: *Angelo Catalani, Valerio Colitta, Alessandro Lo Presti*

## 1.1 Proposal

The paper describes a new approach to compute similarity between documents and queries. Not only does it take into account tf-idf, but also SIMILARITY among terms via a new formula.
Terms are identified through WordNet, where they are linked to each other in different ways. Given a particular synset (sense) for a term, you can traverse the whole network to find related terms (superclasses, subclasses, etc).
The paper introduces a specific metric for quantifying the similarity of two words by measuring shortest path between their senses in the WordNet graph.
We implement a new metric, using the WordNet graph. In particular we establish that the similarity depends on the least common ancestor in the term-hypernym path (the least the better).

## 1.2 Outline

- **VSM** le caratteristiche e le limitazioni con la similarity.

- **GVSM**, introduzione di un nuovo modello vettoriale e di una estensione della vecchia cosine similarity che tiene conto della relazione semantica dei termini.

- **WORDNET**, e il fatto che il paper lo usi per trovare la similarity in base a delle metriche specificate ad-hoc : `SCM`, `SPE`, `SR`.

- **NOI**, che creiamo una nuova similarity measure basata su wordnet, che usa il least commont ancestor.

- **COMPARAZIONE**, Termine-Termine usando tre dataset. Document-Query usando NPL. Quest'ultimo, usando le metriche precision and recall graph.