ICS 3207: Knowledge Discovery and Management
Project 2016/17

Relationship Identification from Text

Introduction

The RelFinder[1] allows users to explore how things are related with each other? Given two user-specified objects (e.g. two persons), RelFinder extracts and visualizes relationships between them. RelFinder uses RDF data from DBpedia and other structured sources.

In this assignment you need to extract relationships and entities from Wikipedia pages and perform community detection to find interesting communities for these entities. You will also need to use some visualisation/s to display the entity-relationship graphs and the communities.

Tasks

You need to find documents (Wikipedia Pages) related to different persons of interest and extract relationships between these persons of interest and other named entities. Then you will need to combine the information into a single entity-relationship graph.

You need to consider the following 8 persons of interest:

- Joseph Muscat
- Konrad Mizzi
- Edward Scicluna
- Louis Grech
- Simon Busuttil
- Tonio Fenech
- Lawrence Gonzi
- Austin Gatt

For each person of interest, you need to:

- Find related Wikipedia Pages about that person of interest. You can use the Wikimedia API to extract:
    - The Wikipedia page about that person
    - The related Wikipedia pages - obtained through the outgoing links from the person's Wikipedia Page.
    - Full text search results for that person of interest as a query.

- Identify the Named Entities (Persons, Locations and Organisations) from these documents (using an NER tool - such as Stanford Core NLP), and perform co-reference resolution to identify how different string instances relate to the same real-life entity.

[1] http://www.visualdataweb.org/relfinder.php accessed October 2016

- Identify relationships between these Named Entities.
    - Relations may be identified using hand-written rules, using a semi-supervised approach; or using an unsupervised approach. Relations between entities are typically represented as entity relation triples.

    - Count the frequency which with different entity relation triples co-occur. Different entity relations triples may count as co-occurring only if they co-occur within the same document / paragraph / sentence / window (the size should be your justified choice).

- Represent these relationships as an entity-relationship graph using some visualization library.

After building these graphs around each person of interest, you need to:

- Perform coreference resolution on the different named entities across the different graphs;

- Combine the different graphs together into 1 single graph;

- Convert this graph into an undirected graph using moralization and ignore the relationship labels;

- Apply a community detection technique such as the Girvan/Newman or the Bron-Kerbosch (other techniques can also be considered) to identify possible communities within the entity-relationship graph;

- The communities will need to be displayed using suitable visualization.

Note that all these tasks need to be performed programmatically.

Useful APIs and Resources

- MediaWiki API - http://www.mediawiki.org/wiki/API:Main_page
- Stanford Core NLP - http://nlp.stanford.edu/software

Evaluation vs Findings

Please note that you are ***not*** expected to perform evaluation of your system. However, you should provide in your report a discussion on the main findings from your project, and your reasoning as to why these findings occur.

Deliverables

1. Program(s), solution files etc

2. Report (6 pages in the format of an application track scientific paper), should include
    a. Abstract (Max 1 page)
    b. Introduction
    c. Overview of relevant background research
    d. Description of method(s) used
    e. Findings & discussion on Findings
    f. Future work & conclusion.

3. A single page that states what parts of the project were fulfilled or not, and how the work was divided between the different members of the group.

A link to a template for the format of the paper will be supplied on the VLE.

Marking Criteria

- Software Artefact (50%):
    - Information Acquisition (5%)
    - Entity Identification (5%)
    - Relationship Identification (15%)
    - Community Detection (15%)
    - Visualisation (10%)

- Report (50 %):
    - Background research depth & coverage (10%)
    - Methodology Used & Justification (20%)
    - Findings and Related Discussion (5%)
    - Future Work & Conclusion (5%)
    - Overall report quality (10%)

References

- Sekine, S., “On-demand Information Extraction”, in Proceedings of the COLING/ACL on Main Conference Poster Sessions COLING-ACL ’06 pp. 731–738, Stroudsburg, PA, USA: Association for Computational Linguistics.

- Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A., “Unsupervised Named-entity Extraction from the Web: An Experimental Study”, in Artif. Intell. 165 (1), 91–134., ACM, 2005.

- Bach, Nguyen, and Sameer Badaskar. "A review of relation extraction." Literature review for Language and Statistics II (2007).

- Gonzalez, Edgar, and Jordi Turmo. "Unsupervised relation extraction by massive clustering." 2009 Ninth IEEE International Conference on Data Mining. IEEE, 2009.

- Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. 2009. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 661-670. DOI=http://dx.doi.org/10.1145/1526709.1526798

- Girvan, M. & Newman, M. E. (2002), 'Community structure in social and biological networks.', Proc Natl Acad Sci U S A 99 (12), 7821--7826.

- Coen Bron and Joep Kerbosch. 1973. Algorithm 457: finding all cliques of an undirected graph. Commun. ACM 16, 9 (September 1973), 575-577. DOI=http://dx.doi.org/10.1145/362342.362367

Groups

You should work in **groups of 2**.

You should include a page in the report that clearly states:
- What parts of the project were fulfilled or not
- How the work was divided between the different members of the group.

Note that the mark given to the individual group members may vary according to their contribution to this project.

Deadlines

Submission deadline: **Friday, 20th January 2017 1200**

A demo session will be scheduled after this deadline for each group to demonstrate the submitted work.

Contact

If you have queries please don't hesitate to contact any of:

- Joel Azzopardi - joel.azzopardi@um.edu.mt
- Chris Staff - chris.staff@um.edu.mt
- Charlie Abela - charlie.abela@um.edu.mt