Analytics for Question Answering:
Graphs & Clusterization

Abstract: *We aim at exploring a question answering dataset in terms of clusterization potential and inter-record relationships, by applying existing or self-implemented widespread algorithms for the former and building analysis graphs for the latter. Our final goal is to be able to compare the visual results with other machine interpreted patterns available (from our complementary project, built on top of the same database, destined for the "Natural Language Processing" subject).*

Dataset options:

- ***https://github.com/allenai/gooaq/tree/main?tab=readme-ov-file*** *(chosen)*
- ***https://github.com/google-research-datasets/natural-questions?tab=readme-ov-file*** *(alternative)*

Main objective: various configurations of graphs and clusterization; step-by-step objectives:

1. research on the HotpotQA dataset's structure, purpose and particularities, with the source article being published at: https://arxiv.org/abs/1809.09600
2. Graphs for exploratory analysis: question-answers (QA), question-paragraph (QP), question-answer-paragraph (QAP: paragraph = source), answer similarity (AS), motifs, connected components
3. Graphs for topic modelling: connections among questions-topics (TPQ), connections among questions-paragraphs grouped by topic (TPQP)
4. Graph for Named-Entity-Recognition (the NLP Entity Graph)
5. Clusterization on topic detection, resource-oriented (variations on the data being fed: questions only, questions and context) – includes ML contribution for topic detection – this is (fully) **unsupervised** clusterization (we try to guess the number of clusters)
6. Clusterization on Named-Entity-Recognition (variations on the data being fed: questions only, questions and context) – includes ML classification contribution for entity recognition – this is **unsupervised** clusterization
7. Optimization of the number of clusters – this is **semi-supervised** clusterization (we try to previously infer the number of clusters) - via recommendation systems – research level, we do not assume full implementation, but domain analysis
8. Further observed to be non-interesting: Clusterization over question lengths (implemented but not included in presentation)

9. Interpretations over graph analysis: interpret Many-to-Many, One-to-Many relationships
10. Compare **“real”** vs. **“interpreted”** (correlation with NLP project) NER graph
11. Graph-based clustering: community detection via Label Propagation
12. Graph-based clustering: community detection via self-implemented distributed Louvain
13. Page Rank over subgraph detached from the answer similarity graph and over subgraphs detached from the NER graph
14. Interpret clusterization results: identify logical grouping patterns and intra-cluster similarities

Contributions:

- ***Corina’s responsibilities: 1, 5, 6, 7, 8, 10, 12, 14***
- ***Remus’s responsibilities: 1, 2, 3, 4, 9, 10, 11, 13***

All components, with code references (in /home/ubuntu/jupyter):

- Intro (slides 1-10) - Description of the datasets used (hotpot_dev_distractor_v1.json and hotpot_train_v1.1.json)
- Graphs for Exploratory Data Analysis (Slides 11-22) – Simple Graphs.ipynb (details on the method in slides, results in the notebook and saved in .html and .txt files in /results/)
- PageRank and Label Propagation Algorithm for Community Detection (Slides 23-26) – In NER Graph.ipynb and Answer Similarity Graph.ipynb (results in the notebook and saved in /results/AS Subgraphs and /results/NER Subgraph dev/ as html)
- Louvain for Community Detection (Slides 27-33) – Louvain.ipynb (details on the method in slides, code in the notebook but ran locally, with results saved locally, uploaded to /results/Clusterization Results/Louvain and previewed in slides);

executed on subgraphs belonging to Answer Similarity Graphs and the NER Graphs (subgraphs saved in /results/Louvain)

- Topic Modelling Clusterization (Slides 34-39) – Topic Modelling Clustering.ipynb and the results in /results/Clusterization Results (details on the method in slides)
- NER Clusterization (Slides 40-46) – NER Clusterization.ipynb and the results in /results/Clusterization Results (details on the method in slides)
- Graphs for NER tasks (Slides 47-56) – NER Graph.ipynb with results saved in /results/NER Subgraphs dev (details on the method in slides)
- Answer Similarity Graph (Slides 57-62) – Answer Similarity Graph.ipynb with results saved in results/AS Subgraphs (details on the method in slides)
- QAP (Slides 63-64) – Topic Modelling Graphs.ipynb with results saved locally and uploaded to RaaS (placed in the archive in /results/ with TQP and TLPQ); details on Topic Modelling can be found within the Topic Modelling Clusterization section
- Slides 65-73 – Study and Further work on Recommendation System for meta-learning of hyperparameters of k-Means, RaaS handling of errors and optimization, Conclusions

Interpretation for the above methods and results can be also found within the corresponding slides.

Acknowledgement:

> *"Procesarea și analiza datelor a fost realizată cu sprijinul proiectului SMIS 124759 - RaaS-IS (Research as a Service Iasi), finanțat prin Programul Operațional Competitivitate".*