05_10_2022

Topic of the systematic review: Topic labeling

Question 1 - Focus of the gathered papers

Should we only gather papers where topic labeling is the **main focus** of the proposed research or should we also analyse those papers where the labeling phase represents as a **secondary step**?

This decision will influence the search/selection strategy specified in the review protocol.

Question 2 - Topic domain

Should we try to cover topic labeling research **as a whole** or should the domain of the analysed papers be limited to topic labeling in **software engineering** (like the systematic review of topic labelling from Silva et al.)?

In this context, choosing the broader route of covering topic labeling as a whole might be beneficial given that not all research that applies topic modelling techniques (LDA, LSI, pLSI, etc.) also applies labels to the generated topics.
For example, of the 111 SE papers covered by Silva et al., only 36 papers applied topic labeling techniques.

Possible answer

Retrieve papers where topic labeling is the **main focus** (Question 1) but **without limitations** on the covered **domain** (Question 2).
If needed, extend the search also to papers where topic labeling is a secondary step (Question 1).

Review Protocol (Methods)

Preparing the review protocol that will establish the guidelines to follow when conducting the review.
The content of the protocol will be described in the methods section of the final report.

0. Establishing a rationale for the review

In this context, the main motivator for this research would be the **lack of systematic reviews** exclusively focusing on topic labeling.
Normally, the topic is tangentially covered/mentioned in topic modeling reviews.

1. Coming up with an initial set of research questions (Objectives)

- **RQ1**: What are the different approaches for topic labeling, how are they used and in which context?
- **RQ2**: Which underlying topic modelling techniques have been used and how have they affected the choice of a topic labeling approach?
- **RQ3**: How are candidate labels ultimately selected and how is the quality of the final label assignments evaluated?
- **RQ4**: (Dependent on Question 2) Which are the prevalent domains on which topic labeling techniques have been applied?

2. Eligibility criteria

- **Time period**
  - How should we establish the time period we will gather papers from? (~3-5)
  - Silva et al. covers a range of **12 years**. The only reason given for this choice is as follows: "Limiting the search to the last twelve years allowed us to focus on more mature and recent works"
- **Journals/Publications restrictions**
  - Conferences: CORE ranking portal
    - Green The GII-GRIN-SCIE (GGS) Conference Rating - About the GGS Rating 2021
  - Journals: Scopus / DBLP / Google Scholar
    - ACM / IEEE
    - Scimago Journal & Country Rank
  - Depending on Question 2, this might include only **Software Engineering** venues (like in Silva et al.) or a broader set consisting also of **Information Retrieval** and **Natural Language Processing** ones.
- Language limitations
- Use of previous reviews

3. Information sources

Sources:

- Dependent on selected Conferences/Journals

4. Search strategy

Keywords

- **Primary**: topic label, topic label[l]ing
- **Secondary** (2021, Silva et al.): topic model[s], topic model[l]ing
- **Tertiary** (2016, Chen et al.): lsi, plsi, lda, latent dirichlet allocation, latent semantic

5. Selection process

Exclusion criteria

- The paper does not actively apply topic labeling techniques
- The paper is a systematic review (secondary/tertiary study)
- (Depending on Question 1) Topic labeling is not the main focus of the paper

#Thesis/Temporary notes#