**PROJECT WORKFLOW**

* CONSIDERED MODELS:
  + Unsupervised document embedding technique
    - Topic Modelling (collocation and distributional similarity[[1]](#footnote-1))
      * Text Classification
    - Information Retrieval
      * Sentiment Analysis
* RESEARCH FOCUS WITHIN SMART HOME:
  + Connectedness
  + Personalisation
  + Relationship
  + Privacy/Security
  + Trust
* TOPIC MODELLING WORKFLOW

|  |  |  |  |
| --- | --- | --- | --- |
| **1. PREPROCESS** | **2. TRAIN** | **3. SCORE** | **4. EVALUATE** |
| Setting the right window of context for the machine.  Context made of topics and words into a topic.  Input: documents related to research focus.  Output: documents and their word distribution, and vectorized words (word-ids). | Model: **L**atent **D**irichlet **A**llocation.  Tells what topics are present in any given document by observing all the words in it and producing a document distribution.  Input: output from phase 1.  Output: topics and words distribution within a topic. | Read text from links as bag-of-words (only understand the words used to train the model)  Input: output from phase 2  Output: topic distributions which explain best how this document could have been generated. | Does the model capture the right aspects of a document?  Text classification.  LDA simplex space where we can calculate distance within similar topic. |

* INFORMATION RETRIEVAL WORKFLOW
  + Automate:
    - extract field of interest in the free text within a topic
      * Name Entity Recognition and Relation Extraction
    - Unsupervised sentiment analysis
  + Not Automate:
    - Read the extracted documents
* QULITATIVE INTERVIEWS:
  + prepare interviews to reinforce the findings
  + labelling of extracted text

**CHALLENGES:**

**1.** **How to extract good quality of topics that are clear, segregated and meaningful**.

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>

<https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/>

<https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/>

<https://www.machinelearningplus.com/nlp/gensim-tutorial/>

<https://github.com/rbkhb/NLP_IMC>

1. Words that frequently appears in similar context are more likely to be semantically related. [↑](#footnote-ref-1)