NLP Hackathon

Problem Statement

Given a corpus of user stories available in the url https://data.mendeley.com/datasets/7zbk8zsd8y/1/files/adf93dac-6b07-492d-b7ec-09848652c713

Do the following:

1. Automatic generation user stories for different stakeholders /roles
    a. Role based Stories extraction can be done using regular expression
    b. The different stakeholders involved in the corpus are data manager, PI, IT manager, IT staff member, researcher, repository manager etc.
2. Identify the tasks for every stakeholder
    a. Data management
    b. Data store
    c. Analysis
    d. Cleaning
    e. privacy
3. Visualization of frequently occurred words based on stakeholders
4. Check the syntax of user stories using any of the parsers
5. Do the above tasks using the following NLP tasks
    a. Language models
    b. POS tagging
    c. NER
    d. Topic modelling
    e. Clustering
6. Analyse and evaluate the models by varying the following
    a. Word representation (tf-idf, word2vec)
    b. Deep learning models
    c. Machine learning models
    d. Statistical models
    e. Semantic Similarity measures (Cosine similarity , PPMI (term and context or term vs stakeholders))