computational social media

assignment #3

29.04.2022

idiap RESEARCH INSTITUTE

EPFL

The assignment:
hands-on exercise with topic modeling

Goals

1. You are asked to apply Latent Dirichlet Allocation (LDA) to discover and display topics on Tweet data.

2. Through the exercise, you will
    (a) Use basic text pre-processing techniques
    (b) Use LDA and reflect upon the topics that can be discovered from the Tweet dataset

Instructions

1. Extract the text of the Tweets (in English) that you collected in assignment #2 and download the readme file from the link provided in Moodle.

2. Preprocessing: You can do pre-processing as follows:
    a. Convert text into lowercase. **Hint**: Use function lower() from string
    b. Remove non-alphabetic characters (e.g., !@#$%^&*()). **Hint**: Use re.sub with expression to filter out.
    c. Break sentences into words. **Hint**: Use function split() to token.
    d. Remove short words (length strictly less than 3 characters) (e.g. a, of, to). **Hint**: Use function len(word).
    e. Lemmatization: convert a word to its base form e.g. car, cars and car's to car. **Hint**: Use WordNetLemmatizer from nltk.stem .
    f. Remove common English words (e.g. the, for, etc.). **Hint**: Use stopwords from nltk.corpus to remove common English words: stopwords.words('english')

Instructions

3. Topic modeling: Use a python implementation of LDA to extract K topics, using the parameters suggested in the readme file. Focus on the top words per topic (e.g., 5 or so top words) to understand what the topics are about.
    a. Remove words that occur in less than 10 documents, and words that occur in more than 90% of the documents.
    b. Build a dictionary.
    c. Transform each document to a vectorized form by computing the TF/IDF of each word.
    d. Apply LDA with different K values and find the optimal number of topics based on the topic coherence measure (see the readme file for the definition.) Plot a curve showing topic coherence vs number of topics.
    e. Examine the topics found by the best model, with your own proposal for the “name” of each topic that is meaningful to you. Display your results in the notebook.

Instructions

4. To aggregate the class results, fill in the questionnaire on Moodle:
    - What was the optimal number of topics?
    - Pick 3 topics that were especially meaningful to you and introduce this in the questionnaire: (1) semantic 'name' you gave to the topic; and (2) top 5 kwords per topic.
    - Discuss the results in the questionnaire. Are the extracted topics expected? Were any topics surprising? What are the limitations of LDA?

5. Submit your Jupyter Notebook via Moodle including code, results & comments.

Logistics and deadline

1. In case of questions, contact Lakmal by email.

2. Deadline to submit assignment: **Mon 16.05.2022, 7pm**
   - please submit your assignment even if it is not complete
   - late assignments will not be given any credit

questions?

daniel.gatica-perez@epfl.ch