COMP 7970

Special Topics: Natural Language Processing

Instructor: Shubhra ("Santu") Karmaker

Assignment #1: Latent Dirichlet Allocation Implementation [100 points]

⚠ **Notice:** This assignment is due **Monday, September 6, 2021 at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

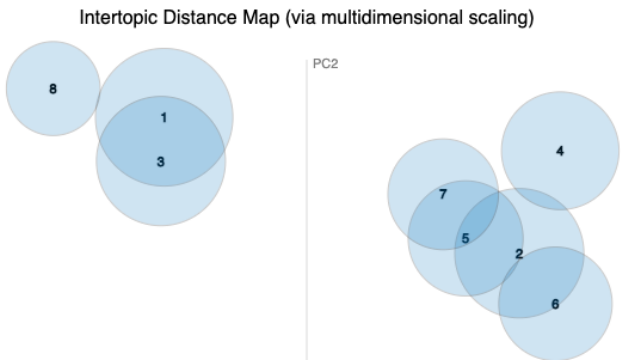
1. Implementation of LDA [40 pts]

LDA (short for Latent Dirichlet Allocation) is an unsupervised machine-learning model that takes documents as input and finds topics as output. The model also says in what percentage each document talks about each topic. Now, let's implement the LDA model. You may use any programming language you would like. [If you are using python, feel free to explore *Gensim* library]

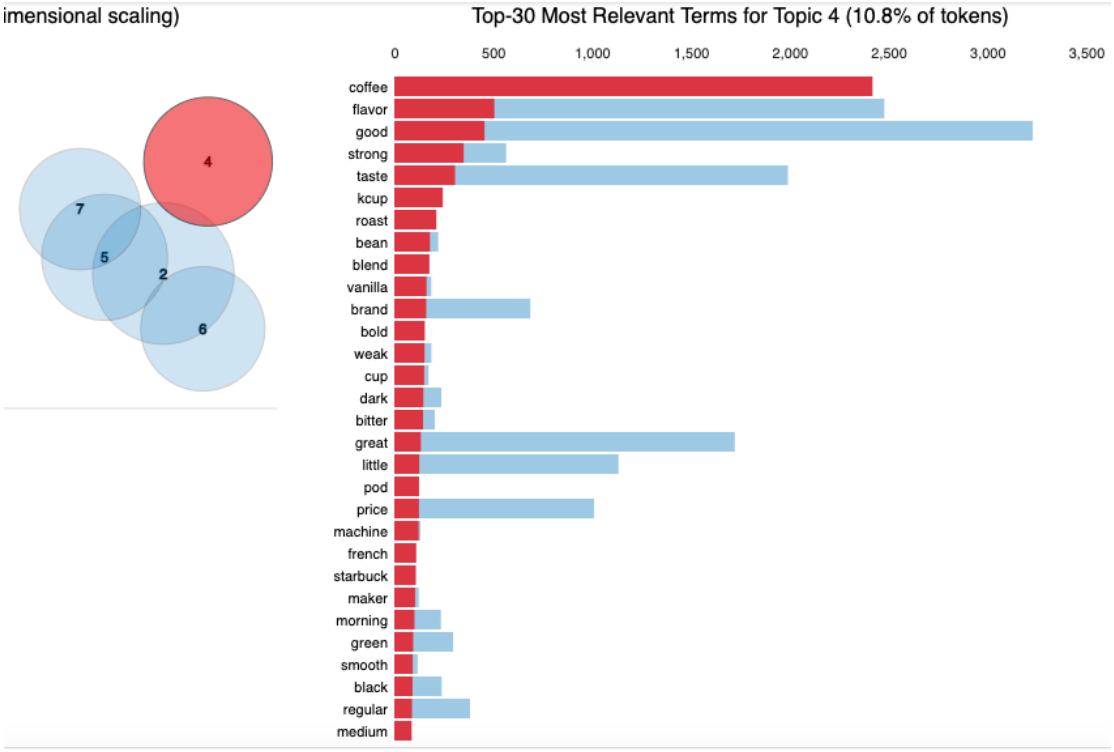
a. Dataset: Uploaded on the Canvas
b. Before implementing LDA do the following steps
    i. Remove stop words.
    ii. Lemmatize the reviews.
    iii. Create document term matrix [if using *Gensim* library].
c. Consider the number of topics is equal 10, alpha=[0.05] eta=[0.05].
d. Print the topics along with top 30 high-probability words for each topic.

2. Visualization Of Topics [40 pts]

a. Once you implemented LDA model and identified topics, create a visualization to check if the topics are overlapping or dispersed. For an example, here is the visualization using *pyLDAvis* which shows the inter-topic distance. Include the visualization in your report.



b. The visualization also shows distribution of words for each topic as below, consider Top-30 most relevant word for each topic. Include these visualizations in your report.



c. Vary the number of topics as [5, 10, 15, 25, 50] and report optimum number of topics based on below assessment
   i. Are your topics interpretable?
   ii. Are your topics unique? (two different topics have different words)
   iii. Are your topics exhaustive? (are all your documents well represented by these topics?)

d. Tweak *alpha* and *eta* to adjust your topics. Start with 0.05 and try other values such as [0.05, 0.10, 0.15, 0.20, 0.25] . Report optimum value of *alpha* and *eta*, based on below assessment
   i. Are your topics interpretable?
   ii. Are your topics unique? (two different topics have different words)
   iii. Are your topics exhaustive? (are all your documents well represented by these topics?)

From step c and d conclude the optimum number of topics, alpha, eta and state reasons behind your conclusion.

3. Implement KL-divergence [20 pts]

KL divergence is a way of measuring the distance between two distributions. For study refer to: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

a. Implement KL Divergence to verify if you notice the similar behavior between the topics identified. For example, if you observe two topics are overlapping, does KL Divergence score indeed capture this? Please note, KL-divergence is not symmetric. Therefore, while calculating KL-divergence between two topics, consider taking average between two values, i.e. $D_{KL}(P||Q)$ and $D_{KL}(P||Q)$.
b. To support your answer, create a visualization and show whether topics are indeed overlapping / dispersed. You can create similar visualization as in question 2 and show inter-topic distances.

Dataset

Your dataset is updated on the canvas. This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews. Reviews include product and user information, ratings, and a plain text review. We have sampled random 10000 short reviews [length between 20-100] from the main dataset.