**Instructions**

For the final assessment of the Web Scraping and Text Mining workshop, you will analyse the content of the 1975 House of Commons debate over the renegotiation of the terms of the British entry into the European Community, which preceded the Referendum on the European Community (Common Market) later that year. The House debated for three days (7th, 8th and 9th of April, 1975), which was followed by a vote on whether to continue in the Common Market under the new renegotiated terms. This question was nonetheless put before the electorate in the form of a referendum.

Using the accompanying R script for this exercise, you will need to scrape the Historical Hansard records of the European Community Membership debate for the 7th, 8th and 9th of April House of Commons sittings, retrieve the speeches by each MP who spoke, aggregate them over the length of the debate (three days), and run an unsupervised method to retrieve the main themes/topics of the debate.

The R script is partially completed. You’ll need to read and understand the code, and complete the sections that lack information:

1. Scraping and parsing the content of the debate over the three days has been written into a loop; the code is complete, but I encourage you to break it down and understand how it operates;
2. You will need some auxiliary information to assess the output of the text analysis. As such, you will scrape a table in a Wikipedia page that gives you the party affiliation for each MP elected in the October 1974 election. You will need to complete the R code by providing the “Xpath” selector for the relevant table giving party affiliation for each MP. Use the “inspect” option in your browser to retrieve the Xpath; alternatively, install the Inspector Gadget extension.
3. After scraping the table, you will need to merge the content of the table with the object containing the speeches data using the first and last names for the MPs. Use regular expressions to harmonise the name strings as they appear in both databases. This step can also be done manually for each MP.
4. Finally, run the standard LDA topic model using quanteda and structural topic modeling using the STM package, conditioning the “content” of part of the model over the party affiliation. The goal is to understand whether the top words for Labour and Conservative MPs differ when talking about the same issue. You will need to write your own code for this step.
5. Finally, justify your choice for the number of topics, produce the top 30 words for each topic, and analyse the results.
6. Save your R script, results and analysis into a word document, and email it to me at p.matos@ub.edu