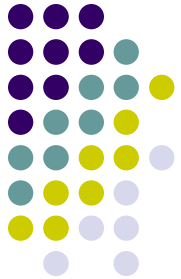
Big Data Analytics

Course Assignment

UNIVERSITÄT LUZERN

Deadline: 23 April 2021



FIRST PART

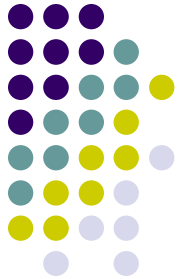
1) Create in Quanteda a corpus starting from the folder that includes a sample of UK party manifestoes since 2010

- ✓ Analyze such corpus via Wordfish
- ✓ Give a substantial interpretation of the latent scale extracted
- ✓ Comments the result (also by taking a look at Wikipedia! For example: what did happen after the 2005 UK elections? What about the movements of parties? Do they make sense to you?)

Some suggestions: a) I would keep in the Dfm only features with a number of characters>1; b) If you have in your corpus this symbol “â” you can get rid of it (before creating the corpus) using the gsub command:

```
gsub("[\u00E2]","",myText$text)
```

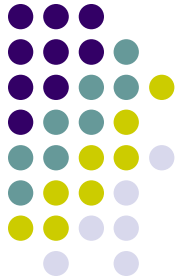
Deadline: 23 April 2021



2) Analyze such corpus via Wordscores, using as reference texts the 2015 party programs along two different dimensions. Economic dimension: CONS=7.85; Lab=3.85; Lib=5.14; UKIP=8.57 (such scores refer to a left-right economic scale). EU dimension (higher score implies being more pro-EU): CONS=3.14; Lab=5.57; Lib=6.71; UKIP=1.14 [source of parties' scores: 2014 Chapel Hill expert survey]

- ✓ Run the analysis to obtain raw scores. Plot and comment your results
- ✓ Compare the results obtained via Wordfish with the results you get via Wordscores

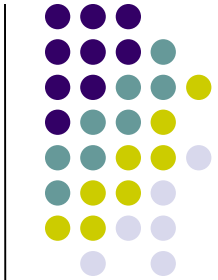
Deadline: 23 April 2021



SECOND PART

1. Retrieve the last 1,500 tweets from the official account of Joe Biden “@JoeBiden” via `rtweet`. Do you remember how to do it? *get_timeline(c(“JoeBiden"), n = 1500, include_rts=TRUE)*

Deadline: 23 April 2021

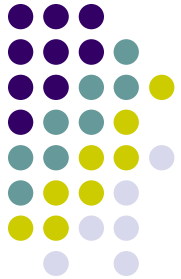


2. Run a STM on the set of tweets using as covariate for assessing the **topic prevalence** in your Structural Topic Model the day in which a tweet has been posted that you treat as a “continuous” variable
   - REMEMBER! When you assess a “continuous” topic prevalence model, your covariate should be a number, not a character or a date! Use the option: *as.numeric(the name of your time variable)* to convert your time-variable into a number. Save the new variable you just created in the data frame you got via your `rtweet` query. Then from this data frame create first your corpus, then your dfm, then convert your dfm into a stm object!

Deadline: 23 April 2021

3. Briefly comment your results (for example, why have you decided to select that specific number of topics? the content of the topics, the number of them, the results you got via topic prevalence, etc.)

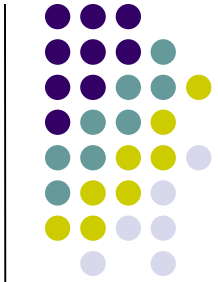
Deadline: 23 April 2021



THIRD PART

1. Make any query you like on Twitter
2. Run a semi-supervised classification analysis on it by identifying the keyword words you think are relevant given your query. Comment the results

Deadline: 23 April 2021



FOURTH PART

- Run a further query on Twitter as you like and download between 5,000 and 10,000 tweets
- Define a set of categories for the tweets you have downloaded (it can be a 2-set categories such as positive/negative, or a 3-set categories such as positive/negative/neutral, or anything else you want!!!)

Deadline: 23 April 2021

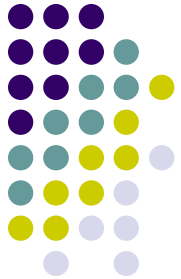
- Define a training-set (around 200 tweets if you have a 2-set categories; 300 if you have a 3-set categories, etc.) and a test-set. Manually codify the tweets
- Each of the student in the group (*if any*) must codify the same tweets

Deadline: 23 April 2021

In case you are working in group:

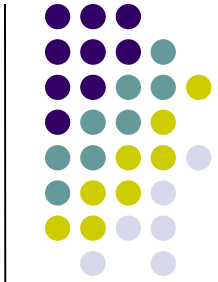
- Check your inter-coder reliability
- If the results are satisfactory (i.e., k>.7/.75) then cool! Each of the coder will use his/her codified tweets to classify the test-set
- If, however, you get an unsatisfactory result (i.e., k<.7) you should go back to the codification stage and find out why did happen and improve the classification so to increase the agreement score
- In each of your assignment, plz write me the ppl belonging to your group, the k-value you get for the inter-coder reliability part, and if you had to repeat the analysis a n-number of rounds

Deadline: 23 April 2021



- Then run the 3 ML algorithms discussed in class on the training-set and pick up the best algorithm via cross-validation
- Finally, classify the test-set

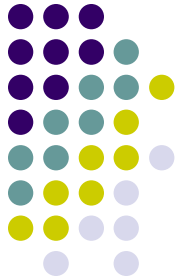
Deadline: 23 April 2021



FIFTH PART

- On the same corpus of tweets, run a local WE
- Explore some analogies
- Run 2 ML algorithms (of your choice) with the Word Embedding estimates

Deadline: 23 April 2021



Send to my email (luigi.curini@unimi.it) a pdf file with the results (that also include: scripts; tables/graphs; comments)