Topic Modelling Competition

The dataset provided are snippets of four books mixed up. Each snippet of 200 words can be considered to be a document, and each book can be considered as a class.

The four books are:

- Frankenstein (FS)
- Les Miserables (LM)
- Walden (WD)
- The Bible - New Testament (NT)

This is a classification problem where the aim is to assign each snippet to a book using topic modelling algorithm (LDA). This competition will be open from the **10th - 31st July (inclusive).** The results of this competition will be published first week of August.

You should:

1) Download the data and instructions from https://github.com/C3/DataScience/tree/master/DataScienceReadingClub/2017-07%20LDA/Competition/Instructions%20%26%20Data

2) Train an LDA model(s) using the training documents provided. Every document in the training set has a class (i.e. book name), and a 200 word snippet (See example below)

DocumentID	Documents	Class
1	of the holy city and from the things which are written in this book he which	NT
2	funeral pile triumphantly and exult in the agony of the torturing flames the lost in darkness and distance	FS
3	bore this kind of fruit and suffered it to drop off as fast as it ripened would prepare the way for a still more perfect and glorious state which also i have imagined but not yet anywhere seen	WD
4	achim begat eliud and eliud begat eleazar and	NT
5	by or the modern prometheus letter st petersburgh dec 11th to mrs saville	FS
6	economy when i wrote the	WD

	following pages or rather the bulk of them i lived	
7	i fantine so long as there shall exist by virtue of law and custom decrees of	LM

3) Using the testing snippet data set, predict the book of each testing document. (I.e. predict the class of each document)

This competition will be ranked on whether or not each document is assigned to the right book.

How to enter:

- Submit your results in a csv file (sample provided below) by email to Nicole (Nicole.pinto@eyc3.com), please make sure you write "LDA competition submission" in the subject of the email.
- Please name your submission file "<First Name>_<Last Name>.csv".
- At the end of the competition, *if you are one of the top 3 submissions, you will be asked to submit your commented code to qualify for a winning position.*

Sample submission:

Sample format of csv file to be submitted. Please make sure you use "LM", "NT", "FS", and "WD" for the book classes.

Example ONLY

DocumentID	Predicted_class
1	FS
2	WD
3	NT
4	LM
5	FS
6	WD
7	NT
8	LM
9	FS
10	WD
11	NT
12	LM
13	LM
14	FS

Resources:

WARNING: If you are planning to use LDA libraries in topic modelling, please make sure you are using the library for Latent Dirichlet Allocation, as LDA in machine learning can also refer to Linear Discriminant Analysis.

- https://github.com/C3/DataScience/tree/master/DataScienceReadingClub/2017-07%20LDA/Resources
- http://tidytextmining.com/topicmodeling.html (R)
- https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/ (Python)
- http://chdoig.github.io/pygotham-topic-modeling/#/ (Python)

Please use your “au.ey.com” email when joining Mattermost. For github please create an account and ask Nicole or