Text Mining and Text Retrieval Group Assignment

2020-2021

General Instructions

- Submit the Jupyter Notebook in electronic form before **Thursday January 21, 23:59 Amsterdam Time**
- The Notebook should be submitted through CANVAS
- No late homework reports will be accepted.
- The Notebook should be submitted by one individual per group.
- The group members should be identified in the Notebook.
- The Notebook text cells should be written in English.
- The Notebook can run from end to end. If necessary, have variables in the first cell that point to the directories where the files are, so I can adjust for my environment
- It is possible to use a Google Colab Notebook
- Take into account the shell commands for downloading / unzipping the data
- Take into account the installation of libraries through pip
- Use Text Cells to discuss your choices and results
- For questions, please use Canvas.

Introduction

You will work with a corpus made of news articles. You can download the data from UvA Sharepoint. This dataset originates from Kaggle.

This is a collection of approximately 140.000 articles from 15 different sources, published between the years 2000 and 2017, the balance between publications is shown in Figure 1.

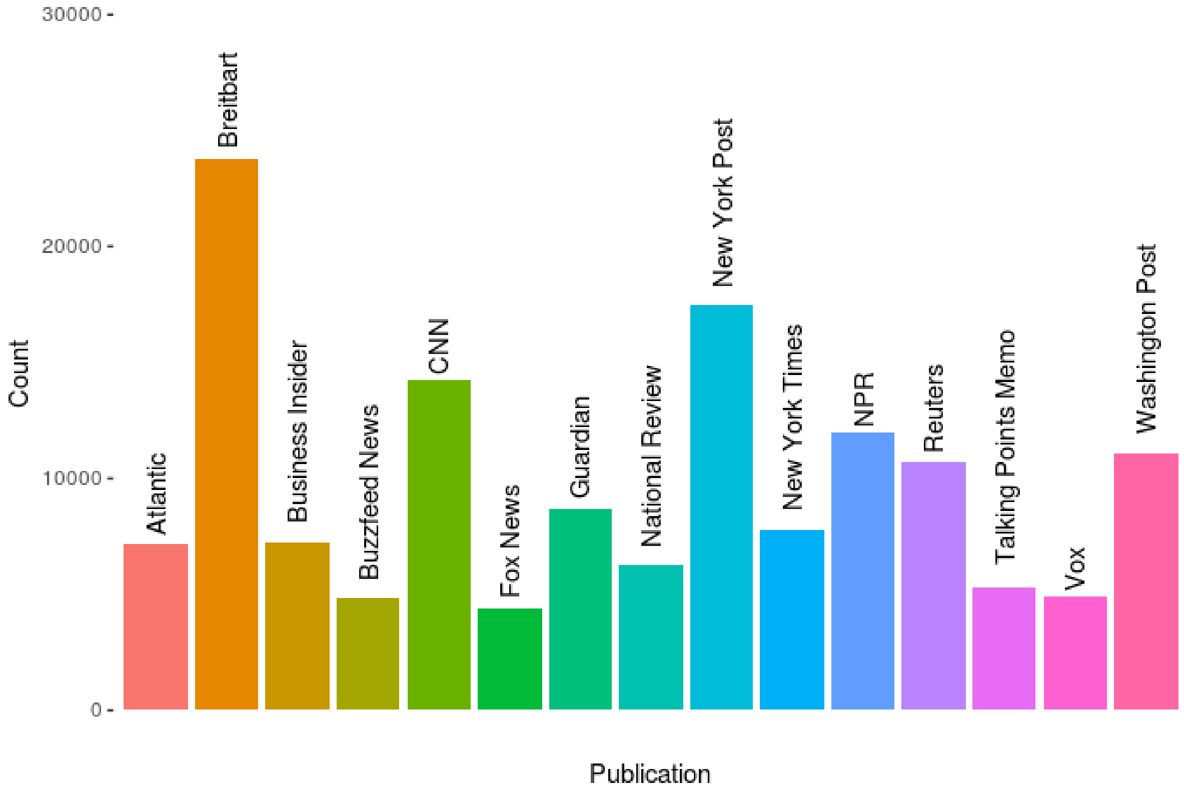

Figure 1: Balance of Publications

The data are provided as 3 CSV files, it is advised to load them as one single pandas dataframe.

The fields can be described as follows:

- Publication: the name of the media that published the article. This is one item from the list above.
- Title: The title of the article
- Date: the publication date
- Year: publication year
- Month: publication month
- Content: the full text of the article
- Author: name of the author
- ID: an id to identify each article

This homework will cover the material presented in Lectures 1 and 2: Lexical and Semantic Representation of Text. You can use any (or all) of the methods described along with anything you covered in the Machine Learning course regarding classification.

You will submit a Jupyter Notebook, in which you shall mention all the processing done to the text (stopping, stemming, . . .). The Jupyter Notebook should work, it will be used as a tool for result reproducibility, but the quality of the code will not be assessed.

Regarding the text cleaning, remember you have the freedom to complete the list of stopwords, to remove words based on document frequency, etc... mention what you do, explain it shortly. If this is an intuition, or an assumption, state it. There are no trap, and most of the time, there are no good/bad decision.

Assignment

Question:	1	2	3	4	Total
Points:	20	20	40	20	100

You are going to analyze the content of the articles, inspired by the Vogue analysis we discussed in class.

1. (20 points) Describe the corpus
   (a) Display the distribution of number of articles published per year
   (b) Select one specific year you will study further, select 4 publications
   (c) Display the split between the 4 publications in your year of study
   (d) From now on, consider only the articles published during the selected year by the selected 4 publications
   (e) Consider each article as a document
   (f) Illustrate your description with statistics: distribution of document length (number of tokens) across publications, most frequent words per publication

2. (20 points) Pre-process the corpus
   (a) Create the Dictionary, reduce it to a viable size
   (b) Use Stopwords and Stemming
   (c) Use Frequency Filtering
   (d) (Optional) Use token pattern

3. (40 points) Study your corpus with Topic Modeling
   (a) Select a value for K (number of topics)
   (b) Perform LDA on your corpus
   (c) Select a topic from LDA, annotate it by giving it a name
   (d) Illustrate with the Top 10 articles with highest saturation in your topic
   (e) Group of 5: perform Hyperparameter optimization for Coherence

4. (20 points) Predict a publication
   (a) You will build document classifiers that can predict which publication originated the article
   (b) Provide at least 2 classifiers: one based on Bag of Words, another based on Topic Modeling
   (c) Explain the difference in accuracy