Visualization project learning diary 1

Heli Huhtilainen - 12.4.2021

Project overview

This spring I finished doing a deep learning project for the Introduction to Deep learning course with my coursemate Jaakko Vilenius. The project was about the Reuters corpus document classification. The corpus data consist of news articles from the year 1997 and are stored as XML files. All the documents have metadata about the topics of the articles, the date, and the city and country. The text and topic part data is already a bit familiar to me after the deep learning project, and I thought it would be fun to see if this data has more to offer and to see what kind of interactive visualizations I can come up with it. Already I know that there are 103 different topics for the documents, and it varies a lot how popular the topics are. There is also a lot of variation in the length of the texts.

I am interested in combining some NLP techniques with visualization. The corpus text could be processed and the most important words shown as word clouds. Also, some distribution data could be shown, and there could be some filtering according to topic, location, or date. The location aspect could be visualized by finding the locations for the documents on a map, and the temporal aspect could be utilised with an animation. I wish to make a Heroku web application to show my visualizations, so it will be publicly available. The copyright limitations of the corpus prevent from showing too much of the actual texts, so this must be taken into account in the process.

Data loading and initial transformations

I already had the XML files from the project and a script that was originally made by Jaakko that I could utilize to make a data frame. In our deep learning project, we had not needed most of the metadata in the XML files and I was happy to find out that there were some interesting features available. I changed the data transforming script to include the date, location, and country. The XML files only had shortcodes for the topics, but there was an additional file where the codes were explained. I mapped the topic descriptions to the data rows to see more clearly what the topics were. I decided to keep the headline and main text in their own separate columns, as it might be an option to use just the headlines in some context. The essentials column is more or less the same text column that was used in our project: in this column, the words from both headline and main text are lemmatized, and stop words are removed. I think this column might be most useful when making word clouds.

An overview of my data frame is shown in figure 1. When taking a glimpse of the data frame, some things still need fixing: the extra index is probably not needed, as there are ids

for the articles. Also, the column names could all be in a singular form. It looks like there are some extra whitespaces in the headlines and texts, so these still need to be cleaned.

	id	date	location	country	codes	descs	headlines	texts	essentials
0	635751	1997-06-04	TORONTO	EU	[G15, GCAT]	[EUROPEAN COMMUNITY, GOVERNMENT/SOCIAL]	Official Journal contents - OJ C 169 of June 4...	* (Note - contents are displayed in reverse or...	COMMISSION 14 4 displayed OJ Commission within...
1	635752	1997-06-04	QUEBEC CITY	EU	[G15, GCAT]	[EUROPEAN COMMUNITY, GOVERNMENT/SOCIAL]	Official Journal contents - OJ L 144 of June 4...	* (Note - contents are displayed in reverse or...	4 displayed 29 Regulation Annexes Latvia Lithu...
2	635753	1997-06-04	OTTAWA	CANADA	[M14, M143, MCAT]	[COMMODITY MARKETS, ENERGY MARKETS, MARKETS]	Suncor lowers Canada heavy oil price.	* (Note - contents are displayed in reverse or...	29.3 oil 142.00 .... BLEND 06/04/97 IMPERIAL B...
3	635754	1997-06-04	TORONTO	CANADA	[M14, M143, MCAT]	[COMMODITY MARKETS, ENERGY MARKETS, MARKETS]	Suncor cuts Canada light oil prices.	* (Note - contents are displayed in reverse or...	25.76 oil .... SWEET 163.00 SOUR 25.44 06/04/9...
4	635755	1997-06-04	TORONTO	CANADA	[M11, MCAT]	[EQUITY MARKETS, MARKETS]	Toronto stocks end softer, more consolidation ...	* (Note - contents are displayed in reverse or...	said earnings DJI earlier * -42.49 -0.70 . mar...
...	...	...	...	...	...	...	...	...	...
281004	699748	1997-07-01	PARIS	UK	[M14, M141, MCAT]	[COMMODITY MARKETS, SOFT COMMODITIES, MARKETS]	Euro veg oils little changed despite U.S. 4 yr...	France's new leftist government, in its first ...	said coconut 4 mixed 512.50 Newsroom south old...
281005	699749	1997-07-01	PARIS	UK	[C31, CCAT, M14, MCAT]	[MARKETS/MARKETING, CORPORATE/INDUSTRIAL, COMM...	London shipsales.	France's new leftist government, in its first ...	10 inspection 23,803 survey ORE vessel final S...
281006	699750	1997-07-01	PARIS	UK	[GCAT, GENV, GWEA]	[GOVERNMENT/SOCIAL, ENVIRONMENT AND NATURAL WO...	Britain's June rainfall highest since 1860.	France's new leftist government, in its first ...	said replenished left come flat reservoir Afri...
281007	699751	1997-07-01	BEIJING	UK	[C21, C24, CCAT]	[PRODUCTION/SERVICES, CAPACITY/FACILITIES, COR...	Hess N.Sea Durward field sees small delay.	France's new leftist government, in its first ...	said 32 Plc Newsroom 7930 half suffered Group ...
281008	699752	1997-07-01	SINGAPORE	UK	[M14, M141, MCAT]	[COMMODITY MARKETS, SOFT COMMODITIES, MARKETS]	LIFFE sugar mostly higher after NY July expiry.	France's new leftist government, in its first ...	said morning 1025 one 7928 +44 newsroom lot Tu...

281009 rows × 9 columns

Figure 1. An overview of the initial data frame shows some of the features of the corpus.

This week I also created a GitHub repository for my project. I thought that it is good to use the repository right from the beginning, and it will be especially needed when I do the web application.

Plan for the next week

Next week I will take a good look at the contents of my data frame, and do all the cleaning that still needs to be done. After that, I will do some exploratory data analysis to get some statistics of the data. I hope this will help me come up with good ideas of what kind of visualizations would be most interesting and useful, what purpose they would serve, and what problems they would solve. Next week I will also start designing visualization layouts and concepts on the ground of these ideas.

References

Source code for the project: https://github.com/apndx/ReutersVisualizer

Reuters Corpus: https://trec.nist.gov/data/reuters/reuters.html