

Project A



Text Analysis : 10K Filings

Elisa Etienne
Hugo Poncelet
Valentin Hamers

STEP 1 - Data Collection



Selenium
WebDriver

1. Get page: Uber and Tesla
 - Using Selenium Webdriver to open our links to Uber and Tesla
 - Making a soup and then looking for the links of the articles.

In Form Type	10-K		
Company Name:	{UBER}		
Central Index Key (CIK):	Central Index Key (CIK)		
Start Date:	jj-mm-aaaa	End Date:	jj-mm-aaaa
Clear		Search	

Search Results

results:2

10-K 10-K 10-K Form 10-K - Period Ending 2020-12-31	Uber Technologies, Inc (UBER) (CIK 0001543151) 2021-03-01
10-K 10-K 10-K Form 10-K - Period Ending 2019-12-31	Uber Technologies, Inc (UBER) (CIK 0001543151) 2020-03-02

STEP 1 - Data Collection

3. Get Data From Page :

- a. Title
- b. Submitted and published dates
- c. Period Ending
- d. Text (without table and hidden/useless information)

Form 10-K Uber Technologies, Inc Annual report [Section 13 and 15(d), not S-K Item 405]

[SEC.report](#) / [Uber Technologies, Inc](#) / [Form 10-K](#)



Published: 2021-03-01 08:06:48

Submitted: 2021-03-01





























Filing Agent: [Wynaendts Alexander R](#)

Period Ending In: 2020-12-31

STEP 2 - Data Persistence (save to Mongo)

All data in the same collection

divided in “classes” (id; title; period ending; text...)

ProjectACollection						
	_id ObjectId	title String	published String	submitted String	period	
1	607008af3c272a4e4060b2c4	"Tesla, Inc. 2020 Annual Report	"2021-02-08"	"2021-02-08"	"2020-	   
2	607008c43c272a4e4060b2c6	"Tesla, Inc. 2019 Annual report	"2020-04-28"	"2020-04-28"	"2019-	   
3	607008f33c272a4e4060b2c8	"Tesla, Inc. 2019 Annual Report	"2020-02-13"	"2020-02-13"	"2019-	   
4	6070092e3c272a4e4060b2ca	"Tesla, Inc. 2018 Annual Report	"2019-02-19"	"2019-02-19"	"2018-	   
5	607009573c272a4e4060b2cc	"Tesla, Inc. 2017 Annual Report	"2018-02-23"	"2018-02-23"	"2017-	   
6	6070098f3c272a4e4060b2ce	"Tesla, Inc. 2016 Annual Report	"2017-03-01"	"2017-03-01"	"2016-	   
7	607009b93c272a4e4060b2d0	"Tesla, Inc. 2015 Annual Report	"2016-02-24"	"2016-02-24"	"2015-	   
8	607009ea3c272a4e4060b2d2	"Tesla, Inc. 2014 Annual Report	"2015-02-26"	"2015-02-26"	"2014-	   

STEP 3 - Total number of times “competition” is mentioned per year (per company)

Database Querying

```
_id: ObjectId("607007029502949b27d7ba98")  
title: "Tesla, Inc. 2020 Annual Report 10-K"  
published: "2021-02-08"  
submitted: "2021-02-08"  
periodEnding: "2020-12-31"  
text: "  tsla-10k_20201231.htm          UNITED STATES SECURITIES AND EXCHANGE ..."
```

total number of documents: 15

total number of documents containing the word competition: 13

STEP 3 - Total number of times “competition” is mentioned per year (per company)

	title	competition occurrence
1	Tesla, Inc. 2020 Annual Report 10-K	8
2	Tesla, Inc. 2019 Annual report 10-K/A	0
3	Tesla, Inc. 2019 Annual Report 10-K	10
4	Tesla, Inc. 2018 Annual Report 10-K	11
5	Tesla, Inc. 2017 Annual Report 10-K	10
6	Tesla, Inc. 2016 Annual Report 10-K	10
7	Tesla, Inc. 2015 Annual Report 10-K	9
8	Tesla, Inc. 2014 Annual Report 10-K	13
9	Tesla, Inc. 2013 Annual Report 10-K	16
10	Tesla, Inc. 2012 Annual Report 10-K	17
11	Tesla, Inc. 2011 Annual report 10-K/A	0
12	Tesla, Inc. 2011 Annual Report 10-K	19
13	Tesla, Inc. 2010 Annual Report 10-K	21
14	Uber Technologies, Inc 2020 Annual Report 10-K	40
15	Uber Technologies, Inc 2019 Annual Report 10-K	43

STEP 4 - extraction of the top-20 most frequent bi-grams

	bigram	occurrence		bigram	occurrence
1	(december, 31)	3317	11	(unite, state)	820
2	(common, stock)	1905	12	(battery, pack)	787
3	(administrative, agent)	1855	13	(financial, condition)	775
4	(end, december)	1704	14	(electric, vehicles)	750
5	(fair, value)	1493	15	(confidential, treatment)	739
6	(year, end)	1182	16	(energy, systems)	711
7	(solar, energy)	1119	17	(treatment, request)	710
8	(shall, mean)	1062	18	(0, million)	1062
9	(financial, statements)	953	19	(pay, agent)	710
10	(operate, result)	871	20	(balance, sheet)	706

Thank you

Sources

- Datacamp, october 23, 2018, Hafsa JABEEN, *Stemming and Lemmatization in Python*,
https://www.datacamp.com/community/tutorials/stemming-lemmatization-python?utm_source=adwords_ppc&utm_campaignid=898687156&utm_adgroupid=48947256715&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adposition=&utm_creative=229765585183&utm_targetid=aud-763347114660:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1001394&gclid=Cj0KCQjw9_mDBhCGARIsAN3PaFMU57aELO1FskM0wD3fwfGwQg8oEJnv_QFmzmkues69WxF3btX-9d0aAgXuEALw_wcB
- freeCodeCamp, april 16, 2018, Dave GRAY, *Better web scraping in python with Selenium, BeautifulSoup and Pandas*,
<https://www.freecodecamp.org/news/better-web-scraping-in-python-with-selenium-beautiful-soup-and-pandas-d6390592e251/?fbclid=IwAR06PDWAtWCrdabCkp9w4oUF7Vo5Yt7gIj33SmOdG68iSfg-eWT98bc12U>
- TutorialsPoint, https://www.tutorialspoint.com/python_text_processing/python_bigrams.htm
- Stackoverflow, <https://stackoverflow.com/questions/50655203/how-to-efficiently-count-bigrams-over-multiple-documents-in-python/50656229>