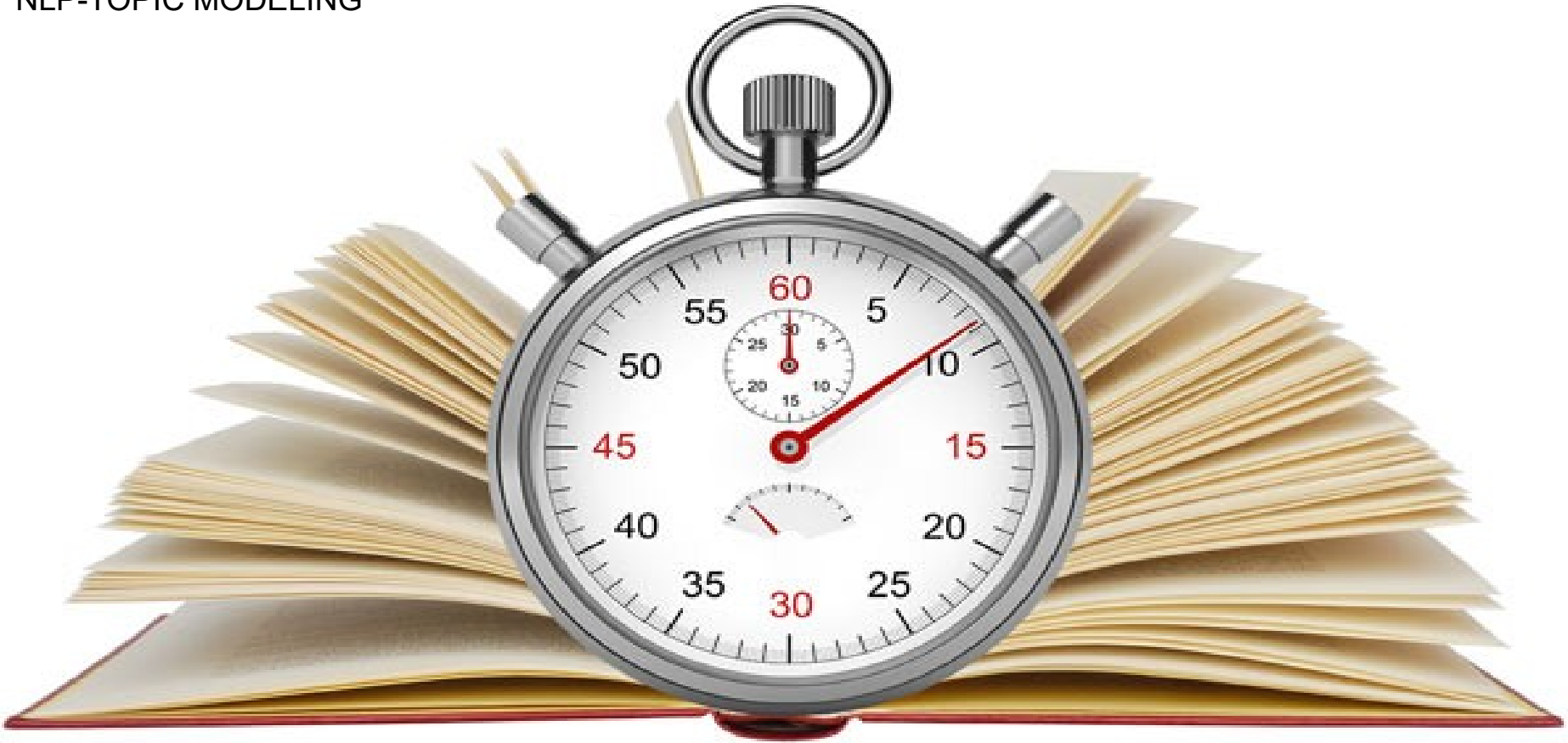
NLP-TOPIC MODELING



Overview

Topic modeling is a context of natural language processing which helps to group or cluster unlabeled texts (unsupervised process) , through hidden structure (similarities) between of them.

GOAL

Use nlp to read big amount of Financial news , in real time , and provide asset managers with information advantage.

Data description

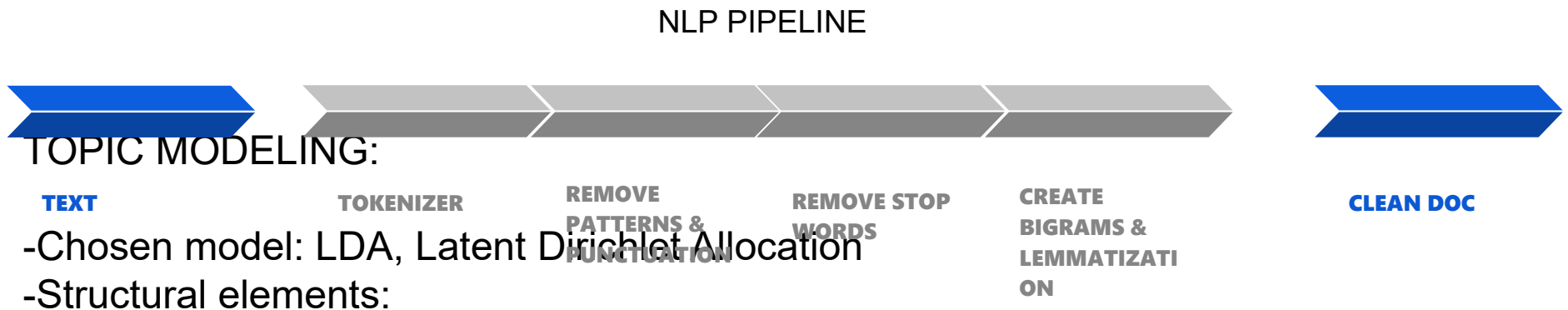
Provided data by consultancy company, collected from rated as reliable sites for News

-Number of text-documents: 362.871

-Time period of collection: 365 days

Implementation

NLP: prepare and clean our data



TOPIC MODELING:

-Chosen model: LDA, Latent Dirichlet Allocation

-Structural elements:

1)Dictionary: assign words to unique id --->turning words to numbers

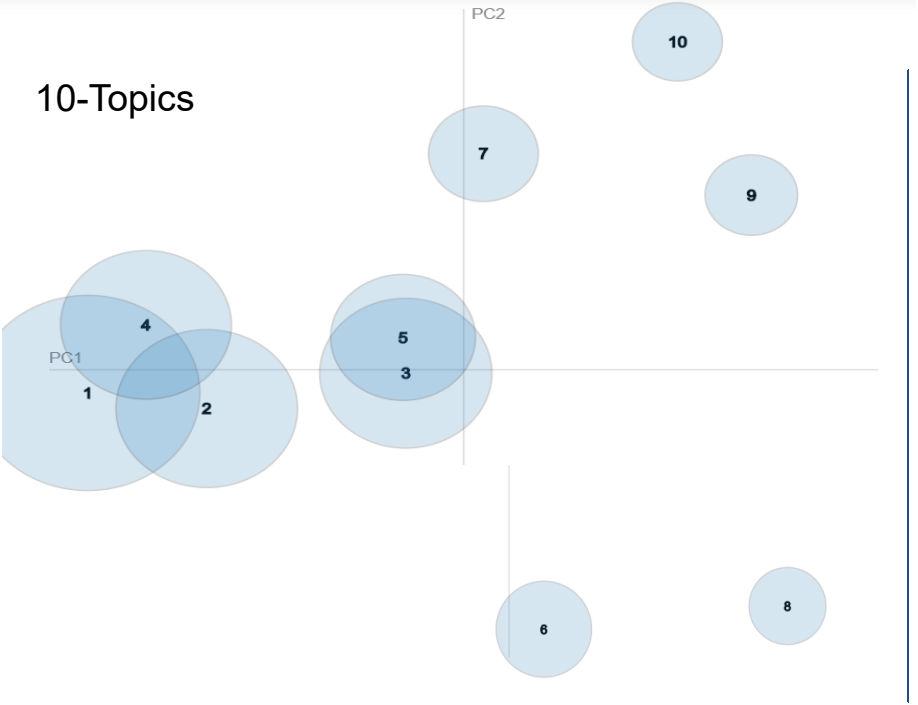
2)Corpus:a union of sub-bags of words for each doc --->vectorization of data

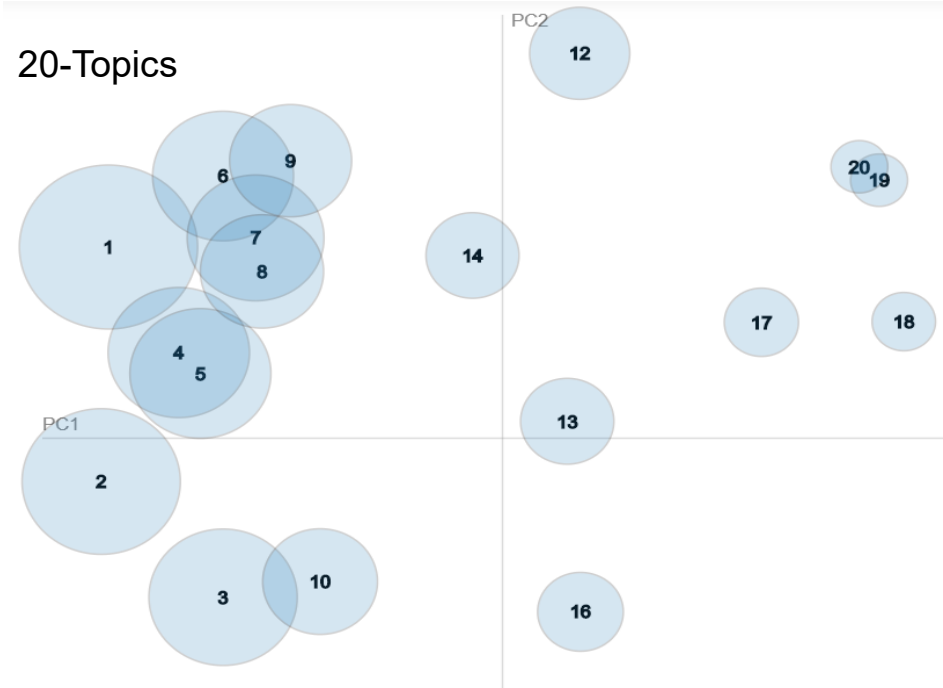
HOW MANY TOPICS?-CAN WE KNOW?

-A way to go is:

1)Through visualization

-The more segmented are, the less overlapped , the more clear Thema(Topic).





2) Coherence metric:

A pipeline that turns a qualitative result to quantitative.
Coherence results:
- 7 Topics = 0.484
-10 Topics = 0.502
-20 Topics = 0.495
-40 Topics = 0.428
-Looks 10 is the number of Topics we have to continue and extract further insights.
-Are these enough reliable criteria?

10-TOPICS VS 20-TOPICS

```
[(0,
  '0.014*"service" + 0.010*"business" + 0.010*"technology" + 0.010*"company" + '
  '0.009*"industry" + 0.009*"information" + 0.008*"include" + 0.008*"india" + '
  '0.008*"provide" + 0.008*"datum"'),
 (1,
  '0.035*"december" + 0.028*"news" + 0.017*"tender" + 0.015*"research" + '
  '0.014*"file" + 0.012*"open" + 0.011*"thursday" + 0.011*"report" + '
  '0.011*"friday" + 0.009*"november"'),
 (2,
  '0.090*"share" + 0.085*"company" + 0.023*"hold" + 0.019*"value" + '
  '0.019*"sell" + 0.019*"group" + 0.017*"price" + 0.016*"inc" + '
  '0.016*"financial" + 0.015*"management"'),
 (3,
  '0.019*"water" + 0.015*"power" + 0.015*"vehicle" + 0.012*"road" + '
  '0.011*"energy" + 0.011*"area" + 0.011*"construction" + 0.010*"plant" + '
  '0.009*"supply" + 0.009*"application"'),
 (4,
  '0.023*"health" + 0.013*"study" + 0.013*"medical" + 0.012*"care" + '
  '0.012*"hospital" + 0.010*"drug" + 0.010*"patient" + 0.009*"report" + '
  '0.009*"research" + 0.008*"treatment"'),
 (5,
  '0.013*"year" + 0.010*"people" + 0.009*"time" + 0.009*"pron" + 0.006*"come" '
  '+ 0.006*"work" + 0.005*"know" + 0.005*"home" + 0.005*"family" + '
  '0.005*"child"'),
 (6,
  '0.011*"year" + 0.010*"plan" + 0.009*"government" + 0.009*"work" + '
  '0.008*"change" + 0.008*"fund" + 0.007*"time" + 0.007*"project" + '
  '0.007*"state" + 0.006*"statement"'),
 (7,
  '0.033*"date" + 0.027*"notice" + 0.026*"contract" + 0.025*"document" + '
  '0.020*"information" + 0.018*"address" + 0.017*"country" + 0.015*"type" + '
  '0.012*"article" + 0.012*"contact"'),
 (8,
  '0.040*"trade" + 0.039*"stock" + 0.032*"rate" + 0.024*"quarter" + '
  '0.023*"market" + 0.023*"year" + 0.018*"report" + 0.017*"dollar" + '
  '0.015*"percent" + 0.015*"price"'),
 (9,
  '0.018*"president" + 0.017*"trump" + 0.013*"government" + 0.010*"state" + '
  '0.009*"election" + 0.008*"police" + 0.008*"party" + 0.008*"monday" + '
  '0.008*"vote" + 0.008*"border"')]
```

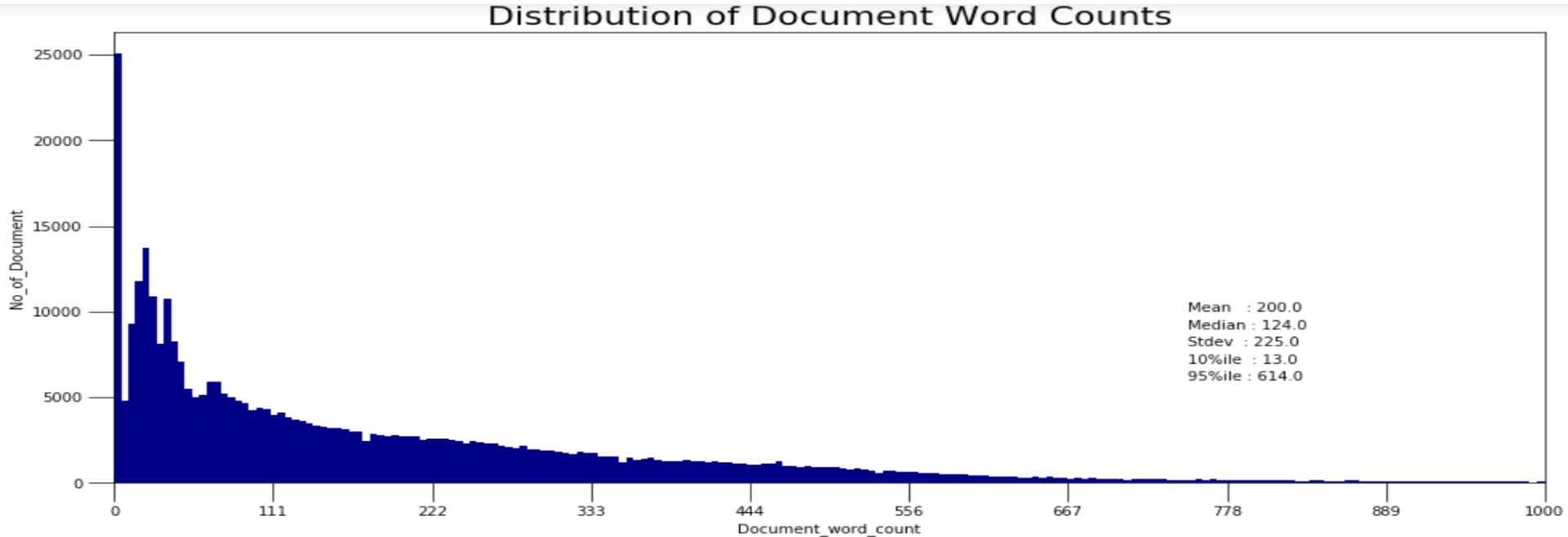
-A direct comparison of topic-9 of the 10-topic model with the topic-8 of the 20-topic model can show us that the thematic in the second one is more clear, and this is what we need. We need more clear and higher coherence between the words that targeting to financial news, even if most of topics are not clear but a couple of them are strong connected for potential financial investments. So in our case will continue with the 20-topics model.

20 TOPICS

```
[(0,
  '0.022*"service" + 0.019*"system" + 0.018*"information" + 0.018*"technology" '
  '+ 0.016*"datum" + 0.012*"provide" + 0.010*"include" + 0.010*"website" + '
  '0.010*"network" + 0.010*"application"'),
 (1,
  '0.012*"photo" + 0.007*"video" + 0.006*"world" + 0.006*"start" + '
  '0.006*"away" + 0.006*"play" + 0.005*"show" + 0.005*"appear" + 0.005*"image" '
  '+ 0.005*"game"'),
 (2,
  '0.079*"contract" + 0.028*"supply" + 0.026*"article_content" + '
  '0.026*"requirement_contact" + 0.018*"australia" + 0.017*"profit" + '
  '0.017*"volume" + 0.015*"value" + 0.015*"hong_kong" + 0.011*"fuel"'),
 (3,
  '0.078*"wall" + 0.023*"gold" + 0.021*"material" + 0.013*"temperature" + '
  '0.012*"heat" + 0.012*"producer" + 0.010*"metal" + 0.009*"maximum" + '
  '0.009*"powell" + 0.008*"inch"'),
 (4,
  '0.016*"case" + 0.016*"issue" + 0.014*"agency" + 0.014*"federal" + '
  '0.013*"department" + 0.013*"order" + 0.013*"public" + 0.012*"information" + '
  '0.011*"state" + 0.011*"report"'),
 (5,
  '0.023*"people" + 0.017*"pron" + 0.016*"time" + 0.011*"tell" + 0.011*"year" '
  '+ 0.011*"want" + 0.011*"come" + 0.011*"know" + 0.010*"work" + '
  '0.009*"think"'),
 (6,
  '0.038*"year" + 0.018*"percent" + 0.014*"high" + 0.011*"monday" + '
  '0.011*"month" + 0.009*"increase" + 0.009*"give" + 0.009*"change" + '
  '0.009*"expect" + 0.009*"economic"'),
 (7,
  '0.066*"december" + 0.025*"news" + 0.022*"open" + 0.020*"inc" + '
  '0.020*"thursday" + 0.020*"november" + 0.019*"address" + 0.017*"service" + '
  '0.016*"comment" + 0.015*"daily"'),
 (8,
  '0.031*"trump" + 0.028*"country" + 0.019*"president" + 0.016*"india" + '
  '0.015*"china" + 0.014*"border" + 0.012*"government" + 0.011*"force" + '
  '0.010*"shutdown" + 0.010*"security"'),
 (9,
  '0.021*"school" + 0.018*"year" + 0.017*"community" + 0.015*"cent" + '
  '0.014*"program" + 0.014*"work" + 0.012*"university" + 0.012*"student" + '
  '0.012*"member" + 0.011*"center"'),
 (10,
  '0.090*"rate" + 0.050*"dollar" + 0.045*"bank" + 0.040*"fund" + '
  '0.035*"financial" + 0.030*"investment" + 0.026*"market" + 0.019*"capital" + '
  '0.019*"revenue" + 0.018*"ratio"'),
 (11,
  '0.098*"share" + 0.072*"company" + 0.070*"stock" + 0.061*"trade" + '
  '0.037*"price" + 0.024*"hold" + 0.023*"sell" + 0.018*"value" + '
  '0.017*"purchase" + 0.014*"christmas"'),
 (12,
  '0.046*"tender" + 0.046*"date" + 0.039*"file" + 0.035*"notice" + '
  '0.034*"document" + 0.034*"court" + 0.019*"district" + 0.018*"type" + '
  '0.012*"person" + 0.011*"time"'),
 (13,
  '0.034*"city" + 0.033*"police" + 0.022*"area" + 0.019*"road" + 0.017*"water" '
  '+ 0.017*"fire" + 0.016*"park" + 0.014*"home" + 0.012*"station" + '
  '0.010*"south"'),
 (14,
  '0.025*"food" + 0.025*"store" + 0.022*"holiday" + 0.019*"sale" + '
  '0.018*"brand" + 0.018*"product" + 0.015*"retail" + 0.014*"shop" + '
  '0.012*"online" + 0.011*"consumer"'),
 (15,
  '0.082*"report" + 0.077*"quarter" + 0.055*"research" + 0.040*"news" + '
  '0.036*"analyst" + 0.033*"target" + 0.030*"year" + 0.030*"worth" + '
  '0.027*"earning" + 0.025*"group"'),
 (16,
  '0.039*"health" + 0.022*"study" + 0.021*"medical" + 0.021*"care" + '
  '0.019*"hospital" + 0.016*"drug" + 0.016*"patient" + 0.014*"treatment" + '
  '0.014*"editor_mailto" + 0.014*"query_respect"'),
 (17,
  '0.031*"company" + 0.023*"business" + 0.018*"project" + 0.016*"market" + '
  '0.016*"development" + 0.014*"industry" + 0.012*"statement" + '
  '0.011*"officer" + 0.010*"energy" + 0.009*"work"'),
 (18,
  '0.076*"index" + 0.036*"unit" + 0.035*"accuse" + 0.032*"vehicle" + '
  '0.017*"patent_classification" + 0.014*"aircraft" + 0.013*"state" + '
  '0.013*"equipment" + 0.012*"table" + 0.012*"valuation"'),
 (19,
  '0.027*"government" + 0.023*"state" + 0.021*"election" + 0.020*"party" + '
  '0.018*"vote" + 0.017*"president" + 0.016*"house" + 0.013*"democrat" + '
  '0.012*"political" + 0.011*"minister"')]
```

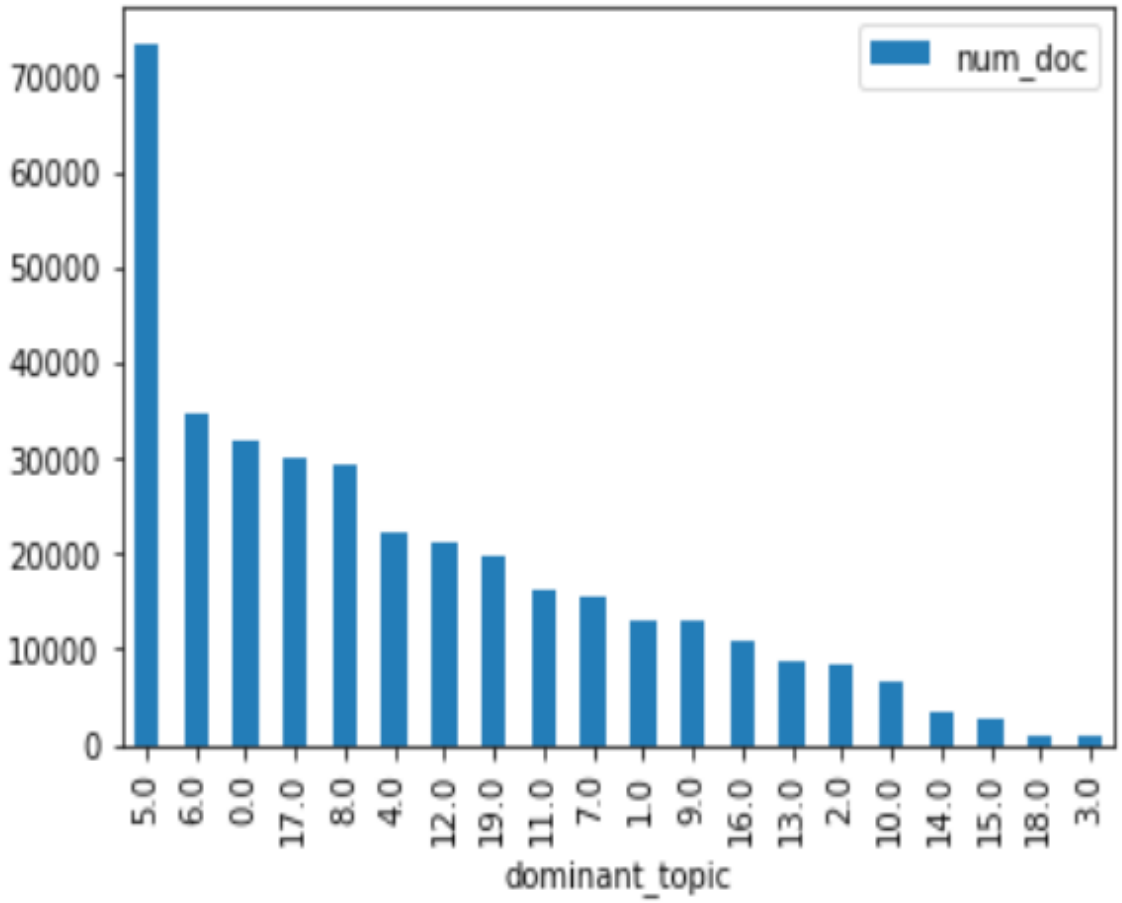
Extract info from the 20-TOPICS model

	doc_num	Dominant_Topic	Topic_Perc_Contrib	Keywords	text	just_date
0	0	5.0	0.1671	people, pron, time, tell, year, want, come, kn...	* Anger over corruption and hardship fuels pro...	2018-01-01
1	1	6.0	0.2462	year, percent, high, monday, month, increase, ...	Bureau of Economic Analysis Table 1.3.1. Perce...	2018-01-01
2	2	17.0	0.2500	company, business, project, market, developmen...	High-profile companies like Wells Fargo, Uber,...	2018-01-01
3	3	8.0	0.1588	trump, country, president, india, china, borde...	President Donald Trump slammed Pakistan for 'l...	2018-01-01
4	4	4.0	0.3455	case, issue, agency, federal, department, orde...	Scope.\nApproved March 28, 2013. 160 pages. A ...	2018-01-01



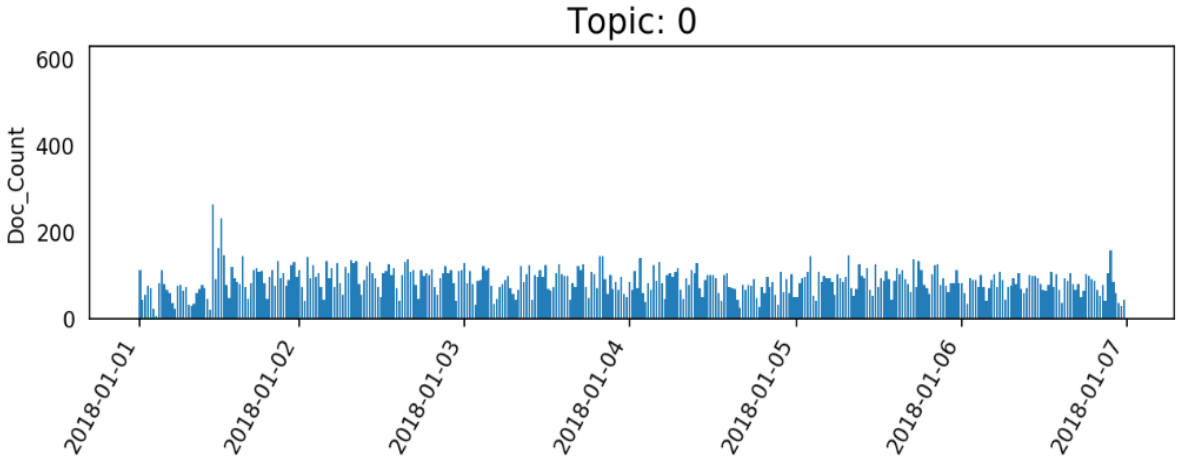
Total number of Doc per Topic

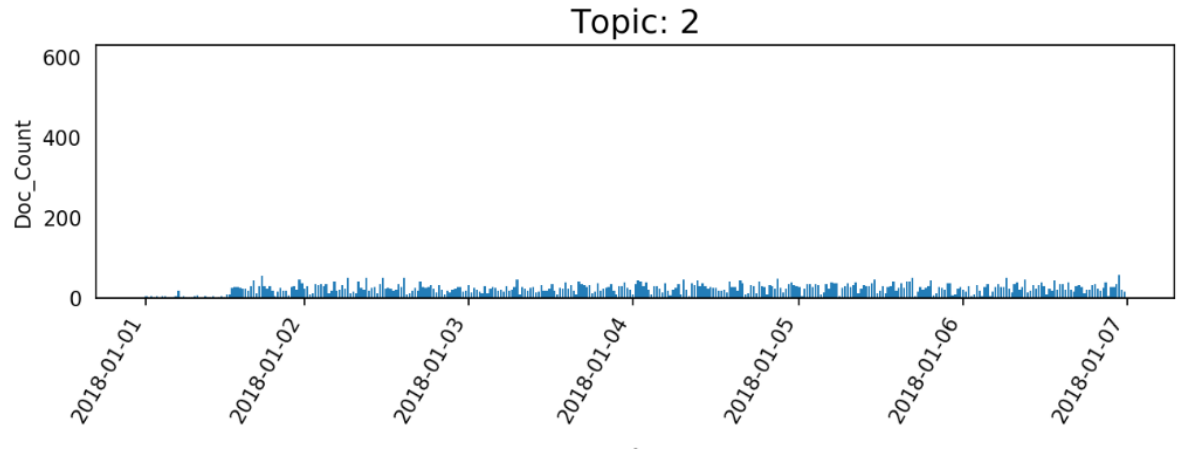
	dominant_topic	num_doc	perc_doc
0	5.0	73481	0.202
1	6.0	34671	0.096
2	0.0	31816	0.088
3	17.0	30059	0.083
4	8.0	29460	0.081
5	4.0	22365	0.062
6	12.0	21132	0.058
7	19.0	19711	0.054
8	11.0	16030	0.044
9	7.0	15498	0.043



Per day / Per Topic sum of Documents

	Dominant_Topic	just_date	sum_of_doc
0	0.0	2018-01-01	112
1	0.0	2018-01-02	45
2	0.0	2018-01-03	55
3	0.0	2018-01-04	76
4	0.0	2018-01-05	71
5	0.0	2018-01-06	24
6	0.0	2018-01-07	8
7	0.0	2018-01-08	82
8	0.0	2018-01-09	113
9	0.0	2018-01-10	80





Most representative Documents

```
[doc_num                                                    209239
 Dominant_Topic                                                  0
 Topic_Perc_Contrib                                         0.7387
 Keywords              service, system, information, technology, datu...
 text                  Alexandria, July 31 -- NEC Corporation has sec...
 just_date                                              2018-07-31
 Name: 209239, dtype: object,
 doc_num                                                    174703
 Dominant_Topic                                                  1
 Topic_Perc_Contrib                                         0.5692
 Keywords              photo, video, world, start, away, play, show, ...
 text                  * window._taboola = window._taboola || []; _ta...
 just_date                                              2018-06-26
 Name: 174703, dtype: object,
 doc_num                                                    325465
 Dominant_Topic                                                  2
 Topic_Perc_Contrib                                         0.5423
 Keywords              contract, supply, article_content, requirement...
 text                  Slovakia, Nov. 23 -- Contract Id: 1350536.\nDe...
 just_date                                              2018-11-24
 Name: 325465, dtype: object,
 doc_num                                                    358876
 Dominant_Topic                                                  3
 Topic_Perc_Contrib                                         0.8529
 Keywords              wall, gold, material, temperature, heat, produ...
 text                  28.12.2018 - 14:47 Uhr.\n\n\nDGAP-Ad-hoc: MyBu...
 just_date                                              2018-12-28
```

```
doc_num                                                     72650
Dominant_Topic                                                  4
Topic_Perc_Contrib                                         0.5913
Keywords              case, issue, agency, federal, department, orde...
text                  SUMMARY: In accordance with the Privacy Act of...
just_date                                              2018-03-16
Name: 72650, dtype: object,
doc_num                                                    172112
Dominant_Topic                                                  5
Topic_Perc_Contrib                                         0.6834
Keywords              people, pron, time, tell, year, want, come, kn...
text                  Love Island could live up to its name this eve...
just_date                                              2018-06-24
Name: 172112, dtype: object,
doc_num                                                     37437
Dominant_Topic                                                  6
Topic_Perc_Contrib                                         0.7549
Keywords              year, percent, high, monday, month, increase, ...
text                  Robust demand lifts mentha oil futures by 1.14...
just_date                                              2018-02-09
Name: 37437, dtype: object,
doc_num                                                    203214
Dominant_Topic                                                  7
Topic_Perc_Contrib                                         0.6994
Keywords              december, news, open, inc, thursday, november,...
text                  LANXESS Aktiengesellschaft: Release according ...
just_date                                              2018-07-25
```

Insights-Results-Comments

-We found in which doc each topic is assigned with the highest percentage.So, This document is the most related with each topic.
-Slides 8,9,10 are not giving us now any insights but are very important for the maintenance and stability of the model later.(if we want to run a classification model or do sentiment analysis)
How interpret ,labeling our topics and get benefit of the output.
-Topic 0: manual-label/ information technology
The context speaks about a patent on a communication device.This could be an area for potential investment.
-Topic 8: manual-label/ U.S foreign policy
The context speaks about military actions of U.S. in Persian gulf

GOAL SUCCEED

362.871 Documents

Downsized

20

THANKS