Université de Lille
Faculté d'Ingénierie et Management de la Santé (ILIS)
Master Data Science pour la Santé

A Big Data Framework for Sentiment Analysis in Healthcare: the case of Covid-19

Par Ulrich TCHUENKAM
Sous la direction de Wajdi DHIFLI

Data de soutenance : 20 Octobre 2020

ANNÉE UNIVERSITAIRE 2019/2020

OUTLINE

- MOTIVATION AND CONTEXT
- METHODS AND MATERIAL
  - Layers of the framework
  - Data management in the framework
- CASE STUDY ON COVID-19
- RESULTS
- DISCUSSION
- CONCLUSION

MOTIVATION AND CONTEXT

Huge amount of data on healthcare subjects

Feelings and thoughts classification models

Scalable, distributed and parallelizable technologies

Topics generation models

MOTIVATION AND CONTEXT

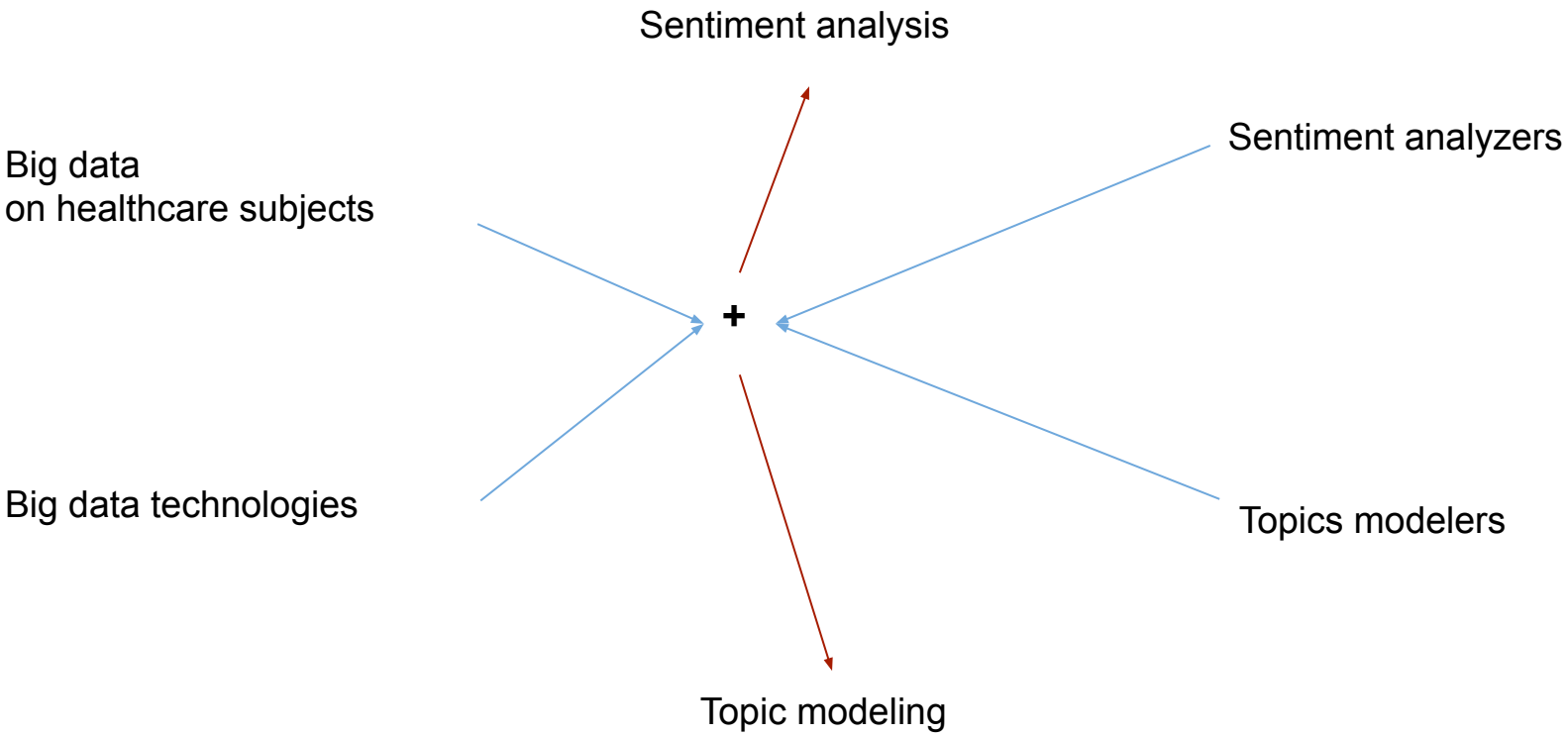
Huge amount of data on healthcare subjects → **Big data**

Feelings and thoughts classification models → **Sentiment analyzers**

Scalable, distributed and parallelizable technologies → **Big data technologies**

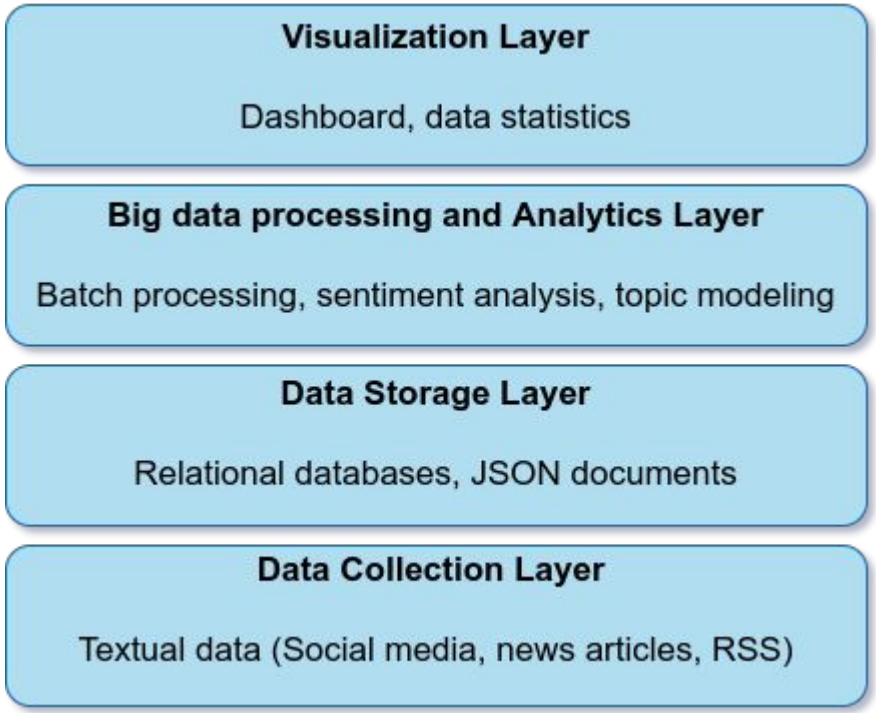
Topics generation models → **Topic modelers**

OBJECTIVE


Sentiment analysis
Sentiment analyzers
Big data
on healthcare subjects
+
Big data technologies
Topics modelers
Topic modeling

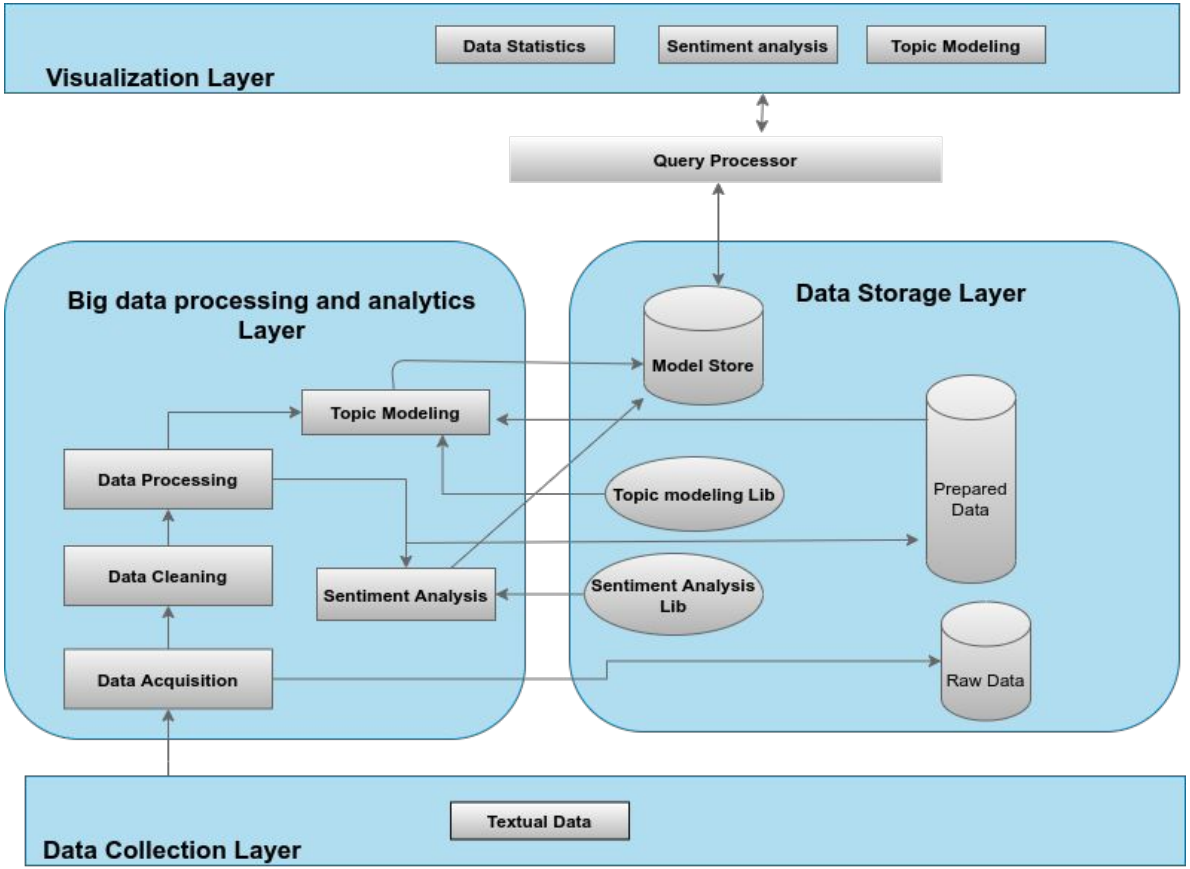
METHODS AND MATERIAL

Layers architecture



METHODS AND MATERIAL

Data management in the framework



CASE STUDY ON COVID-19

Data acquisition

- Using Twitter's API extract tweets about covid-19.

  +

- A public dataset containing tweets about corona.

Data storage

- PostgreSQL
- Elasticsearch

CASE STUDY ON COVID-19

Data description

- *Size* : 633549 tweets
- *Language* : English
- *Period* : April 1st to April 14th
- *Topic* : CORONA
- *Variables* :

Text	*Location*	*Date*

Data wrangling

- Removal of non-alphabetical characters
- Removal of duplicates
- Removal of emoticons
- Removal of punctuation marks

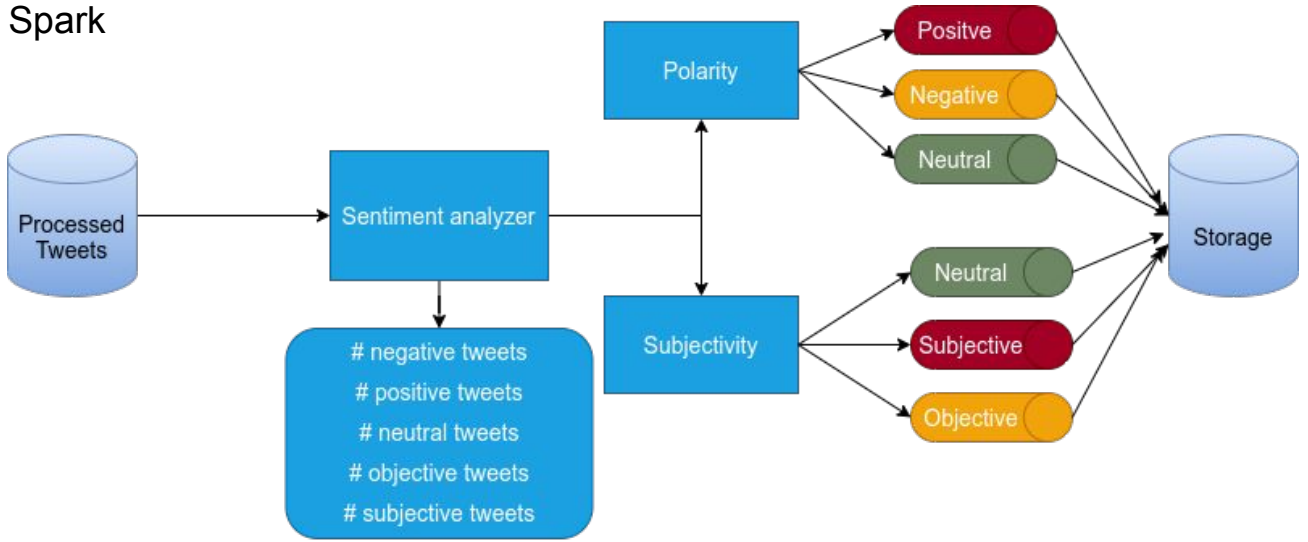
Data processing

- Tokenization
- Removal of stopwords
- Converting tokens to lowercase characters
- Creation of bigrams
- Stemming
- Lemmatization
- Creating corpus and dictionary.

CASE STUDY ON COVID-19

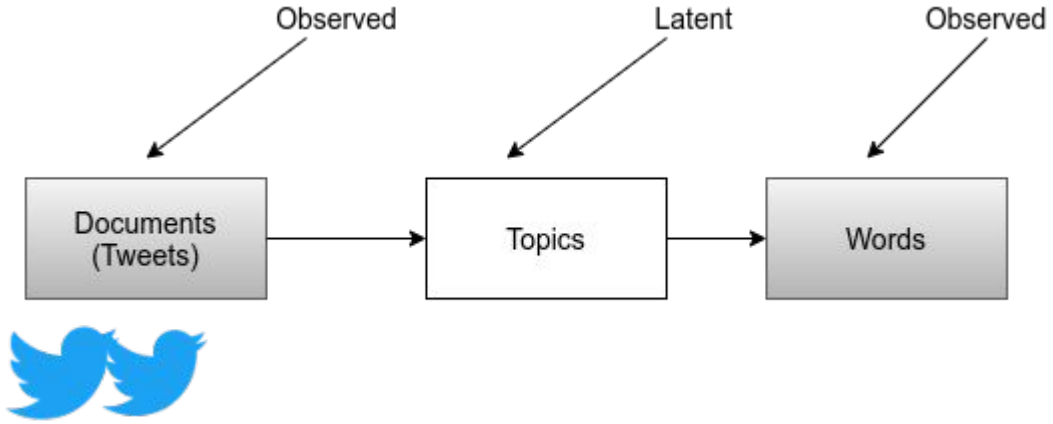
Sentiment analysis

Packages : TextBlob, Spark



CASE STUDY ON COVID-19

Topic Modeling



- Documents are about several topics at the same time.
- Topics are associated with different words.
- Topics in the documents are expressed through the words used.

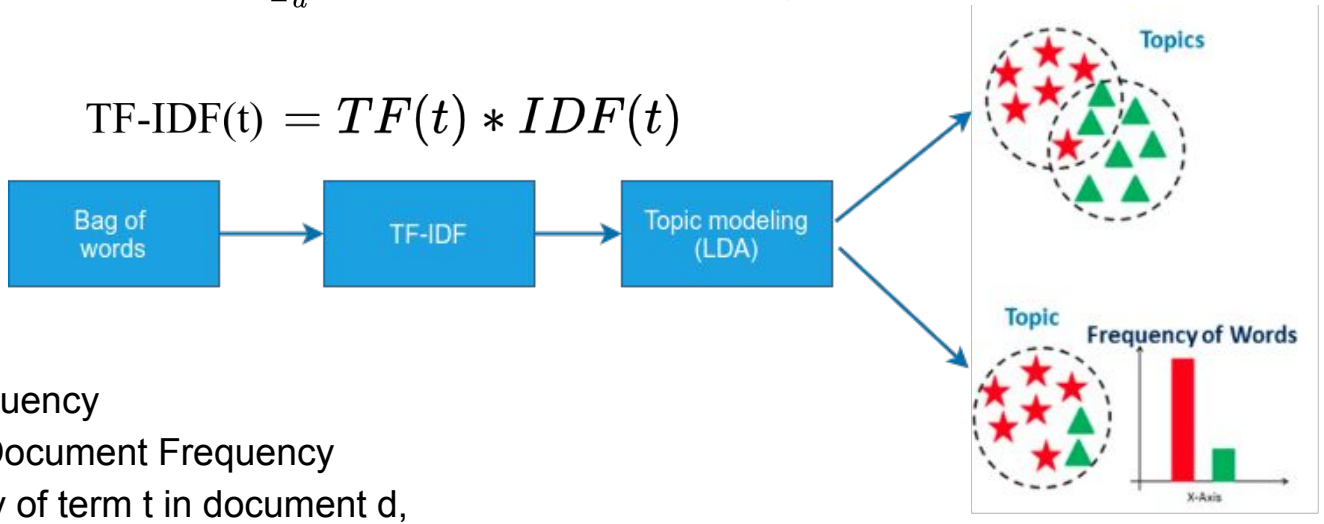
CASE STUDY ON COVID-19

Topic Modeling

$$\text{TF(t)} = \frac{f_{t,d}}{T_d} \quad , \quad \text{IDF(t)} = log(\frac{N}{D_t})$$

Package : Gensim

$$\text{TF-IDF(t)} = TF(t) * IDF(t)$$



Notations:

- TF = Term Frequency
- IDF = Inverse Document Frequency
- $f_{t,d}$ = Frequency of term t in document d,
- T_d = Number of terms in document d,
- N = Total number of documents,
- D_t = Number of documents with term t

CASE STUDY ON COVID-19

Evaluation metric : Topic coherence

$$\text{Coherence} = \sum_{i<j} \text{score}(w_i, w_j)$$

$$\text{score}_{\text{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j)+1}{D(w_i)}$$

for $w_1 \ldots . w_k$ k top words

Notations

- $D(w_i)$: count of documents containing the word w_i
- $D(w_i, w_j)$: count of documents containing both words w_i and w_j
- D : The total number of documents in the corpus.

RESULTS

Data Summary tab

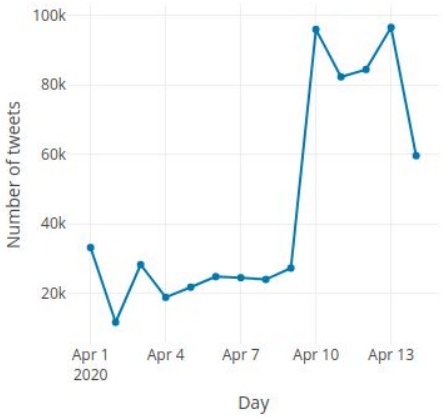
Package : Dash

Drag and Drop or Select dataset file

Preprocessing april_1_14_2020.csv completed and now using it

Total Number of Tweets: 633549

Size of dataframe: 324.22MB



RESULTS

Topic modeling tab (Model 1)

How many topics to model?

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

How many top terms per topic?

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Slide to select the document-topic density (ALPHA)

ALPHA = 0.001

0.001 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Slide to select the topic-word density (ETA)

ETA = 0.001

0.001 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Slide to select minimum Probability of topic?

Topics with Probability < 0.01 shall be rejected?

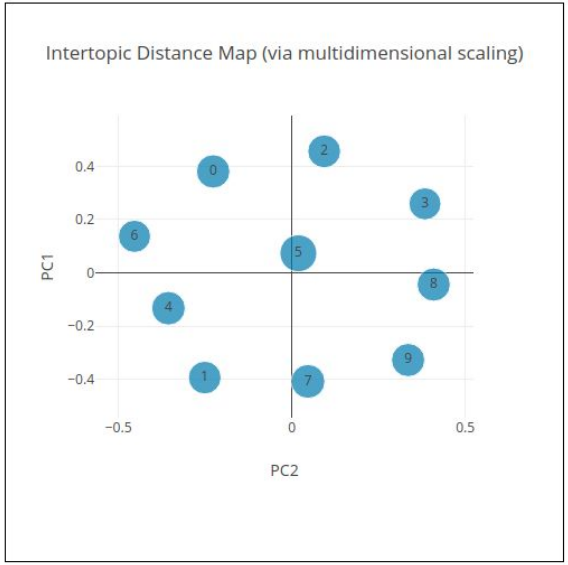
0.01 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Run Model(Basic) Run Model(Advance)

Metrics

Perplexity = -626.21

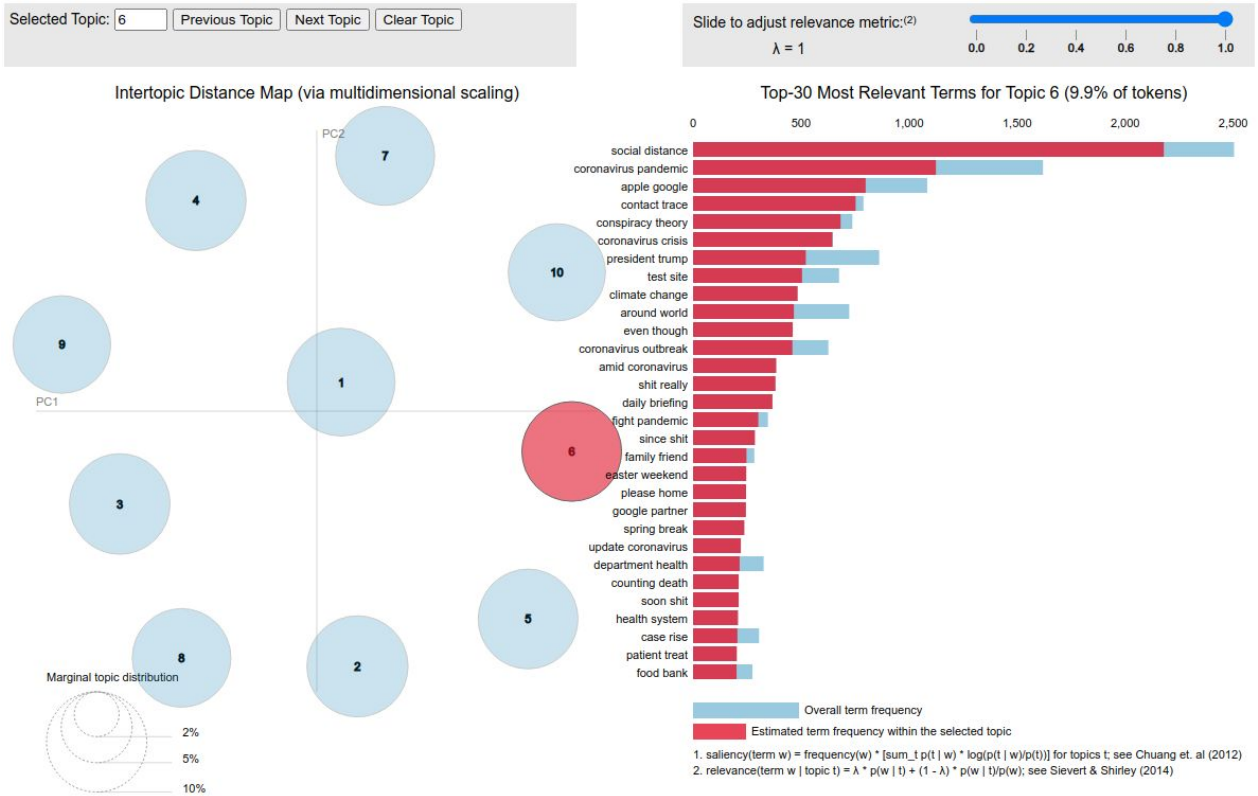
Coherence score = 0.71



Rename Topic 0	Rename Topic 1	Rename Topic 2	Rename Topic 3
Rename Topic 4	Rename Topic 5	Rename Topic 6	Rename Topic 7
Rename Topic 8	Rename Topic 9		

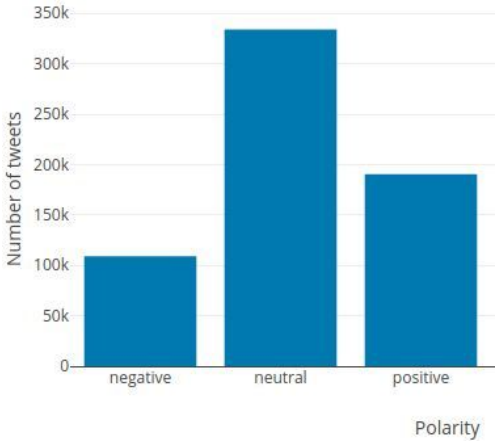
RESULTS

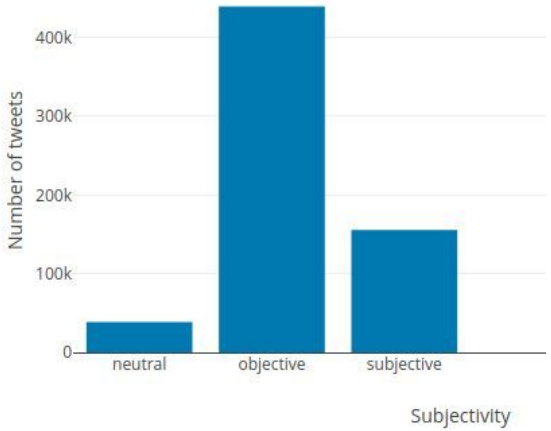
Topic modeling tab (Model 2)

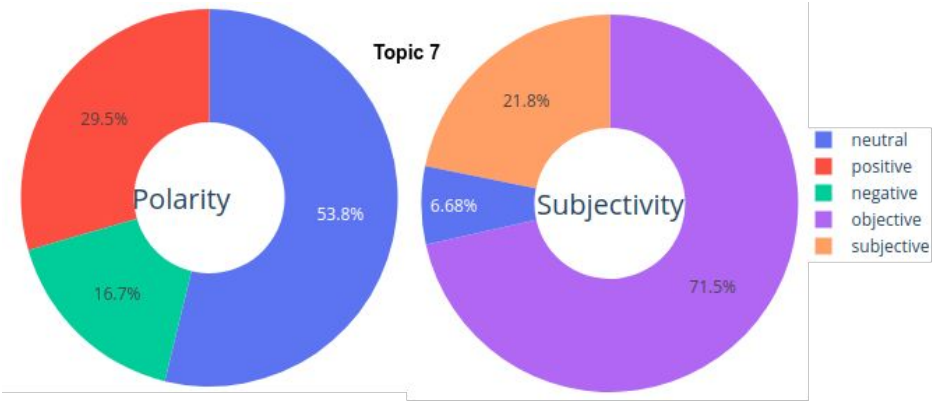


RESULTS

Sentiment analysis tab







DISCUSSION

- The API of Twitter provides a great way to retrieve tweets but has some limitations.
- Twitterscraper and GetOldTweets3 give access to older tweets but they only get a random number of them.

POTENTIAL FUTURE WORKS

- Leveraging Spark Streaming API.
- Add a user mapping feature to the framework.
- Add support for other corpus and languages.
- Use ontologies and knowledge graphs to enhance sentiment analysis and topic modeling performances.

CONCLUSION

- The work leverages big data tools and provides an abstracted data collection layer which can be adapted to several data sources (twitter, Facebook posts, news articles, RSS fluxes, scraping modules).
- The proposed framework is scalable, ***big data*** ready, and has been applied on real-world covid-19 data.
- And I have been able to use the technologies taught in class (Dash, Spark, Elasticsearch, PostgreSQL, machine learning, *etc.*) to develop a prototype of the framework.

Université de Lille

ILIS

Thank you for your attention !

REFERENCES

Twitterscraper : https://pypi.org/project/twitterscraper/0.2.7/

GetOldTweets3 : https://pypi.org/project/GetOldTweets3/

Spacy : https://pypi.org/project/spacy/

Spark : https://spark.apache.org/