Big Data

Spark ML

Instructor: Trong-Hop Do

June 14th 2021

S3Lab
Smart Software System Laboratory


“Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”
– Chris Lynch, Vertica Systems

Spark Machine Learning

Spark MLlib NLP

MLlib: Main Guide

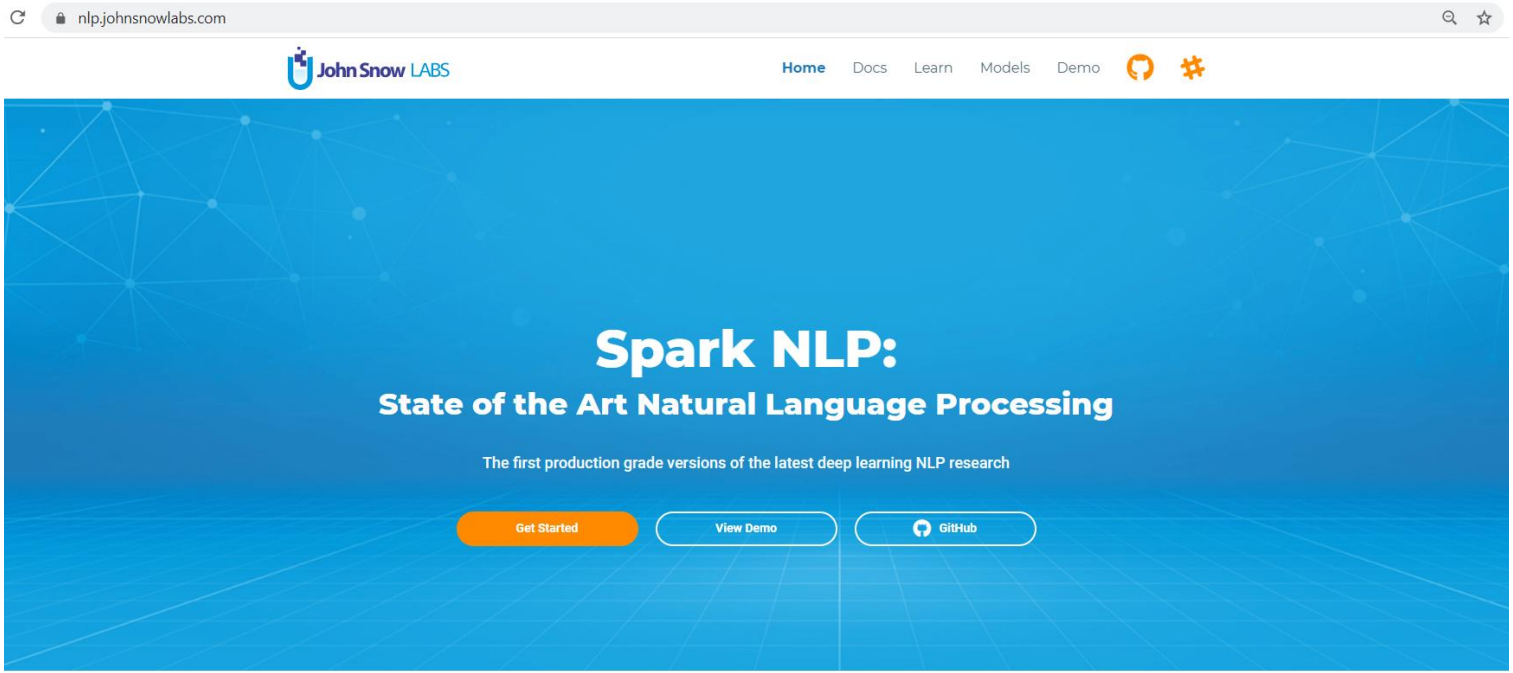
- Basic statistics
- Data sources
- Pipelines
- **Extracting, transforming and selecting features**
- Classification and Regressior
- Clustering
- Collaborative filtering
- Frequent Pattern Mining
- Model selection and tuning
- Advanced topics

«

- Feature Extractors
  - TF-IDF
  - Word2Vec
  - CountVectorizer
  - FeatureHasher
- Feature Transformers
  - Tokenizer
  - StopWordsRemover
  - n-gram
  - Binarizer
  - PCA
  - PolynomialExpansion
  - Discrete Cosine Transform (DCT)
  - StringIndexer
  - IndexToString
  - OneHotEncoder
  - VectorIndexer
  - Interaction
  - Normalizer
  - StandardScaler
  - RobustScaler
  - MinMaxScaler
  - MaxAbsScaler
  - Bucketizer
  - ElementwiseProduct
  - SQLTransformer
  - VectorAssembler
  - VectorSizeHint
  - QuantileDiscretizer
  - Imputer

Spark MLlib NLP

- John Snow LABS' Spark NLP



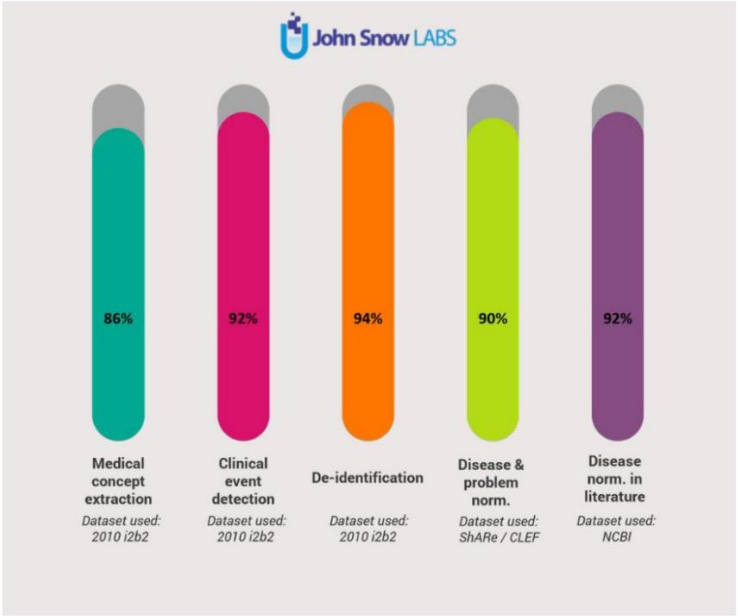
Spark MLlib NLP

- John Snow LABS' Spark NLP

Spark NLP is already in use in enterprise projects for various use cases (2019)

Spark MLlib NLP

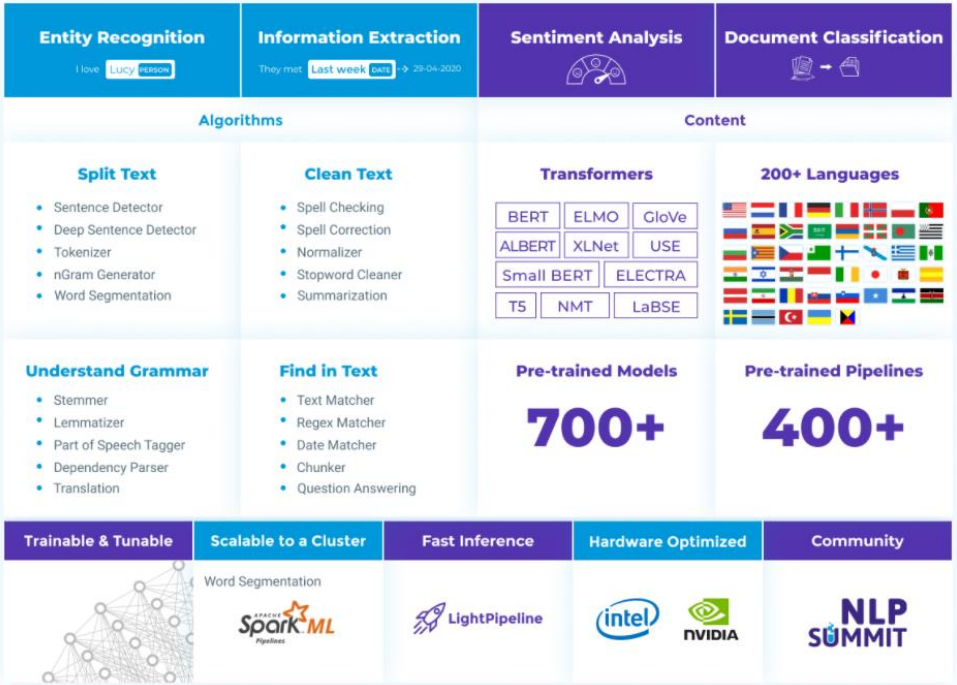
- John Snow LABS' Spark NLP



Spark NLP provides licensed annotators and models that are already trained by SOTA algorithms for Healthcare Analytics

Spark MLlib NLP

- John Snow LABS' Spark NLP



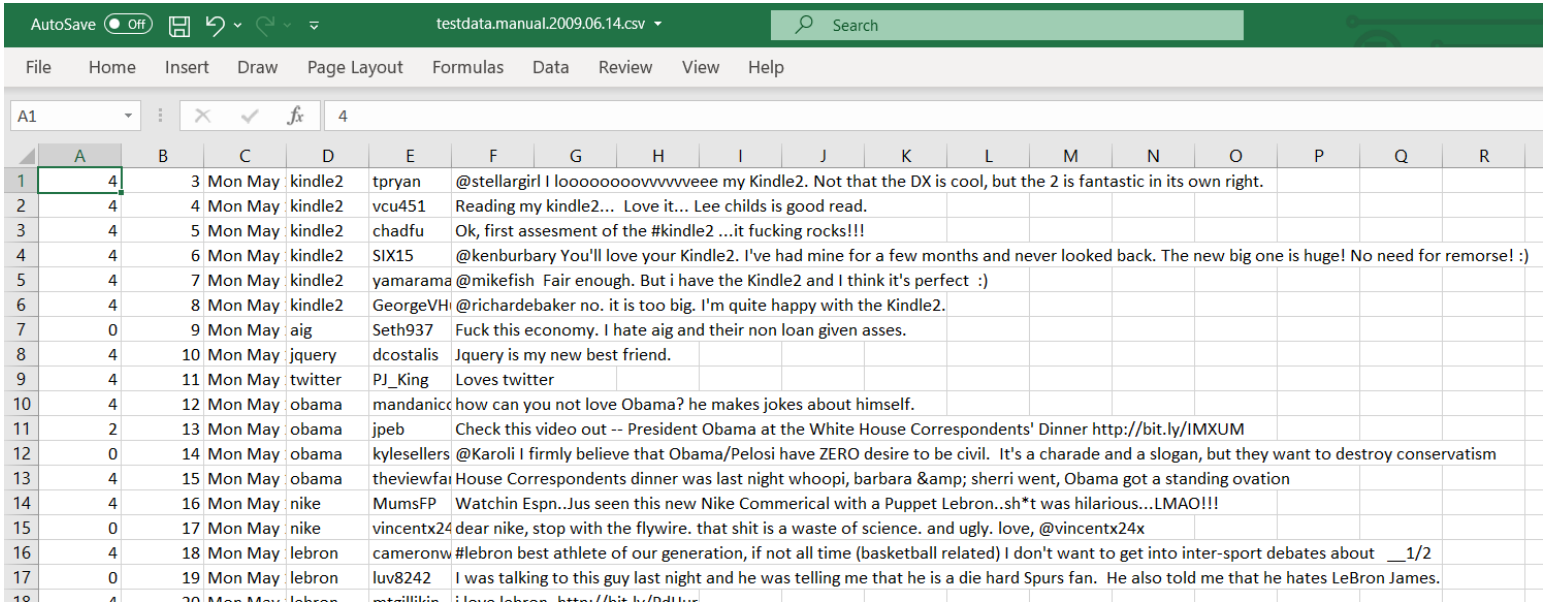
Spark MLlib NLP

Data exploration

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
from pyspark.sql import functions as f

import pandas as pd
from IPython.core.display import display
import seaborn as sns


spark = SparkSession.builder.getOrCreate()

schema = "polarity FLOAT, id LONG, date_time STRING, query STRING, user STRING, text STRING"
```

Spark MLlib NLP

Data exploration: test data



AutoSave Off | testdata.manual.2009.06.14.csv | Search

File Home Insert Draw Page Layout Formulas Data Review View Help

A1 | 4

	A	B	C	D	E	F
1	4	3	Mon May	kindle2	tpryan	@stellargirl I looooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right.
2	4	4	Mon May	kindle2	vcu451	Reading my kindle2... Love it... Lee childs is good read.
3	4	5	Mon May	kindle2	chadfu	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
4	4	6	Mon May	kindle2	SIX15	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
5	4	7	Mon May	kindle2	yamarama	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)
6	4	8	Mon May	kindle2	GeorgeVH	@richardebaker no. it is too big. I'm quite happy with the Kindle2.
7	0	9	Mon May	aig	Seth937	Fuck this economy. I hate aig and their non loan given asses.
8	4	10	Mon May	jquery	dcostalis	Jquery is my new best friend.
9	4	11	Mon May	twitter	PJ_King	Loves twitter
10	4	12	Mon May	obama	mandanicc	how can you not love Obama? he makes jokes about himself.
11	2	13	Mon May	obama	jpeb	Check this video out -- President Obama at the White House Correspondents' Dinner http://bit.ly/IMXUM
12	0	14	Mon May	obama	kylesellers	@Karoli I firmly believe that Obama/Pelosi have ZERO desire to be civil. It's a charade and a slogan, but they want to destroy conservatism
13	4	15	Mon May	obama	theviewfar	House Correspondents dinner was last night whoopi, barbara & sherri went, Obama got a standing ovation
14	4	16	Mon May	nike	MumsFP	Watchin Espn..Jus seen this new Nike Commerical with a Puppet Lebron..sh*t was hilarious...LMAO!!!
15	0	17	Mon May	nike	vincentx24	dear nike, stop with the flywire. that shit is a waste of science. and ugly. love, @vincentx24x
16	4	18	Mon May	lebron	cameronw	#lebron best athlete of our generation, if not all time (basketball related) I don't want to get into inter-sport debates about __1/2
17	0	19	Mon May	lebron	luv8242	I was talking to this guy last night and he was telling me that he is a die hard Spurs fan. He also told me that he hates LeBron James.

Spark MLlib NLP

Data exploration: test data

```
Overall data count: 498
```

	summary	polarity	id	date_time	query	user	text
0	count	498	498	498	498	498	498
1	mean	2.0200803212851404	1867.2269076305222	None	46.0	None	None
2	stddev	1.6996858490577658	2834.891681137318	None	5.163977794943222	None	None
3	min	0.0	3	Fri May 15 06:45:54 UTC 2009	"""booz allen"""	5x1llz	"""The Republican party is a bunch of anti-abortion zealots who couldn't draw flies to a dump."" -- Neal Boortz (just now
4	25%	0.0	388	None	40.0	None	None
5	50%	2.0	1013	None	50.0	None	None
6	75%	4.0	2367	None	50.0	None	None
7	max	4.0	14076	Wed May 27 23:59:18 UTC 2009	yankees	zedomax	zomg!!! I have a G2!!!!!!!

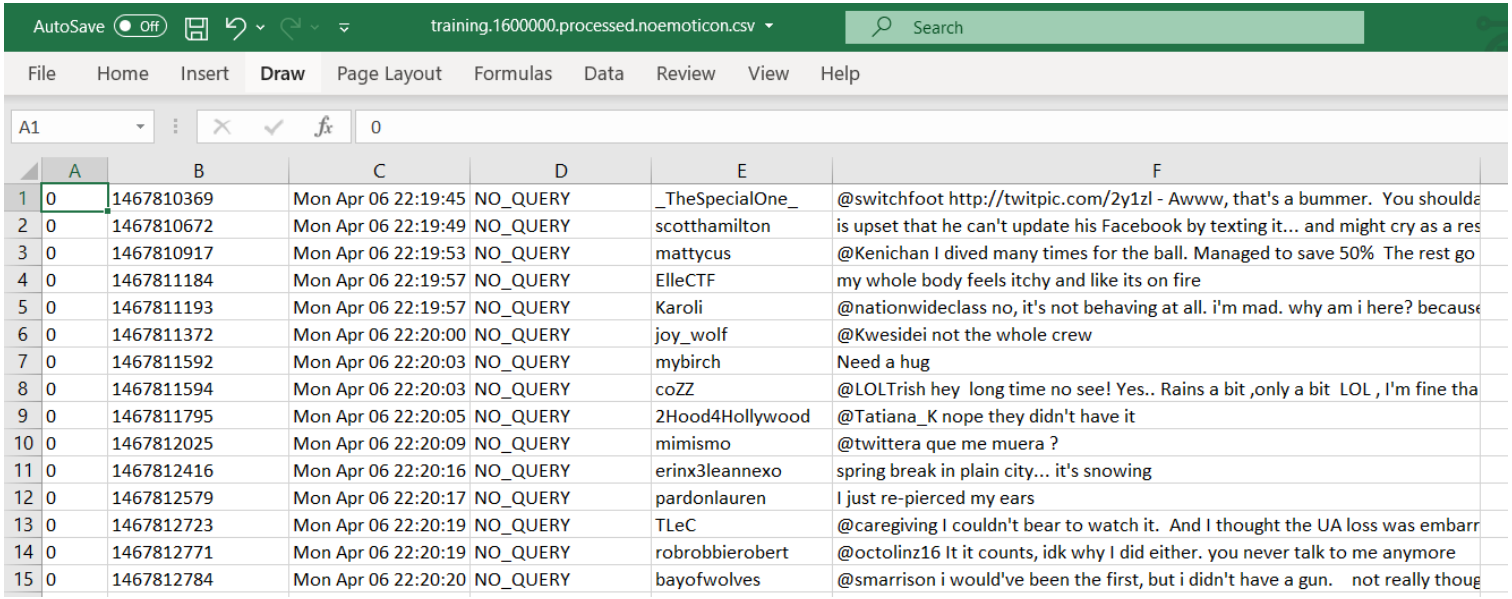
Spark MLlib NLP

Data exploration: test data

	polarity	id	date_time	query	user	text
0	4.0	3	Mon May 11 03:17:40 UTC 2009	kindle2	tpryan	@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right.
1	4.0	4	Mon May 11 03:18:03 UTC 2009	kindle2	vcu451	Reading my kindle2... Love it... Lee childs is good read.
2	4.0	5	Mon May 11 03:18:54 UTC 2009	kindle2	chadfu	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
3	4.0	6	Mon May 11 03:19:04 UTC 2009	kindle2	SIX15	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
4	4.0	7	Mon May 11 03:21:41 UTC 2009	kindle2	yamarama	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)
5	4.0	8	Mon May 11 03:22:00 UTC 2009	kindle2	GeorgeVHulme	@richardebaker no. it is too big. I'm quite happy with the Kindle2.

Spark MLlib NLP

Data exploration: train data



	A	B	C	D	E	F
1	0	1467810369	Mon Apr 06 22:19:45	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda
2	0	1467810672	Mon Apr 06 22:19:49	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a res
3	0	1467810917	Mon Apr 06 22:19:53	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go
4	0	1467811184	Mon Apr 06 22:19:57	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
5	0	1467811193	Mon Apr 06 22:19:57	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because
6	0	1467811372	Mon Apr 06 22:20:00	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
7	0	1467811592	Mon Apr 06 22:20:03	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine tha
9	0	1467811795	Mon Apr 06 22:20:05	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
10	0	1467812025	Mon Apr 06 22:20:09	NO_QUERY	mimismo	@twittera que me muera ?
11	0	1467812416	Mon Apr 06 22:20:16	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
12	0	1467812579	Mon Apr 06 22:20:17	NO_QUERY	pardonlauren	I just re-pierced my ears
13	0	1467812723	Mon Apr 06 22:20:19	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarr
14	0	1467812771	Mon Apr 06 22:20:19	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore
15	0	1467812784	Mon Apr 06 22:20:20	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really thoug

Spark MLlib NLP

Data exploration: train data

```
Overall data count: 1600000
```

	summary	polarity	id	date_time	query	user	text
0	count	1600000	1600000	1600000	1600000	1600000	1600000
1	mean	2.0	1.99881755229562 76E9	None	None	4.325887521835714E9	None
2	stddev	2.000000625000293	1.93576073622678 97E8	None	None	5.16273321845489E10	None
3	min	0.0	1467810369	Fri Apr 17 20:30:31 PDT 2009	NO_QUERY	000catnap000	exhausted
4	25%	0.0	1956912114	None	None	32508.0	None
5	50%	0.0	2002096128	None	None	130587.0	None
6	75%	4.0	2177066219	None	None	1100101.0	None
7	max	4.0	2329205794	Wed May 27 07:27:38 PDT 2009	NO_QUERY	zzzzeus111	����ᵠ��ó������ж���� <<----I DID NOT KNOW I CUD or HOW TO DO ALL DAT ON MY PHONE TIL NOW. WOW..MY LIFE IS NOW COMPLETE. JK.

Spark MLlib NLP

Data exploration: train data

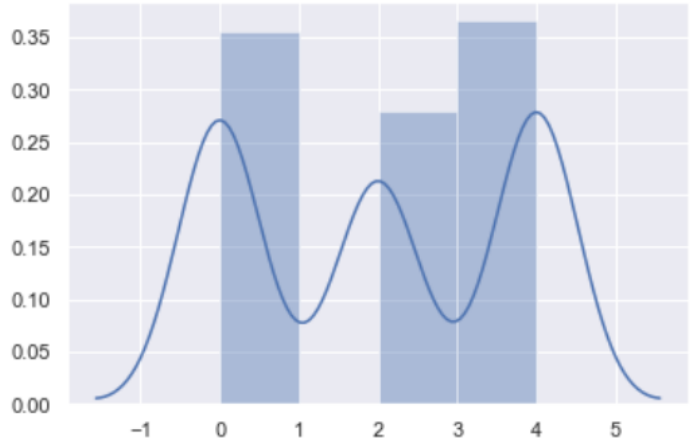
	polarity	id	date_time	query	user	text
0	0.0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
1	0.0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
2	0.0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
3	0.0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0.0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
5	0.0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
6	0.0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug

Spark MLlib NLP

Data exploration: distribution of polarity

```
df = raw_test_data.select("polarity").na.drop()
print(f"No of rows with Polarity: {df.count()}/{raw_test_data.count()}")

sns.distplot(df.toPandas())
```

```
No of rows with Polarity: 498/498
<matplotlib.axes._subplots.AxesSubplot at 0x26b618bd1d0>
```

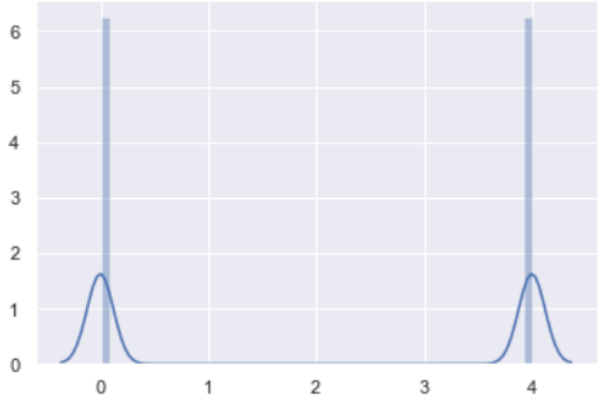


Spark MLlib NLP

Data exploration: distribution of polarity

```
df = raw_training_data.select("polarity").na.drop()
print(f"No of rows with Polarity: {df.count()} / {raw_training_data.count()}")

sns.distplot(df.toPandas())
```

```
No of rows with Polarity: 1600000 / 1600000
<matplotlib.axes._subplots.AxesSubplot at 0x26b61bc9358>
```




```
polarity_df = raw_training_data.select("polarity").cache()

polarity_df.groupBy("polarity").count().toPandas()
```

	polarity	count
0	4.0	800000
1	0.0	800000

Spark MLlib NLP

Data exploration

Now it's time for us to write the raw data we intend to use to disk.
We're going to:

- keep the format CSV
- partition the data by polarity, this will create 2 subfolders inside our output folder
- repartition the data in 20 partitions: This will ensure that we have 20 smaller csv files per partition

Spark MLlib NLP

Data cleaning - Converting Date column

```
spark.sql("set spark.sql.legacy.timeParserPolicy=LEGACY")

schema_ddl = " polarity STRING, id LONG, date TIMESTAMP, query STRING , user string, text string "
spark_reader = spark.read.schema(schema_ddl)

simple_date_format = "EEE MMM dd HH:mm:ss zzz yyyy"

raw_data = spark_reader.csv(RAW_PATH, timestampFormat=simple_date_format)
raw_data.show(10)
raw_data.printSchema()

raw_data.summary().show()
```

Spark MLlib NLP

Data cleaning - Converting Date column

```
+----------+-------------------+--------+---------+--------------------+--------+
|        id|               date|   query|     user|                text|polarity|
+----------+-------------------+--------+---------+--------------------+--------+
|1833543437|2009-05-18 13:37:41|NO_QUERY|  Kbrodes|Hi all. Have woke...|     0.0|
|1692704388|2009-05-04 10:35:20|NO_QUERY|spentwent|@kristindoll YOU ...|     0.0|
+----------+-------------------+--------+---------+--------------------+--------+
only showing top 2 rows

root
 |-- id: long (nullable = true)
 |-- date: timestamp (nullable = true)
 |-- query: string (nullable = true)
 |-- user: string (nullable = true)
 |-- text: string (nullable = true)
 |-- polarity: string (nullable = true)
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

- Remove email-addresses and URLs
- Extract and then remove username (@mentions)
- Extract and then remove hastag (#hash-tag)

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
# extract up to 6 twitter user names/handles to the output column `users_mentioned`

user_regex = r"(@\w{1,15})"

raw_data.select(
    f.array_remove(
        f.array(
            f.regexp_extract(f.col("text"), user_regex, 1),
            f.regexp_extract(
                f.col("text"), "".join([f"{user_regex}.*?" for i in range(0, 2)]), 2
            ),
            f.regexp_extract(
                f.col("text"), "".join([f"{user_regex}.*?" for i in range(0, 3)]), 3
            ),
            f.regexp_extract(
                f.col("text"), "".join([f"{user_regex}.*?" for i in range(0, 4)]), 4
            ),
            f.regexp_extract(
                f.col("text"), "".join([f"{user_regex}.*?" for i in range(0, 5)]), 5
            ),
            f.regexp_extract(
                f.col("text"), "".join([f"{user_regex}.*?" for i in range(0, 6)]), 6
            ),
        ),
        "",
    ).alias("users_mentioned"),
    "text",
).toPandas().head(35)
```

	users_mentioned	text
0	[]	Hi all. Have woken up with a cold. Boo! No tim...
1	[@kristindoll]	@kristindoll YOU NEVER ANSER ME
2	[]	Craaaaamps
3	[]	Our human mom just called. Her brand new Pruis...
4	[]	stomach growling. would give anything to be ab...
5	[]	I miss the sea, nak gi diving badly....but no ...
6	[@gradytwin]	@gradytwin i wish it was true for me too
7	[]	Why is it so hot??
8	[]	I need a hug
9	[]	Just saw a guy with the same (except black) &q...
10	[]	doinnnn' some hmwk, projects, studying, what a...
11	[]	Gosh there is nothing on tv
12	[@rrunyan]	I'm ignoring @rrunyan this morning because it'...
13	[@lesley007]	@lesley007 scales are evil. FACT!!! xxx

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
raw_data.select(
    f.regexp_replace(f.col("text"), user_regex, "").alias("text"),
    f.col("text").alias("original_text"),
).toPandas().head(20)
```

	text	original_text
0	Hi all. Have woken up with a cold. Boo! No tim...	Hi all. Have woken up with a cold. Boo! No tim...
1	YOU NEVER ANSER ME	@kristindoll YOU NEVER ANSER ME
2	Craaaaamps	Craaaaamps
3	Our human mom just called. Her brand new Pruis...	Our human mom just called. Her brand new Pruis...
4	stomach growling. would give anything to be ab...	stomach growling. would give anything to be ab...
5	I miss the sea, nak gi diving badly....but no ...	I miss the sea, nak gi diving badly....but no ...
6	i wish it was true for me too	@gradytwin i wish it was true for me too
7	Why is it so hot??	Why is it so hot??
8	I need a hug	I need a hug
9	Just saw a guy with the same (except black) &q...	Just saw a guy with the same (except black) &q...

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

text	hashtags
@JasonCalacanis hey! we have to wait 'til sunday in the UK for the finale #lost	[#lost]
... and then something mega urgent and mega important gatecrashes the list #GTD	[#GTD]
Darn, my earphone cable snapped. Can't listen to music on my #iPhone while walking	[#iPhone]
wow #revision3.com is one big ad	[#revision3]
@AlexLJ alas I think me may have the see flu that be going round this fine ship arhhh #Twittarrr	[#Twittarrr]
@ScottMonty wishing we had an official #SXSEMIA Ford Car. Stuck with crappy Dodge Charger rental.	[#SXSEMIA]
grr, latelatelateeee! i have to shower, then get ready and leave at 10:15 to get into school. i FU...	[#mcfly]
@desertsong1 i only said #shitstack will be over because it's nearly 3am and i need sleep before ...	[#shitstack]
Where's the sunshine gone!???? #fb	[#fb]
still sad the #Mavs season is over. someone cheer me up	[#Mavs]
House small biz meeting not on CSPAN online #CPSIA	[#CPSIA]
@CXXG good night! #fixreplies #twitterfail #fixreplies #twitterfail #fixreplies #twitterfail #f...	[#fixreplies, #twitterfail]
@gerritv What's more, we get a slew of hideous adverts every 23 seconds. The race is as much adv...	null
@robbarry @bugabundo We'll soon find out if #ubuntu uses one core or not on the PS3 soon. I used ...	[#ubuntu]
@machineplay I'm so sorry you're having to go through this. Again. #therapyfail	[#therapyfail]

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
hashtag_replace_regex = "#(\w{1,})"

_.select(f.regexp_replace(f.col("text"), hashtag_replace_regex, "$1"), "hashtags").show(35, 100)
```

```
+----------------------------------------------------------------------------------------------------+----------------------------+
|                                                                    regexp_replace(text, #(\w{1,}), $1)|                    hashtags|
+----------------------------------------------------------------------------------------------------+----------------------------+
|                     @JasonCalacanis hey! we have to wait 'til sunday in the UK for the finale  lost|                     [#lost]|
|                     ... and then something mega urgent and mega important gatecrashes the list  GTD|                      [#GTD]|
|                   Darn, my earphone cable snapped. Can't listen to music on my iPhone while walking|                   [#iPhone]|
|                                                                 wow revision3.com is one big ad|                [#revision3]|
|    @AlexLJ alas I think me may have the see flu that be going round this fine ship arhhh  Twittarrr|                [#Twittarrr]|
|    @ScottMonty wishing we had an official SXSEMIA Ford Car. Stuck with crappy Dodge Charger rental.|                 [#SXSEMIA]|
|grr, latelatelateeee! i have to shower, then get ready and leave at 10:15 to get into school. i FU...|                   [#mcfly]|
|@desertsong1 i only said shitstack will be over because it's nearly 3am and i need sleep before u...|                [#shitstack]|
|                                                          Where's the sunshine gone!????  fb|                        [#fb]|
|                                        still sad the Mavs season is over. someone cheer me up|                      [#Mavs]|
|                                            House small biz meeting not on CSPAN online CPSIA|                     [#CPSIA]|
|@CXXG  good night!  fixreplies twitterfail fixreplies twitterfail fixreplies twitterfail fixrepli...|[#fixreplies, #twitterfail]|
|@gerritv What's more, we get a slew of hideous adverts every 23 seconds.  The race is as much adv...|                        null|
|@robbarry @bugabundo We'll soon find out if ubuntu uses one core or not on the PS3 soon. I used a...|                     [#ubuntu]|
|                     @machineplay I'm so sorry you're having to go through this. Again.  therapyfail|              [#therapyfail]|
|(via @etanowitz) There is an @orlandotweetup photo gallery http://tr.im/kf7h otweet   OH YEAH PIC...|                     [#otweet]|
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
url_regex=r"((https?|ftp|file):\/{2,3})+([-\w+&@#/%=~|$?!:,.]*)|(www.)+([-\w+&@#/%=~|$?!:,.]*)"
email_regex=r"[\w.-]+@[\w.-]+\.[a-zA-Z]{1,}"
```

```
raw_data.select(
    f.regexp_replace(f.col("text"), email_regex, "").alias("text_no_email"),
    f.regexp_replace(f.col("text"), url_regex, "").alias("text_no_url"),
    f.col("text").alias("original_text"),
).toPandas().head(20)
```

	text_no_email	text_no_url	original_text
0	Hi all. Have woken up with a cold. Boo! No tim...	Hi all. Have woken up with a cold. Boo! No tim...	Hi all. Have woken up with a cold. Boo! No tim...
1	@kristindoll YOU NEVER ANSER ME	@kristindoll YOU NEVER ANSER ME	@kristindoll YOU NEVER ANSER ME
2	Craaaaamps	Craaaaamps	Craaaaamps
3	Our human mom just called. Her brand new Pruis...	Our human mom just called. Her brand new Pruis...	Our human mom just called. Her brand new Pruis...
4	stomach growling. would give anything to be ab...	stomach growling. would give anything to be ab...	stomach growling. would give anything to be ab...
5	I miss the sea, nak gi diving badly....but no ...	I miss the sea, nak gi diving badly....but no ...	I miss the sea, nak gi diving badly....but no ...
6	@gradytwin i wish it was true for me too	@gradytwin i wish it was true for me too	@gradytwin i wish it was true for me too
7	Why is it so hot??	Why is it so hot??	Why is it so hot??
8	I need a hug	I need a hug	I need a hug
9	Just saw a guy with the same (except black) &q...	Just saw a guy with the same (except black) &q...	Just saw a guy with the same (except black) &q...

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
started to think that Citi is in really deep s&amp;^t. Are they gonna survive the turmoil or are they gonna be the next AIG?
"I'm listening to ""P.Y.T"" by Danny Gokey &lt;3 &lt;3 &lt;3 Aww
```

```python
from pyspark.sql.functions import udf
import html

@udf
def html_unescape(s: str):
    return html.unescape(s)

raw_data.select(html_unescape("text")).show(35, 150)
```

```
+------------------------------------------------------------------------------------------------------+
|                                                                                     html_unescape(text)|
+------------------------------------------------------------------------------------------------------+
|              Hi all. Have woken up with a cold. Boo! No time to feel sorry for myself. My family are coming today. Hot lemon for me.|
|                                                                          @kristindoll YOU NEVER   ANSER ME|
|                                                                                              Craaaaamps|
|                                                   Our human mom just called. Her brand new Pruis just died on the causeway!!|
|              stomach growling. would give anything to be able to open mouth wide and chomp down on a burger, or pizza, or even rice!|
|                                                                     I miss the sea, nak gi diving badly....but no cuti|
|                                                                             @gradytwin i wish it was true for me too|
|                                                                                                  Why is it so hot??|
|                                                                                                         I need a hug|
|                                  Just saw a guy with the same (except black) "toppu" t-shirt as I have. I'm no longer unique!|
|                                                      doinnnn' some hmwk, projects, studying, what a nice way to end this long weekend|
|                                                                                          Gosh there is nothing on tv|
|                I'm ignoring @rrunyan this morning because it's Sunday and that's his weigh-in day...and he didn't do well this week.|
|                                                                             @lesley007 scales are evil. FACT!!!  xxx|
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
raw_data = spark.read.schema(schema).csv(RAW_PATH)
clean_data = cleaning_process(raw_data)
clean_data.show()
clean_data.select("text").show(50, False)
```

```
+----------+--------------------+--------+--------------+--------------------+--------+--------------+--------+--------------------+
|        id|                date|   query|          user|                text|polarity|users_mentioned|hashtags|       original_text|
+----------+--------------------+--------+--------------+--------------------+--------+--------------+--------+--------------------+
|1833543437|Sun May 17 23:37:...|NO_QUERY|        Kbrodes|Hi all. Have woke...|     0.0|          null|    null|Hi all. Have woke...|
|1692704388|Sun May 03 20:35:...|NO_QUERY|      spentwent| YOU NEVER   ANSE...|     0.0| [@kristindoll]|    null|@kristindoll YOU ...|
|1677671697|Sat May 02 03:21:...|NO_QUERY|      Honey_Nut|          Craaaaamps|     0.0|          null|    null|          Craaaaamps|
|1573908755|Tue Apr 21 03:25:...|NO_QUERY|dirtydogsofsobe|Our human mom jus...|     0.0|          null|    null|Our human mom jus...|
|1836234330|Mon May 18 07:37:...|NO_QUERY|       chiewmei|stomach growling....|     0.0|          null|    null|stomach growling....|
|1833057837|Sun May 17 22:10:...|NO_QUERY|     tini_hotfm|I miss the sea, n...|     0.0|          null|    null|I miss the sea, n...|
|1695573341|Mon May 04 06:27:...|NO_QUERY|      D3vouring| i wish it was tr...|     0.0|   [@gradytwin]|    null|@gradytwin i wish...|
|1834138334|Mon May 18 01:46:...|NO_QUERY|    jessyflores|  Why is it so hot??|     0.0|          null|    null|  Why is it so hot??|
|1760069887|Sun May 10 20:35:...|NO_QUERY|  DarianFroseth|        I need a hug|     0.0|          null|    null|        I need a hug|
|1825085705|Sun May 17 04:16:...|NO_QUERY|         mmazur|Just saw a guy wi...|     0.0|          null|    null|Just saw a guy wi...|
|1836140621|Mon May 18 07:27:...|NO_QUERY|     meljonasxo|doinnnn' some hmw...|     0.0|          null|    null|doinnnn' some hmw...|
|1556303497|Sat Apr 18 22:20:...|NO_QUERY|    Bridgetgarz|Gosh there is not...|     0.0|          null|    null|Gosh there is not...|
|1686547885|Sun May 03 05:53:...|NO_QUERY|     willow1999|I'm ignoring  thi...|     0.0|     [@rrunyan]|    null|I'm ignoring @rru...|
|1879980021|Thu May 21 23:43:...|NO_QUERY|    iSlayer2009| scales are evil....|     0.0|   [@lesley007]|    null|@lesley007 scales...|
|1826030215|Sun May 17 07:31:...|NO_QUERY|      gaminette|omg sinus infecti...|     0.0|          null|    null|omg sinus infecti...|
```

Spark MLlib NLP

Data exploration

```
+----------------------------------------------------------------------------------------------------------------------------------------------------------------+
|text                                                                                                                                                            |
+----------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Hi all. Have woken up with a cold. Boo! No time to feel sorry for myself. My family are coming today. Hot lemon for me.                                         |
| YOU NEVER   ANSER ME                                                                                                                                           |
|Craaaaamps                                                                                                                                                      |
|Our human mom just called. Her brand new Pruis just died on the causeway!!                                                                                      |
|stomach growling. would give anything to be able to open mouth wide and chomp down on a burger, or pizza, or even rice!                                        |
|I miss the sea, nak gi diving badly....but no cuti                                                                                                              |
| i wish it was true for me too                                                                                                                                  |
|Why is it so hot??                                                                                                                                              |
|I need a hug                                                                                                                                                    |
|Just saw a guy with the same (except black) "toppu" t-shirt as I have. I'm no longer unique!                                                                    |
|doinnnn' some hmwk, projects, studying, what a nice way to end this long weekend                                                                                |
|Gosh there is nothing on tv                                                                                                                                     |
|I'm ignoring  this morning because it's Sunday and that's his weigh-in day...and he didn't do well this week.                                                  |
| scales are evil. FACT!!!  xxx                                                                                                                                  |
|omg sinus infection, so that was you lurking behind yesterday's headache. just don't stay too long okay?                                                       |
|alone and sad                                                                                                                                                   |
|Working until close tonight. I'm going to miss the Hell's Kitchen finale...nooo                                                                                |
|i miss my ugly wight!!!                                                                                                                                         |
|I think I lost my best friend today...feeling blue                                                                                                              |
| Its also knows as glandular fever () I got ill about 2 weeks ago, still feeling miserable                                                                      |
|Gahh, I just spilt hot chocolate all down my top and burnt myslef a bit                                                                                         |
| hiiiiiiiiiiiiiii!!! i am following u! LOL ps. i miss u guys and my neighbor! LOL dark knight isnt following me on twitter!                                     |
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
raw_data.count()
```

```
1600000
```

```
clean_data.count()
```

```
1600000
```

```
clean_data.filter("text == ''").show(1000)
```

```
+----------+--------------------+--------+---------------+----+--------+------------------+--------+--------------------+
|        id|                date|   query|           user|text|polarity|   users_mentioned|hashtags|       original_text|
+----------+--------------------+--------+---------------+----+--------+------------------+--------+--------------------+
|1823933200|Sat May 16 23:27:...|NO_QUERY|     NOTjanelle|    |     0.0| [@chriswantsfood]|    null|     @chriswantsfood|
|1753166696|Sun May 10 00:24:...|NO_QUERY|   Sweet_Candii|    |     0.0| [@Stealth_Tricia]|    null|     @Stealth_Tricia|
|1794571767|Thu May 14 06:06:...|NO_QUERY|     msfitznham|    |     0.0|     [@demongirly]|    null|         @demongirly|
|1693024371|Sun May 03 21:21:...|NO_QUERY|       juuleeya|    |     0.0|   [@fanficaholic]|    null|       @fanficaholic|
|1823473643|Sat May 16 22:04:...|NO_QUERY|michelle_dunlap|    |     0.0| [@smoulderingsea]|    null|     @smoulderingsea|
|1556285000|Sat Apr 18 22:16:...|NO_QUERY|  xcassiegottox|    |     0.0|     [@staticxage]|    null|         @staticxage|
|1565687892|Mon Apr 20 07:20:...|NO_QUERY|     7arfal3ain|    |     0.0|     [@pearly_uae]|    null|         @pearly_uae|
|1834489373|Mon May 18 03:07:...|NO_QUERY|         rehmxo|    |     0.0|   [@shaunjumpnow]|    null|       @shaunjumpnow|
|1468112539|Mon Apr 06 23:46:...|NO_QUERY|    MissPassion|    |     0.0|  [@thecoolestout]|    null|      @thecoolestout|
|1956202634|Thu May 28 21:23:...|NO_QUERY|       BLAK_OUT|    |     0.0|   [@shortyjunior]|    null|       @shortyjunior|
|1978920897|Sun May 31 00:56:...|NO_QUERY|      desplesda|    |     0.0|    [@TheRealBnut]|    null|        @TheRealBnut|
|1982409707|Sun May 31 11:09:...|NO_QUERY|          lpt21|    |     0.0|    [@The_Brew_Co]|    null|        @The_Brew_Co|
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
df = (
    df_clean
    # Remove all numbers
    .withColumn("text", f.regexp_replace(f.col("text"), "[^a-zA-Z']", " "))
    # Remove all double/multiple spaces
    .withColumn("text", f.regexp_replace(f.col("text"), " +", " "))
    # Remove leading and trailing whitespaces
    .withColumn("text", f.trim(f.col("text")))
    # Ensure we don't end up with empty rows
    .filter("text != ''")
)

data = df.select("text", "polarity").coalesce(3).cache()
```

Spark MLlib NLP

Data cleaning - Cleaning the tweet text

```
print(df_clean.count())
print(df.count())
```

```
1600000
1596232
```

```
df.toPandas()
```

	polarity	id	date_time	query	user	text	original_text
0	4.0	3	2009-05-11 03:17:40	kindle2	tpryan	I looooooooovvvvveee my Kindle Not that the DX is cool but the is fantastic in its own right	@stellargirl I looooooooovvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right.
1	4.0	4	2009-05-11 03:18:03	kindle2	vcu451	Reading my kindle Love it Lee childs is good read	Reading my kindle2... Love it... Lee childs is good read.
2	4.0	5	2009-05-11 03:18:54	kindle2	chadfu	Ok first assesment of the kindle it fucking rocks	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
3	4.0	6	2009-05-11 03:19:04	kindle2	SIX15	You'll love your Kindle I've had mine for a few months and never looked back The new big one is huge No need for remorse	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
4	4.0	7	2009-05-11 03:21:41	kindle2	yamarama	Fair enough But i have the Kindle and I think it's perfect	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)
...	...	...	...	...	...	...	...
493	2.0	14072	2009-06-14 04:31:43	latex	proggit	Ask Programming LaTeX or InDesign submitted by calcio link comment	Ask Programming: LaTeX or InDesign?: submitted by calcio1 [link] [1 comment] http://tinyurl.com/myfmf7
494	0.0	14073	2009-06-14 04:32:17	latex	sam33r	On that note I hate Word I hate Pages I hate LaTeX There I said it I hate LaTeX All you TEXN RDS can come kill me now	On that note, I hate Word. I hate Pages. I hate LaTeX. There, I said it. I hate LaTeX. All you TEXN3RDS can come kill me now.
495	4.0	14074	2009-06-14 04:36:34	latex	iamtheonlyjosie	Ahhh back in a real text editing environment I LaTeX	Ahhh... back in a *real* text editing environment. I <3 LaTeX.
496	0.0	14075	2009-06-14 21:36:07	iran	plutopup7	Trouble in Iran I see Hmm Iran Iran so far away flockofseagullsweregeopoliticallycorrect	Trouble in Iran, I see. Hmm. Iran. Iran so far away. #flockofseagullsweregeopoliticallycorrect
497	0.0	14076	2009-06-14 21:36:17	iran	captain_pete	Reading the tweets coming out of Iran The whole thing is terrifying and incredibly sad	Reading the tweets coming out of Iran... The whole thing is terrifying and incredibly sad...

Spark MLlib NLP

Building model

```
(training_data, validation_data, test_data) = data.randomSplit([0.98, 0.01, 0.01], seed=2020)
```

+ Code + Markdown

```
%%time
from pyspark.ml.feature import (
    StopWordsRemover,
    Tokenizer,
    HashingTF,
    IDF,
)
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline

tokenizer = Tokenizer(inputCol="text", outputCol="words1")
stopwords_remover = StopWordsRemover(
    inputCol="words1",
    outputCol="words2",
    stopWords=StopWordsRemover.loadDefaultStopWords("english")
)
hashing_tf = HashingTF(
    inputCol="words2",
    outputCol="term_frequency",
)
idf = IDF(
    inputCol="term_frequency",
    outputCol="features",
    minDocFreq=5,
)
lr = LogisticRegression(labelCol="polarity")

semantic_analysis_pipeline = Pipeline(
    stages=[tokenizer, stopwords_remover, hashing_tf, idf, lr]
)
```

Spark MLlib NLP

Building model

```
df1 = tokenizer.transform(validation_data)
df1.show()
```

```
+--------------------+--------+--------------------+
|                text|polarity|              words1|
+--------------------+--------+--------------------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|
|'Honey' the chick...|     0.0|['honey', the, ch...|
|'What Canadians H...|     4.0|['what, canadians...|
|'allo Davina welc...|     4.0|['allo, davina, w...|
|'course it's not ...|     4.0|['course, it's, n...|
|'s bus just broke...|     0.0|['s, bus, just, b...|
|'s happy thought ...|     0.0|['s, happy, thoug...|
|'s screen size is...|     0.0|['s, screen, size...|
|'s voice makes me...|     4.0|['s, voice, makes...|
|'tis ok I posted ...|     4.0|['tis, ok, i, pos...|
|A BIG Welcome to ...|     4.0|[a, big, welcome,...|
|      A BOMB INATION|     4.0|  [a, bomb, ination]|
|A Demi I hope no ...|     0.0|[a, demi, i, hope...|
|A Don't worry abo...|     4.0|[a, don't, worry,...|
|A GIRL IS SO LUCK...|     0.0|[a, girl, is, so,...|
|A Gabrielle wishe...|     0.0|[a, gabrielle, wi...|
|A How babies just...|     4.0|[a, how, babies, ...|
|A I am sorry abou...|     0.0|[a, i, am, sorry,...|
|A I hella miss yo...|     0.0|[a, i, hella, mis...|
|A I hope they cal...|     0.0|[a, i, hope, they...|
+--------------------+--------+--------------------+
only showing top 20 rows
```

Spark MLlib NLP

Building model

```
df2 = stopwords_remover.transform(df1)
df2.show()
```

```
+--------------------+--------+--------------------+--------------------+
|                text|polarity|              words1|              words2|
+--------------------+--------+--------------------+--------------------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|['ayye, pass, mon...|
|'Honey' the chick...|     0.0|['honey', the, ch...|['honey', chicken...|
|'What Canadians H...|     4.0|['what, canadians...|['what, canadians...|
|'allo Davina welc...|     4.0|['allo, davina, w...|['allo, davina, w...|
|'course it's not ...|     4.0|['course, it's, n...|['course, quantit...|
|'s bus just broke...|     0.0|['s, bus, just, b...|    ['s, bus, broke]|
|'s happy thought ...|     0.0|['s, happy, thoug...|['s, happy, thoug...|
|'s screen size is...|     0.0|['s, screen, size...|['s, screen, size...|
|'s voice makes me...|     4.0|['s, voice, makes...|['s, voice, makes...|
|'tis ok I posted ...|     4.0|['tis, ok, i, pos...|['tis, ok, posted...|
|A BIG Welcome to ...|     4.0|[a, big, welcome,...|[big, welcome, ne...|
|      A BOMB INATION|     4.0|  [a, bomb, ination]|     [bomb, ination]|
|A Demi I hope no ...|     0.0|[a, demi, i, hope...|[demi, hope, one,...|
|A Don't worry abo...|     4.0|[a, don't, worry,...|[worry, hun, wors...|
|A GIRL IS SO LUCK...|     0.0|[a, girl, is, so,...|[girl, lucky, won...|
|A Gabrielle wishe...|     0.0|[a, gabrielle, wi...|[gabrielle, wishe...|
|A How babies just...|     4.0|[a, how, babies, ...|[babies, light, g...|
|A I am sorry abou...|     0.0|[a, i, am, sorry,...|             [sorry]|
|A I hella miss yo...|     0.0|[a, i, hella, mis...|       [hella, miss]|
|A I hope they cal...|     0.0|[a, i, hope, they...|[hope, calm, fun,...|
+--------------------+--------+--------------------+--------------------+
only showing top 20 rows
```

Spark MLlib NLP

Building model

```
df3 = hashing_tf.transform(df2)
df3.show()
```

```
+--------------------+--------+--------------------+--------------------+--------------------+
|                text|polarity|              words1|              words2|      term_frequency|
+--------------------+--------+--------------------+--------------------+--------------------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|['ayye, pass, mon...|(262144,[31536,64...|
|'Honey' the chick...|     0.0|['honey', the, ch...|['honey', chicken...|(262144,[55627,64...|
|'What Canadians H...|     4.0|['what, canadians...|['what, canadians...|(262144,[16415,10...|
|'allo Davina welc...|     4.0|['allo, davina, w...|['allo, davina, w...|(262144,[1512,124...|
|'course it's not ...|     4.0|['course, it's, n...|['course, quantit...|(262144,[43890,70...|
|'s bus just broke...|     0.0|['s, bus, just, b...|    ['s, bus, broke]|(262144,[91694,92...|
|'s happy thought ...|     0.0|['s, happy, thoug...|['s, happy, thoug...|(262144,[12409,29...|
|'s screen size is...|     0.0|['s, screen, size...|['s, screen, size...|(262144,[92492,14...|
|'s voice makes me...|     4.0|['s, voice, makes...|['s, voice, makes...|(262144,[92492,10...|
|'tis ok I posted ...|     4.0|['tis, ok, i, pos...|['tis, ok, posted...|(262144,[21894,64...|
|A BIG Welcome to ...|     4.0|[a, big, welcome,...|[big, welcome, ne...|(262144,[64358,10...|
|      A BOMB INATION|     4.0|  [a, bomb, ination]|     [bomb, ination]|(262144,[26648,66...|
|A Demi I hope no ...|     0.0|[a, demi, i, hope...|[demi, hope, one,...|(262144,[21823,61...|
|A Don't worry abo...|     4.0|[a, don't, worry,...|[worry, hun, wors...|(262144,[117975,1...|
|A GIRL IS SO LUCK...|     0.0|[a, girl, is, so,...|[girl, lucky, won...|(262144,[13781,33...|
|A Gabrielle wishe...|     0.0|[a, gabrielle, wi...|[gabrielle, wishe...|(262144,[991,2078...|
|A How babies just...|     4.0|[a, how, babies, ...|[babies, light, g...|(262144,[109208,2...|
|A I am sorry abou...|     0.0|[a, i, am, sorry,...|            [sorry]|(262144,[144961],...|
|A I hella miss yo...|     0.0|[a, i, hella, mis...|       [hella, miss]|(262144,[197515,2...|
|A I hope they cal...|     0.0|[a, i, hope, they...|[hope, calm, fun,...|(262144,[23087,46...|
+--------------------+--------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Spark MLlib NLP

Building model

```
df4 = idf.fit(df3).transform(df3)
df4.show()
```

```
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+
|                text|polarity|              words1|              words2|      term_frequency|            features|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|['ayye, pass, mon...|(262144,[31536,64...|(262144,[31536,64...|
|'Honey' the chick...|     0.0|['honey', the, ch...|['honey', chicken...|(262144,[55627,64...|(262144,[55627,64...|
|'What Canadians H...|     4.0|['what, canadians...|['what, canadians...|(262144,[16415,10...|(262144,[16415,10...|
|'allo Davina welc...|     4.0|['allo, davina, w...|['allo, davina, w...|(262144,[1512,124...|(262144,[1512,124...|
|'course it's not ...|     4.0|['course, it's, n...|['course, quantit...|(262144,[43890,70...|(262144,[43890,70...|
|'s bus just broke...|     0.0|['s, bus, just, b...|    ['s, bus, broke]|(262144,[91694,92...|(262144,[91694,92...|
|'s happy thought ...|     0.0|['s, happy, thoug...|['s, happy, thoug...|(262144,[12409,29...|(262144,[12409,29...|
|'s screen size is...|     0.0|['s, screen, size...|['s, screen, size...|(262144,[92492,14...|(262144,[92492,14...|
|'s voice makes me...|     4.0|['s, voice, makes...|['s, voice, makes...|(262144,[92492,10...|(262144,[92492,10...|
|'tis ok I posted ...|     4.0|['tis, ok, i, pos...|['tis, ok, posted...|(262144,[21894,64...|(262144,[21894,64...|
|A BIG Welcome to ...|     4.0|[a, big, welcome,...|[big, welcome, ne...|(262144,[64358,10...|(262144,[64358,10...|
|      A BOMB INATION|     4.0|  [a, bomb, ination]|     [bomb, ination]|(262144,[26648,66...|(262144,[26648,66...|
|A Demi I hope no ...|     0.0|[a, demi, i, hope...|[demi, hope, one,...|(262144,[21823,61...|(262144,[21823,61...|
|A Don't worry abo...|     4.0|[a, don't, worry,...|[worry, hun, wors...|(262144,[117975,1...|(262144,[117975,1...|
|A GIRL IS SO LUCK...|     0.0|[a, girl, is, so,...|[girl, lucky, won...|(262144,[13781,33...|(262144,[13781,33...|
|A Gabrielle wishe...|     0.0|[a, gabrielle, wi...|[gabrielle, wishe...|(262144,[991,2078...|(262144,[991,2078...|
|A How babies just...|     4.0|[a, how, babies, ...|[babies, light, g...|(262144,[109208,2...|(262144,[109208,2...|
|A I am sorry abou...|     0.0|[a, i, am, sorry,...|             [sorry]|(262144,[144961],...|(262144,[144961],...|
|A I hella miss yo...|     0.0|[a, i, hella, mis...|       [hella, miss]|(262144,[197515,2...|(262144,[197515,2...|
|A I hope they cal...|     0.0|[a, i, hope, they...|[hope, calm, fun,...|(262144,[23087,46...|(262144,[23087,46...|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Spark MLlib NLP

Building model

```
lr.fit(df4).transform(df4).show()
```

```
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|                text|polarity|              words1|              words2|      term_frequency|            features|       rawPrediction|         probability|prediction|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|['ayye, pass, mon...|(262144,[31536,64...|(262144,[31536,64...|[12.9780656017038...|[0.21901779229395...|       4.0|
|'Honey' the chick...|     0.0|['honey', the, ch...|['honey', chicken...|(262144,[55627,64...|(262144,[55627,64...|[12.0945489104097...|[0.94421689167222...|       0.0|
|'What Canadians H...|     4.0|['what, canadians...|['what, canadians...|(262144,[16415,10...|(262144,[16415,10...|[5.99369343435033...|[0.18325323124123...|       4.0|
|'allo Davina welc...|     4.0|['allo, davina, w...|['allo, davina, w...|(262144,[1512,124...|(262144,[1512,124...|[4.37111819424014...|[0.00369593761715...|       4.0|
|'course it's not ...|     4.0|['course, it's, n...|['course, quantit...|(262144,[43890,70...|(262144,[43890,70...|[7.52256788373932...|[0.01878302997629...|       4.0|
|'s bus just broke...|     0.0|['s, bus, just, b...|    ['s, bus, broke]|(262144,[91694,92...|(262144,[91694,92...|[10.5011690396112...|[0.99354789740347...|       0.0|
|'s happy thought ...|     0.0|['s, happy, thoug...|['s, happy, thoug...|(262144,[12409,29...|(262144,[12409,29...|[11.7400887292877...|[0.55704370276780...|       0.0|
|'s screen size is...|     0.0|['s, screen, size...|['s, screen, size...|(262144,[92492,14...|(262144,[92492,14...|[10.0581323946005...|[0.60877308779168...|       0.0|
|'s voice makes me...|     4.0|['s, voice, makes...|['s, voice, makes...|(262144,[92492,10...|(262144,[92492,10...|[8.03725845045519...|[0.17077618299793...|       4.0|
|'tis ok I posted ...|     4.0|['tis, ok, i, pos...|['tis, ok, posted...|(262144,[21894,64...|(262144,[21894,64...|[12.1513795255169...|[0.89768486529536...|       0.0|
|A BIG Welcome to ...|     4.0|[a, big, welcome,...|[big, welcome, ne...|(262144,[64358,10...|(262144,[64358,10...|[4.74536197045959...|[7.71752216152728...|       4.0|
|      A BOMB INATION|     4.0|  [a, bomb, ination]|     [bomb, ination]|(262144,[26648,66...|(262144,[26648,66...|[6.63379263319305...|[0.40468526673953...|       4.0|
|A Demi I hope no ...|     0.0|[a, demi, i, hope...|[demi, hope, one,...|(262144,[21823,61...|(262144,[21823,61...|[11.7315306643212...|[0.95018414511760...|       0.0|
|A Don't worry abo...|     4.0|[a, don't, worry,...|[worry, hun, wors...|(262144,[117975,1...|(262144,[117975,1...|[9.12771068870963...|[0.12421053641147...|       4.0|
|A GIRL IS SO LUCK...|     0.0|[a, girl, is, so,...|[girl, lucky, won...|(262144,[13781,33...|(262144,[13781,33...|[12.4060325488055...|[0.42720958546993...|       4.0|
|A Gabrielle wishe...|     0.0|[a, gabrielle, wi...|[gabrielle, wishe...|(262144,[991,2078...|(262144,[991,2078...|[12.8566097993347...|[0.86975949662995...|       0.0|
|A How babies just...|     4.0|[a, how, babies, ...|[babies, light, g...|(262144,[109208,2...|(262144,[109208,2...|[5.62931364182358...|[0.00755680436664...|       4.0|
|A I am sorry abou...|     0.0|[a, i, am, sorry,...|            [sorry]|(262144,[144961],...|(262144,[144961],...|[7.97099816820809...|[0.91272636828932...|       0.0|
|A I hella miss yo...|     0.0|[a, i, hella, mis...|      [hella, miss]|(262144,[197515,2...|(262144,[197515,2...|[8.58203790611087...|[0.88237741144112...|       0.0|
|A I hope they cal...|     0.0|[a, i, hope, they...|[hope, calm, fun,...|(262144,[23087,46...|(262144,[23087,46...|[7.65951633148478...|[0.08779867011967...|       4.0|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

Spark MLlib NLP

Building model

```
semantic_analysis_pipeline = Pipeline(
    stages=[tokenizer, stopwords_remover, hashing_tf, idf, lr]
)

semantic_analysis_model = semantic_analysis_pipeline.fit(training_data)
```

```
spark = (
    SparkSession.builder.appName("ModelTraining")
    .config("spark.executor.memory", "4g")
    .getOrCreate()
)
```

Spark MLlib NLP

Building model

```
%%time
trained_df = semantic_analysis_model.transform(training_data)
val_df = semantic_analysis_model.transform(validation_data)
test_df = semantic_analysis_model.transform(test_data)

trained_df.show()
val_df.show()
test_df.show()
```

```
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|                text|polarity|              words1|              words2|      term_frequency|            features|       rawPrediction|         probability|prediction|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|                   '|     0.0|                 [']|                 [']|(262144,[186171],...|(262144,[186171],...|[8.07815652290997...|[0.46734275935458...|       4.0|
|                   '|     4.0|                 [']|                 [']|(262144,[186171],...|(262144,[186171],...|[8.07815652290997...|[0.46734275935458...|       4.0|
|         ' ' ' ' ' '|     4.0|  [', ', ', ', ', ']|  [', ', ', ', ', ']|(262144,[186171],...|(262144,[186171],...|[8.45144104491861...|[0.64774113228464...|       0.0|
|           ' ' PIMS|     4.0|         [', ', pims]|         [', ', pims]|(262144,[186171,2...|(262144,[186171,2...|[8.25408146085305...|[0.55428514973836...|       0.0|
|  ' ' cute hug facey|     0.0|[', ', cute, hug,...|[', ', cute, hug,...|(262144,[23837,65...|(262144,[23837,65...|[7.97939735133558...|[0.41707283827387...|       4.0|
| ' i didn't mean...|     0.0|[', ', i, didn't,...|[', ', mean, anyt...|(262144,[991,3710...|(262144,[991,3710...|[8.77058491428643...|[0.77414425733478...|       0.0|
| ' it's a monkey...|     4.0|[', ', it's, a, m...|[', ', monkey, re...|(262144,[31536,60...|(262144,[31536,60...|[9.03477535710885...|[0.85428003703438...|       0.0|
| ' it's good to ...|     4.0|[', ', it's, good...|[', ', good, clea...|(262144,[3955,113...|(262144,[3955,113...|[7.75654967633660...|[0.31408966375802...|       4.0|
| '' is actually ...|     4.0|[', '', is, actua...|[', '', actually,...|(262144,[109068,1...|(262144,[109068,1...|[7.88455743582298...|[0.37211015997989...|       4.0|
| AMAZING heey ho...|     4.0|[', amazing, heey...|[', amazing, heey...|(262144,[23087,51...|(262144,[23087,51...|[6.27953674659944...|[0.02321840068064...|       4.0|
| And she broke m...|     0.0|[', and, she, bro...|   [', broke, heart]|(262144,[112352,1...|(262144,[112352,1...|[9.17251117294008...|[0.88644832647143...|       0.0|
| Birthday Sex' i...|     4.0|[', birthday, sex...|[', birthday, sex...|(262144,[86553,12...|(262144,[86553,12...|[7.59186795756178...|[0.24673277052895...|       4.0|
| Bored on aim wa...|     4.0|[', bored, on, ai...|[', bored, aim, w...|(262144,[9958,180...|(262144,[9958,180...|[7.16161288850605...|[0.12072903391416...|       4.0|
| But u were doin...|     0.0|[', but, u, were,...|[', u, something,...|(262144,[51783,57...|(262144,[51783,57...|[7.38601865334590...|[0.17947636705783...|       4.0|
| CRAZY IN LOVE I...|     0.0|[', crazy, in, lo...|[', crazy, love, ...|(262144,[113024,1...|(262144,[113024,1...|[7.09678520033637...|[0.10894595917827...|       4.0|
|                ' D|     0.0|            [', d]|             [', d]|(262144,[89530,18...|(262144,[89530,18...|[8.05038771619375...|[0.45314700890916...|       4.0|
| DVD out in US t...|     0.0|[', dvd, out, in,...|[', dvd, us, mont...|(262144,[39216,62...|(262144,[39216,62...|[8.89832313333979...|[0.81767537991268...|       0.0|
| Dancing with th...|     4.0|[', dancing, with...|[', dancing, devi...|(262144,[12409,17...|(262144,[12409,17...|[7.72910547164212...|[0.30057766094310...|       4.0|
| GERD is acting ...|     0.0|[', gerd, is, act...|[', gerd, acting,...|(262144,[62224,95...|(262144,[62224,95...|[9.02607468309538...|[0.85322284832704...|       0.0|
| Go to settings ...|     4.0|[', go, to, setti...|[', go, settings,...|(262144,[64465,87...|(262144,[64465,87...|[7.44173772490499...|[0.19552947334757...|       4.0|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

```
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|                text|polarity|              words1|              words2|      term_frequency|            features|       rawPrediction|         probability|prediction|
+--------------------+--------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|'AYYE PASS ME THE...|     4.0|['ayye, pass, me,...|['ayye, pass, mon...|(262144,[31536,64...|(262144,[31536,64...|[7.54757712507432...|[0.22977054253765...|       4.0|
|'Honey' the chick...|     0.0|['honey', the, ch...|['honey', chicken...|(262144,[55627,64...|(262144,[55627,64...|[11.6334177126232...|[0.99905297148233...|       0.0|
|'What Canadians H...|     4.0|['what, canadians...|['what, canadians...|(262144,[16415,10...|(262144,[16415,10...|[6.92111820388093...|[0.07941775606509...|       4.0|
|'allo Davina welc...|     4.0|['allo, davina, w...|['allo, davina, w...|(262144,[1512,124...|(262144,[1512,124...|[7.32487646757031...|[0.16216385641909...|       4.0|
|'course it's not ...|     4.0|['course, it's, n...|['course, quantit...|(262144,[43890,70...|(262144,[43890,70...|[5.44397183072022...|[0.00445302843612...|       4.0|
|'s bus just broke...|     0.0|['s, bus, just, b...|    ['s, bus, broke]|(262144,[91694,92...|(262144,[91694,92...|[9.11167400468925...|[0.87355933245613...|       0.0|
|'s happy thought ...|     0.0|['s, happy, thoug...|['s, happy, thoug...|(262144,[12409,29...|(262144,[12409,29...|[8.31311281522257...|[0.58091120318997...|       0.0|
|'s screen size is...|     0.0|['s, screen, size...|['s, screen, size...|(262144,[92492,14...|(262144,[92492,14...|[8.35122411375337...|[0.60059380009412...|       0.0|
|'s voice makes me...|     4.0|['s, voice, makes...|['s, voice, makes...|(262144,[92492,10...|(262144,[92492,10...|[7.39931990449596...|[0.18311418974805...|       4.0|
```

Spark MLlib NLP

Building model

```
%%time
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

evaluator = MulticlassClassificationEvaluator(labelCol="polarity", metricName="accuracy")
accuracy_val = evaluator.evaluate(val_df)
accuracy_test = evaluator.evaluate(test_df)
print("Validation Data:")
print(f"Accuracy: {accuracy_val*100:.5f}%")
print("Testing Data:")
print(f"Accuracy: {accuracy_test*100:.5f}%")
```

```
Validation Data:
Accuracy: 77.33275%
Testing Data:
Accuracy: 76.87971%
CPU times: user 14.1 ms, sys: 1.91 ms, total: 16 ms
Wall time: 19.7 s
```

Spark MLlib NLP

Building model

```
final_model = semantic_analysis_pipeline.fit(data)
final_model.save(MODEL_PATH)
```

Q & A

Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.