FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

RESEARCH GROUP **DATA ANALYTICS / BIG DATA**

BIG DATA ANALYTICS USING (PY)SPARK FOR ANALYZING IPO TWEETS

Dirk Van den Poel (Special thanks: Giselle Van Dongen & Sacha Dubrulle
Jolien Jackers, Sami Schyvinck & Diederick Van Damme)

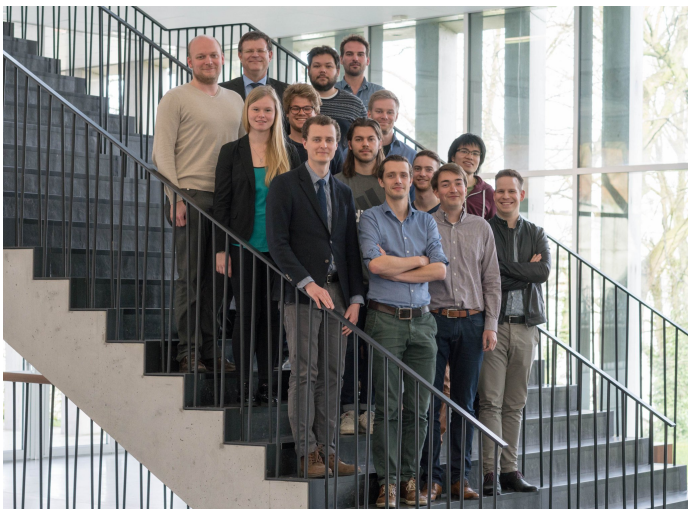
GHENT UNIVERSITY

APACHE: BIG_DATA NORTH_AMERICA

Apache Big Data North America in Miami, FL #ApacheBigData

UGENT DATA ANALYTICS TEAM

- Data Analytics
  www.dataanalytics.UGent.be
- Big Data
  www.bigdata.UGent.be



GHENT UNIVERSITY

DATA ANALYTICS @ UGENT

Descriptive | Predictive | Prescriptive

Master of Science in Marketing Analytics

1999 www.mma.ugent.be

Master of Science in Business Engineering / Data Analytics

2005 www.dataaanalytics.ugent.be

GHENT UNIVERSITY

3

RESEARCH QUESTION

GHENT UNIVERSITY

4

RESEARCH QUESTIONS

- Are we able to predict, based on tweet characteristics, whether a message will be **retweeted** or not?
- Is there a relation between the buzz on social media during an IPO (**I**nitial **P**ublic **O**ffering) and the evolution of the IPO price?

GHENT UNIVERSITY

5

HARDWARE

- Commodity hardware: Tweet streaming & fetching retweets
- HP Proliant C7000 Blade cluster (16x BL460c G6, 32x Quad core CPUs, 512 GB DDR3): ETL big data operations, mainly filtering and joining original tweets with retweets
- HP Z600 workstation to handle tweet analysis
- Databricks: Student projects

GHENT UNIVERSITY

SOFTWARE STACK

GHENT UNIVERSITY

7

STARTING POINT: DATA ACQUISITION

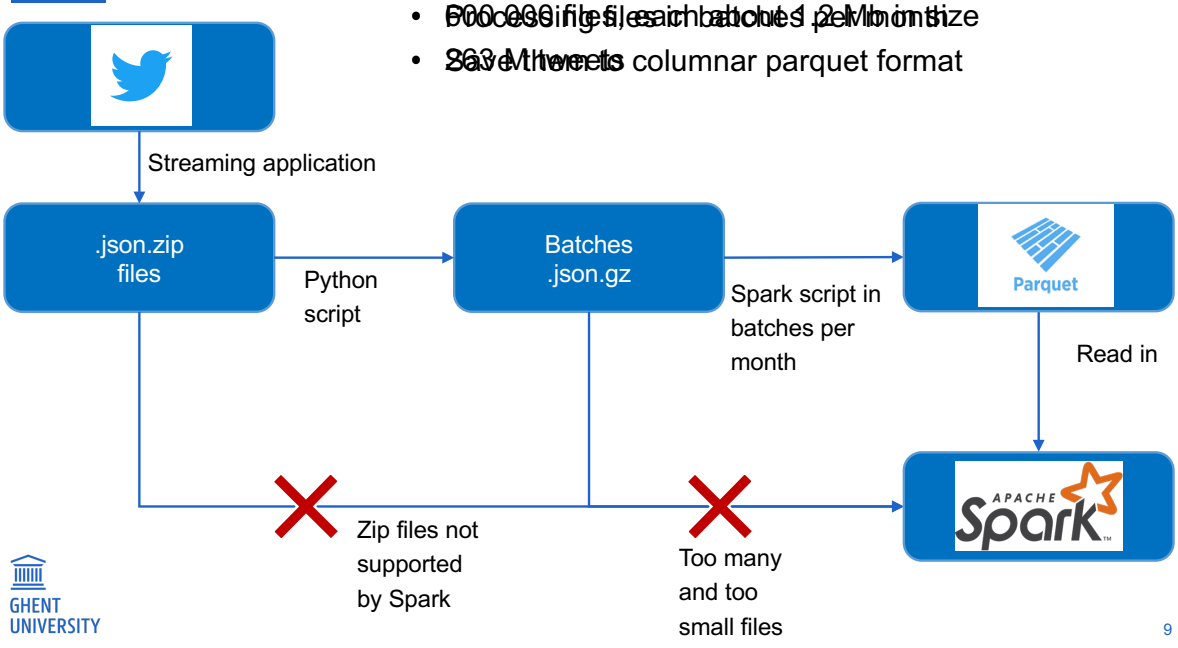
- Single-threaded java application on commodity hardware
- Saved as JSON files per 300 tweets, about 2 - 3 JSON files per minute ➔ about 600 000 in total
- List of keywords related to IPO's
  - Company names: humana, editas
  - Tickers: HUM, EDIT
  - General: #ipo, ipo

GHENT UNIVERSITY

8

ETL

- 800 000 files, each about 1.2 Mb in size
- 263 Mtweets



- Processing steps in batches per month
- Save tweets in columnar parquet format

GHENT UNIVERSITY

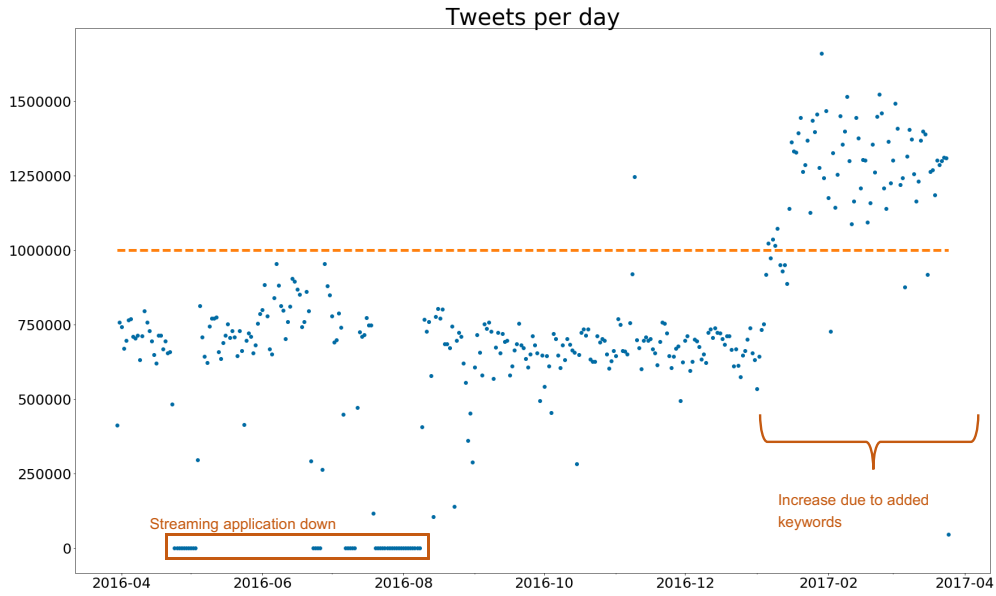
9

PARQUET

- Distributed columnar data storage format
- Excellent integration with Spark
- High compression ratio
- Fast read

GHENT UNIVERSITY

TIME DISTRIBUTION



GHENT UNIVERSITY

11

FILTERING

- 263 M tweets (03/30/2016 – 03/30/2017)
- Problems:
  - Keywords are common words
  - Keywords are part of common words

GHENT UNIVERSITY

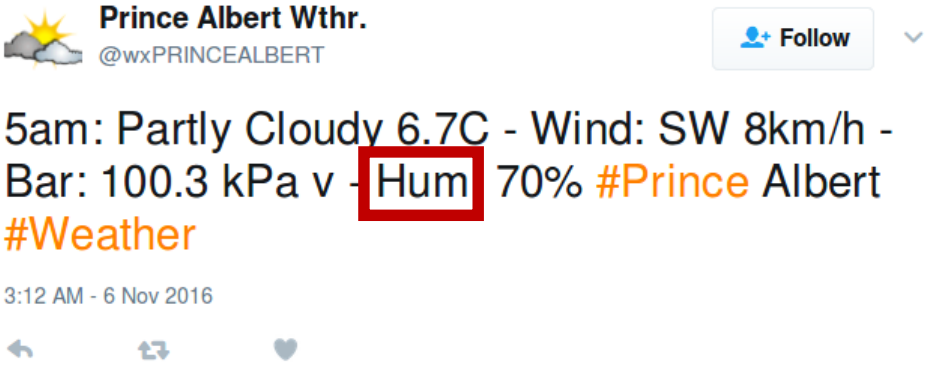
12

EXAMPLES

GHENT UNIVERSITY

13

EXAMPLES



GHENT UNIVERSITY

14

FILTERING

- This leaves us with 263 M tweets (03/30/2016 – 03/30/2017)

  Problems:
  - Tickers/company names are (part of) common words
  - ticker ➔ cashtags: $SNAP, $EDIT
- Regex filtering in Spark

  ➔ 1.7 M tweets remaining
  - Contains ipo
  - Contains cashtag
  - Contains name

GHENT UNIVERSITY

15

FILTERING

```
# returns 0 if column is empty (no regexp match found), 1 otherwise (at least one regexp match found)
def non_empty_column(ipo_cashtag_column):
    if len(ipo_cashtag_column) > 0:
        return 1
    else:
        return 0

non_empty_column_udf = udf(non_empty_column)

# REGEX to extract tickers or presence of IPO
cashtag_expression = '\$[a-z]{1,6}'
ipo_expression = ' #ipo | ipo '

augmented_df = (df.select("tweet_id_str", "text", lower(df.text).alias("text_lower"))
                  .select("*",
                          non_empty_column_udf(regexp_extract('text_lower', ipo_expression, 0)).alias("contains_ipo"),
                          non_empty_column_udf(regexp_extract('text_lower', cashtag_expression, 0)).alias("contains_cashtag"),
                          udf_contains_company_name("text_lower").alias("contains_company_name")
                         )
               )

filtered_df = augmented_df.filter((augmented_df.contains_ipo == 1) |
                                  (augmented_df.contains_cashtag == 1) |
                                  (augmented_df.contains_company_name == 1)
                                 )
```

GHENT UNIVERSITY

16

FILTERING, IN RETROSPECT

- Loose constraints
  e.g. EDIT, ABS, HUM, SNAP, IPO
  + Catch all buzz, filtering afterwards still possible
  – High storage volume

- Tight constraints
  e.g. $EDIT, $ABS, $HUM, $SNAP, #IPO
  + Limited storage volume
  - Dataset limited to financial tweets

GHENT UNIVERSITY

17

FILTERING, RESULT

```
ML_BT.sample(withReplacement=False, fraction = 0.1).select("text").show(10, truncate = False)
```

```
+--------------------------------------------------------------------------------------------------------------------------------------------------------+
|text                                                                                                                                                    |
+--------------------------------------------------------------------------------------------------------------------------------------------------------+
|"A new article on vaccination by Mapi and Sanofi Pasteur: https://t.co/knI3pvuRje" https://t.co/gooPspgk6t                                              |
|#3: VIZIO SS2521-C6 25-Inch 2.1 Channel Sound Stand (2015 Model): VIZIO SS2521-C6 25-Inch 2.1 Channel Sound S... https://t.co/e5FsMJ78jD|
|#5: VIZIO  D43-D1 D-Series 43" Class Full Array LED Smart TV (Black) https://t.co/ol5BYruofz                                                            |
|#AudentesTherapeutics Why the Key to Gene Therapy Could be in an Old Bay Area Warehouse  . Read more: https://t.co/F7v61NSVIO $BOLD                    |
|#Bayshore Petroleum Corp. Announces Changes to Board of Directors and Officers. Read more: https://t.co/TOqbEEgbZD $BSH                                 |
|#Bestseller Vizio E32-C1 32-inch 1080p 60Hz Full-Array Smart LED HDTV with… https://t.co/wq3ECncdV7 #Deal #F4F https://t.co/3841xW7dwn                  |
|#Business | Snap arrives in London to woo skeptical investors ahead of IPO #RTGFinance                                                                  |
|#Cisco AppDynamics' pre-IPO sale may not be the last this year. Read more: https://t.co/gdjDWGD0nn $CSCO                                               |
|#Deals #Ebay Vizio M55-D0 - 55-Inch 4K Ultra HD HDR TV Home Theater Display https://t.co/7JqhWb7HGG                                                    |
|#Electronics  Vizio E48-D0 1080p 48" LED Smart TV, Black https://t.co/i859DGyflB  #Gifts https://t.co/pAgYwXfiQO                                       |
+--------------------------------------------------------------------------------------------------------------------------------------------------------+
only showing top 10 rows
```

GHENT UNIVERSITY

18

RETWEETS

- At the time of streaming, the number of retweets is not yet available
  => Fetch the tweet again after waiting period for tweets to accumulate retweets
- 8 CentOS machines with multiple Authentication Keys each to overcome rate limits using simple Python scripts
- BUT:
  - Account deleted/blocked?
  - Account not publicly available anymore
  - Tweet deleted

GHENT UNIVERSITY

19

RETWEETS, IN RETROSPECT

- Depending on the goal of the analysis, keeping all tweet information at initial streaming might be redundant

GHENT UNIVERSITY

20

MODELING RETWEETS

GHENT UNIVERSITY

21

FEATURES OVERVIEW

TEXT FEATURES

- Contains cashtag
- Cashtag count
- Contains hashtag
- Hashtag count
- Language
- Contains URL
- Contains cashtag with spaces

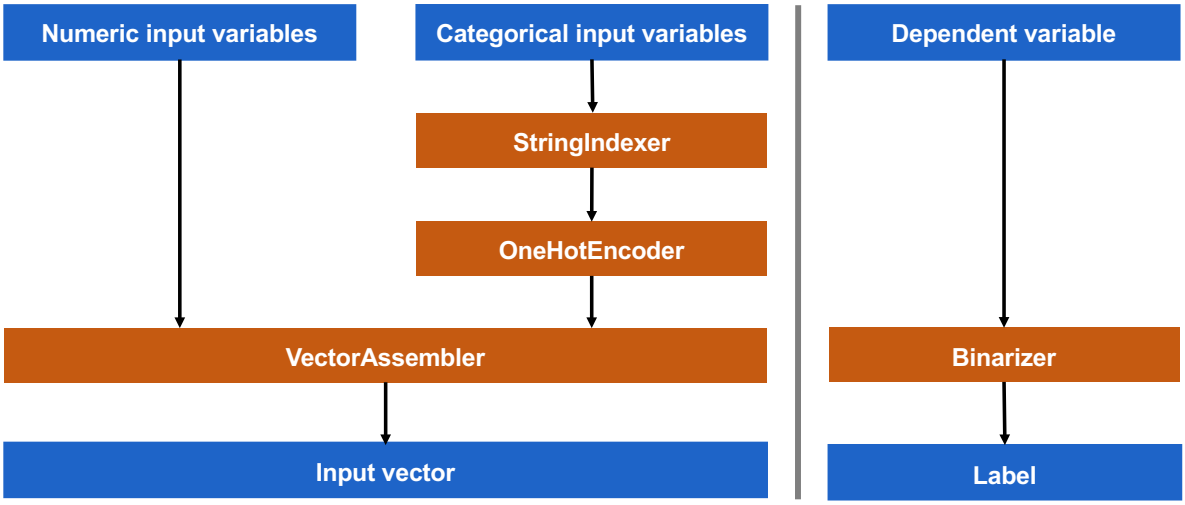
NON-TEXT FEATURES

- User seniority
- User favorite count
- User friends count
- Users follower count
- Day of week

GHENT UNIVERSITY

22

FEATURE PREPROCESSING



GHENT UNIVERSITY

23

DATA MINING METHODOLOGY

- Logistic regression
- Random forest

GHENT UNIVERSITY

24

MODEL DESCRIPTIONS

```
def RF_classification(model_description, df, cat_columns, num_columns, output_column = "retweet_binary"):
    stringIndexers = []; indexed_cols = []; encoders = []; encoded_cols = []

    for column in cat_columns:
        inCol= column
        outCol = column + "indexed"
        stringIndexer = StringIndexer(inputCol = inCol, outputCol=outCol)
        stringIndexers.append(stringIndexer)
        indexed_cols.append(outCol)

    for column in indexed_cols:
        inCol= column
        outCol = column + "encoded"
        encoder = OneHotEncoder(inputCol=inCol, outputCol=outCol)
        encoders.append(encoder)
        encoded_cols.append(outCol)

    binarizer = Binarizer(inputCol = "retweet_count_double", outputCol = output_column, threshold=0.9)
    assembler = VectorAssembler(inputCols= num_columns + encoded_cols, outputCol="features")

    pipeline = Pipeline(stages= stringIndexers + encoders + [assembler] + [binarizer])
    pipelineModel = pipeline.fit(df)

    transformed_df = (pipelineModel.transform(df)
                                   .select("features", col("retweet_binary").cast("integer"), "retweet_count")
                      )

    train, validation, test = transformed_df.randomSplit([0.8, 0.1, 0.1], seed = 42)

    rf = RandomForestClassifier(numTrees=500, labelCol="retweet_binary")
    rf_model = rf.fit(train)

    train_predictions = rf_model.transform(train)
    validation_predictions = rf_model.transform(validation)

    AUC_evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction", labelCol="retweet_binary", metricName="areaUnderROC")

    AUC_train = AUC_evaluator.evaluate(train_predictions)
    AUC_validation = AUC_evaluator.evaluate(validation_predictions)
    print "##################" + model_description.upper() + "##################"
    print "TRAIN AUC:       {}".format(AUC_train)
    print "VALIDATION AUC:        {}".format(AUC_validation)
```

StringIndexer and OneHotEncoder

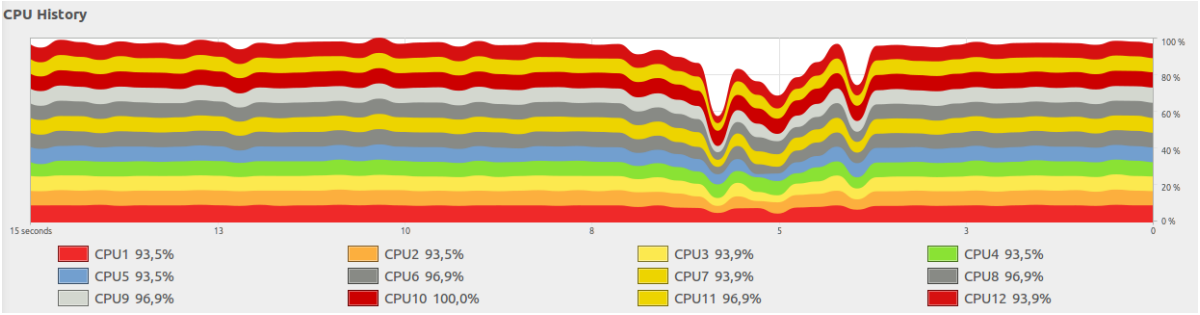
Binarizer & VectorAssembler

Pipeline

Data split, train, predict and evaluate

25

MODEL DESCRIPTIONS



GHENT UNIVERSITY

26

MODEL PERFORMANCE

Algorithm	Features	Train AUC	Test AUC
Logistic Regression	Text	0.6340	0.6290
Random Forest	Text	0.6469	0.6416
Logistic Regression	Non-text	0.6688	0.6661
Random Forest	Non-text	**0.7948**	**0.7923**
Logistic Regression	Text + Non-text	0.7090	0.7049
Random forest	Text + Non-text	**0.7994**	**0.7970**

GHENT UNIVERSITY

27

LINK: TWITTER & STOCK PRICE

GHENT UNIVERSITY

28

LITERATURE REVIEW: TWITTER -> RETURNS

	Lin. R	Log. R	NN	SVM	SVR**	SOFNN	NB	Accuracy
Bollen, J., Mao, H., & Zeng, X. (2011)***						X		86.70%
Au, Benjamin; Zhang, Qian; Zhang, Wanlu (2013)		X		X				62.32%
Paglia, S. a M. (2013)			X*	X			X**	61%
Sumbureru, P.T. (2015)				X				54%-59%

GHENT UNIVERSITY

29

LITERATURE REVIEW: IPO TWEETS -> RETURNS

– Julie Zhang (Nov. 2015, Univ. of Massachusetts): 'Social Media Whispers During The IPO Quiet Period', INFORMS-conference in Philadelphia.

GHENT UNIVERSITY

30

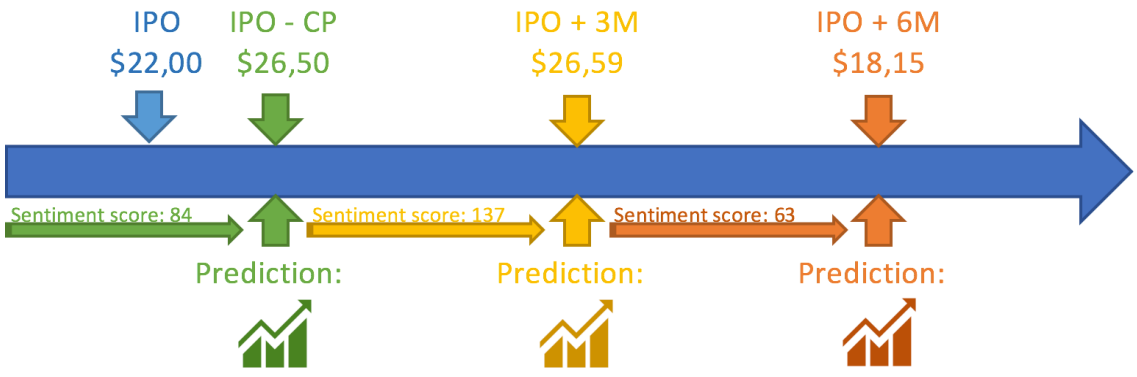
STOCKS

GHENT UNIVERSITY

31

ARA : AMERICAN RENAL ASSOCIATES HOLDINGS

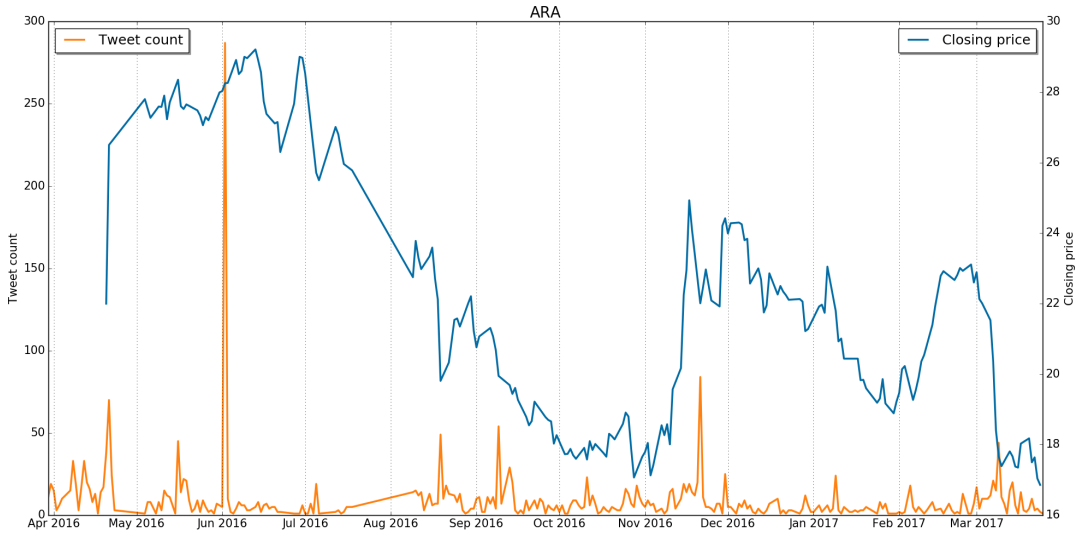
IPO date: April 21, 2016



GHENT UNIVERSITY

32

ARA: AMERICAN RENAL ASSOCIATES HOLDINGS



GHENT UNIVERSITY

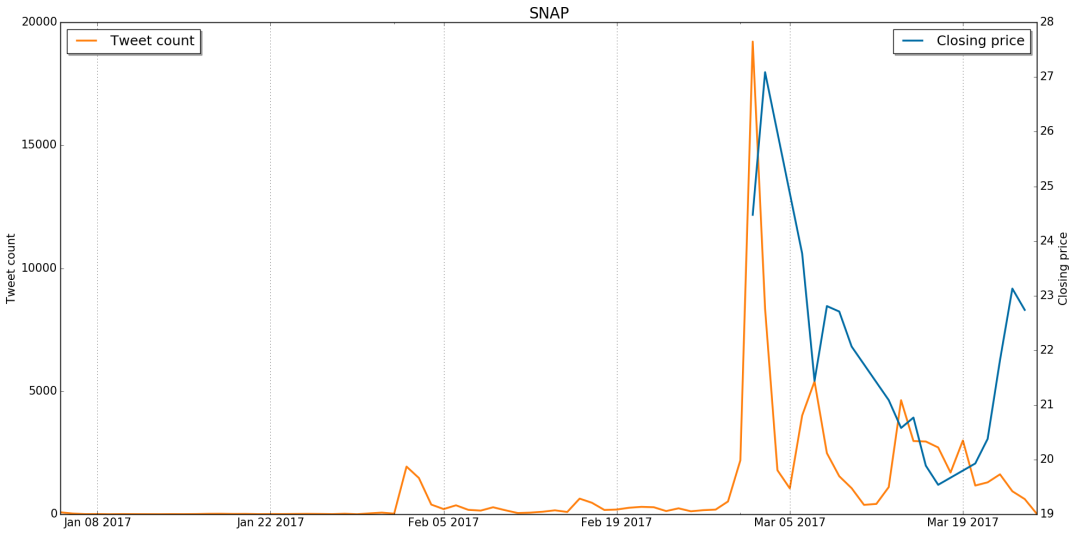
33

ARA

GHENT UNIVERSITY

34

SNAP



GHENT UNIVERSITY

35

ANSWERS TO RESEARCH QUESTION

GHENT UNIVERSITY

36

RESEARCH QUESTION: IPO VS FIRST-DAY CP

	IPO	IPO - CP	UP_1/DOWN_0	Sentiment score	POS_1/DOWN_0
ARA	$22.00	$26.50	1	84	1
BOLD	$15.00	$15.13	1	45	1
MGP	$21.00	$22.01	1	71	1
NTLA	$18.00	$22.10	1	24	1
PLSE	$4.00	$4.17	1	0	0
RETA	$11.00	$13.07	1	-6	0
RRR	$19.50	$18.70	0	80	1
SBPH	$12.00	$11.10	0	0	0
SCWX	$14.00	$14.00	0	56	1
AZRE	$18.00	$14.60	0	12	1
SUPV	$11.00	$11.50	1	51	1
TUSK	$15.00	$13.26	0	8	1
TRHC	$12.00	$14.88	1	1	1
TCMD	$15.00	$11.08	0	69	1
SNAP	$17.00	$24.48	1	6087	1

	Up	Down	%	Up	Down
Pos	7	5	Pos	47%	33%
Neg	2	1	Neg	13%	7%

15

GHENT UNIVERSITY

37

RESEARCH QUESTION: FIRST-DAY CP VS 3M LATER

	PRICE 3M AFTER IPO				
	IPO - CP	IPO - 3M later	UP_1/DOWN_0	Sentiment score	POS_1/DOWN_0
ARA	$26.50	$26.59	1	137	1
BOLD	$15.13	$17.42	1	103	1
MGP	$22.01	$26.70	1	40	1
NTLA	$22.10	$20.19	0	111	1
PLSE	$4.17	$4.62	1	-9	0
RETA	$13.07	$19.21	1	294	1
RRR	$18.70	$23.14	1	167	1
SBPH	$11.10	$10.77	0	11	1
SCWX	$14.00	$14.51	1	238	1
AZRE	$14.60	$16.10	1	254	1
SUPV	$11.50	$13.50	1	89	1
TUSK	$13.26	$17.46	1	269	1
TRHC	$14.88	$14.63	0	452	1
TCMD	$11.08	$17.52	1	258	1

	Up	Down	%	Up	Down
Pos	10	3	Pos	71%	21%
Neg	1	0	Neg	7%	0%

14

GHENT UNIVERSITY

38

RESEARCH QUESTION: 3M VS 6M LATER

	PRICE 6M AFTER IPO				
	IPO - 3M later	IPO - 6M later	UP_1/DOWN_0	Sentiment score	POS_1/DOWN_0
ARA	26.59	18.15	0	63	1
MGP	26.70	25.90	0	1	1
NTLA	20.19	14.67	0	63	1
PLSE	4.62	5.97	1	66	1
RETA	19.21	27.83	1	123	1
RRR	23.14	22.12	0	365	1
SBPH	10.77	7.85	0	206	1
SCWX	14.51	11.29	0	6	1

	Up	Down	%	Up	Down
Pos	2	6	Pos	25%	75%
Neg	0	0	Neg	0%	0%

8

GHENT UNIVERSITY

39

FUTURE RESEARCH & RECOMMENDATIONS

GHENT UNIVERSITY

40

FUTURE RESEARCH

- Take into account who is tweeting (cf. pump and dump)
- Deep learning approach to focus on text of the tweet
- Recommendations for tweet success
- Can we extend stock-trading models using twitter information, either in the IPO of in a more general stock-trading perspective?

GHENT UNIVERSITY

41

IMPROVEMENTS OVER LAST YEAR

- Random Forest received major upgrade
- Findspark Python package

```
import findspark
findspark.init(spark_home = HOME_DIR + "/spark-2.1.0")
import pyspark

from pyspark.sql import SQLContext, SparkSession

try:
    sc = pyspark.SparkContext(appName = 'App')
    spark = SparkSession(sparkContext=sc)
    print "SparkSession initialized"
except ValueError:
    print "SparkSession already initialized"
```

GHENT UNIVERSITY

42

RECOMMENDATIONS

- Extending number of state-of-the-art data mining methodologies in Spark: e.g. XGBoost, Hybrid Ensembles
- Facilitating the interpretation of output (e.g. variables importances)

GHENT UNIVERSITY

43

DATABRICKS

GHENT UNIVERSITY

44

BIG DATA COURSE @ GHENT UNIVERSITY

- Target audience:
  - Business Engineering students
  - Exchange students
  - Students taking course as elective
- Heterogeneous backgrounds
- SQL + Python + Spark

GHENT UNIVERSITY

45

DATABRICKS STUDENT GROUP ASSIGNMENT

- Previously, job submits on HPC cluster
- Stable
- Easy notebook import
- Availability without need of powerful machine
- Notebook environment with nice GUI + nice graphs

GHENT UNIVERSITY

46

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Dirk Van den Poel

Senior Full Professor of Data Analytics / Big Data

DEPARTMENT OF MARKETING

E Dirk.VandenPoel@UGent.Be

Dirk Van den Poel
@dirkvandenpoel
Dirk VAN den POEL

www.crm.ugent.be
www.bigdata.ugent.be
www.mma.ugent.be
www.dataanalytics.ugent.be

GHENT UNIVERSITY

informs