LINKÖPINGS UNIVERSITET

[732A47] Text Mining

Text Mining Project

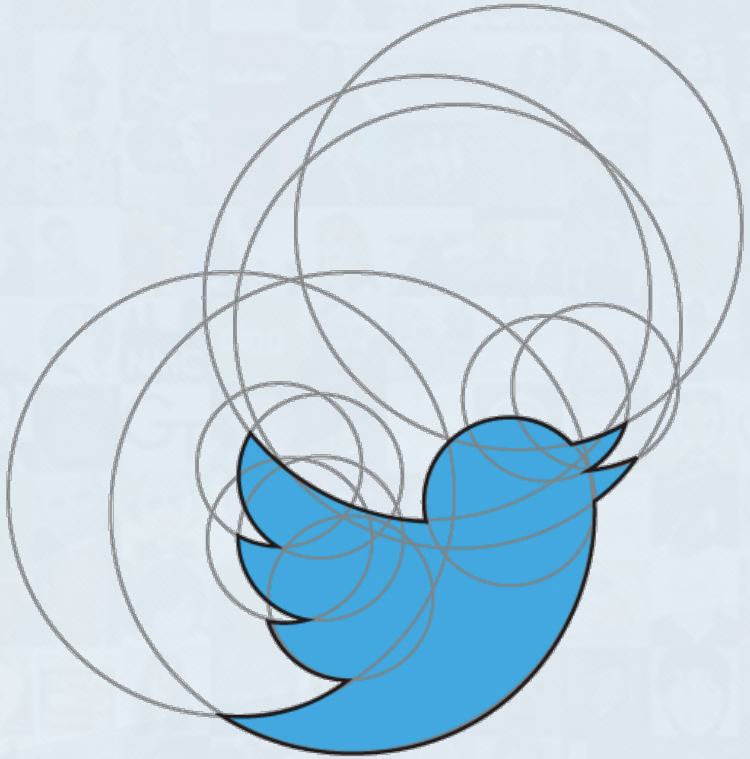
Retweets

student

Sebastiano Milardo

Outline

- *Problem Description and Data*
- *Analysis*
- *Conclusion*



Problem Description

- *What characterizes a tweet that gets retweeted by many people?*

Problem Description

- *What characterizes a tweet that gets retweeted by many people?*

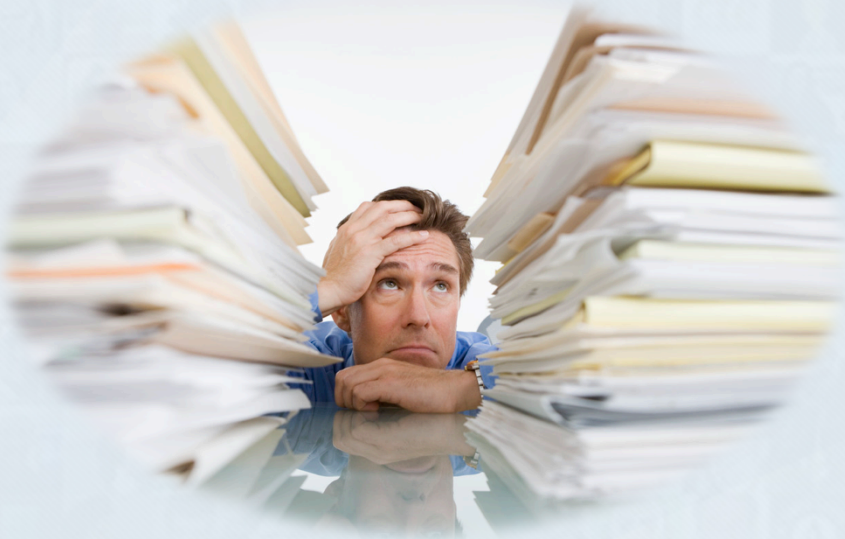
Successful Example

Souce: https://2012.twitter.com/en/golden-tweets.html

Successful Example

Problem Description

Too many…



…too few!

Data

- Google Analytics
  - Queries: 10-9 Cerrone, Abortion, Acne, Alex Morgan, AllyFollowSpree, Amanda Todd, Apple, Audi, Aurora Shooting, Back Pain, Bank, Barack Obama, Barcelona, Bared to you, Baseball, Basket, Big Bang Theory, Bloody Mary, BMW, Cancer, Cat, Chuck, CNN, Converse, Dad, Dance, Defendig Jacob, Depression, Derrick Rose, Diabetes, Dodge, Dog, Economic, Election, Enviroment, ESPN, Fast & Furious 6, Fifty Shades Darker, Fifty Shades Freed, Flowers, Food, Football, Ford, Fox News, Froch, Galaxy, Game of Thrones, Gangnam Style, Gas, Gay Marriage, Glee, Golf, Good Morning, Heart, Hello, Herpes, Honda, Hotel, Huffington Post, Hunger Games, Hurricane, Hurricane Isaac, Immigration, Instagram, iPhone, Jeremy Lin, Joe Paterno, Justin Bieber, Kate Middleton, Kevin Durant, KJ Noons, Kobe, KONY 2012, Lebron, Lolo Jones, Love, Magic Mike, Marc Gasol, Margarita, McDonald, MentionSomeoneYouLove, Michael Clarcke Duncan, Michael Phelps, Mitt Romney, Mojito, Mom, Money, Morgan Freeman, MSN, Music, NBC, Neil Armstrong, News, Nexus, Nicki Minaj, Nike, Nokia, NY, Obama, Obamacare, ObamacareKidsBooks, Olympics, One Direction, Parking, Paul Ryan, Peyton Manning, Photo, Pizza, Pizza Hut, President, Presidential Election, Prometheus, PSY, Restaurant, Ron Paul, Selena Gomez, Serena Williams, Snooki, Son, Song, SOPA, Sport, Starbucks, Super Bowl, Swimming , Ted, Tennis, ThatOneExWho, The Avengers, The Dark Knight Rises, The Serpent's Shadow, Toyota, Twilight, UEFA Euro 2012 , UFC160, Walmart, WeAllKnowThatOnePersonWho, Whitney Houston, WI, X Factor

Information retrieval

- Topsy.com
- Twitter Streamer API
- Twitter API
  - **GET search**
  - **GET users/lookup**
  - **GET statuses/user_timeline**

Information retrieval

- ~~Topsy.com~~
- ~~Twitter Streamer API~~
- Twitter API
  - **GET search**
  - **GET users/lookup**
  - **GET statuses/user_timeline**

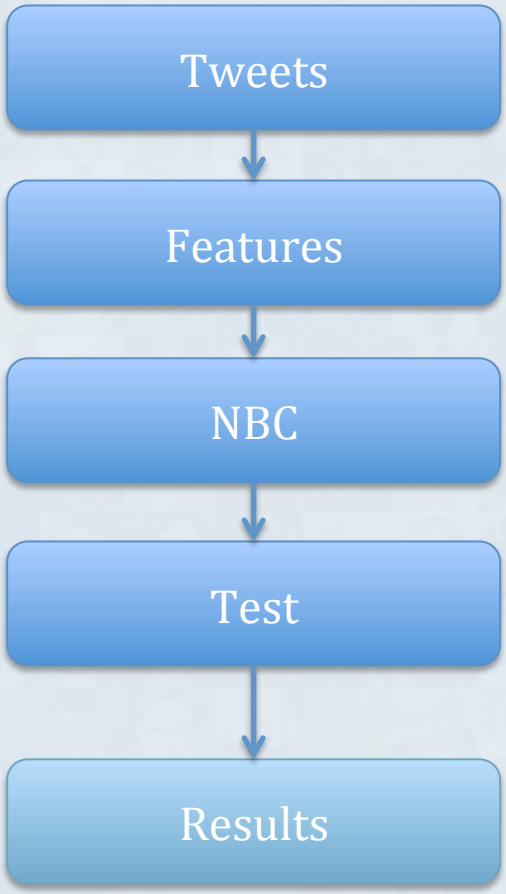
Information retrieval

- 92 010 Unique tweets
  - 9 606 Retweeted
  - 82 404 Non-Retweeted
- Hardware Limits…
  - 9 000 Retweeted
  - 9 000 Non-Retweeted

Features

- Text Mining:
  - Normalization (lowercase)
  - Removing Stop Words
- Features A:
  - Top 600 Most common words
  - Top 300 Most common Screen Names (@)
  - Top 400 Most common Hash Tag (#)
  - Average word length
  - Lexical diversity

Naive Bayes classifier



92 010 unique tweets

(9000+9000) tweets

Trained on 14400 tweets

Tested on 3600 tweets

Results

F-measure: 0.61 - Accuracy: 0.58 - Precision: 0.56 - Recall: 0.66

has(enviroment) = True	False : True =	16.3 : 1.0
has(prometheus) = True	False : True =	11.1 : 1.0
has(f1) = True	False : True =	9.1 : 1.0
has(retweet) = True	True : False =	8.7 : 1.0
has(sale) = True	False : True =	7.5 : 1.0
has(mojito) = True	False : True =	7.3 : 1.0
avg<3 = True	False : True =	7.1 : 1.0
has(syria) = True	False : True =	6.8 : 1.0
has(sopa) = True	False : True =	6.5 : 1.0
has(mi) = True	False : True =	6.4 : 1.0
has(honda) = True	False : True =	5.1 : 1.0
has(toyota) = True	False : True =	4.9 : 1.0
has(nexus) = True	False : True =	3.9 : 1.0
has(manchester) = True	True : False =	3.7 : 1.0
has(alcohol) = True	False : True =	3.7 : 1.0
has(ufc160) = True	False : True =	3.7 : 1.0
has(architecture) = True	False : True =	3.7 : 1.0
has(barcelona) = True	True : False =	3.7 : 1.0
has(iran) = True	True : False =	3.6 : 1.0
has(snooki) = True	False : True =	3.3 : 1.0

	F a l s e	T r u e
False	<917>	912
True	598	<1173>

Features

- Features related to a particular tweet

- Features B: A + Tweet Properties
  - N. of Links
  - N. of Hash Tags

Results

F-measure: 0.61 - Accuracy: 0.62 - Precision: 0.62 - Recall: 0.60

has(enviroment) = True	False : True =	16.3 : 1.0
has(prometheus) = True	False : True =	11.1 : 1.0
has(f1) = True	False : True =	9.1 : 1.0
has(retweet) = True	True : False =	8.7 : 1.0
has(sale) = True	False : True =	7.5 : 1.0
has(mojito) = True	False : True =	7.3 : 1.0
avg<3 = True	False : True =	7.1 : 1.0
has(syria) = True	False : True =	6.8 : 1.0
has(sopa) = True	False : True =	6.5 : 1.0
has(mi) = True	False : True =	6.4 : 1.0
has(honda) = True	False : True =	5.1 : 1.0
has(toyota) = True	False : True =	4.9 : 1.0
has(nexus) = True	False : True =	3.9 : 1.0
has(manchester) = True	True : False =	3.7 : 1.0
has(architecture) = True	False : True =	3.7 : 1.0
has(alcohol) = True	False : True =	3.7 : 1.0
has(ufc160) = True	False : True =	3.7 : 1.0
has(barcelona) = True	True : False =	3.7 : 1.0
has(iran) = True	True : False =	3.6 : 1.0
has(snooki) = True	False : True =	3.3 : 1.0

	F a l s e	T r u e
False	<1180>	650
True	706	<1064>

Followers

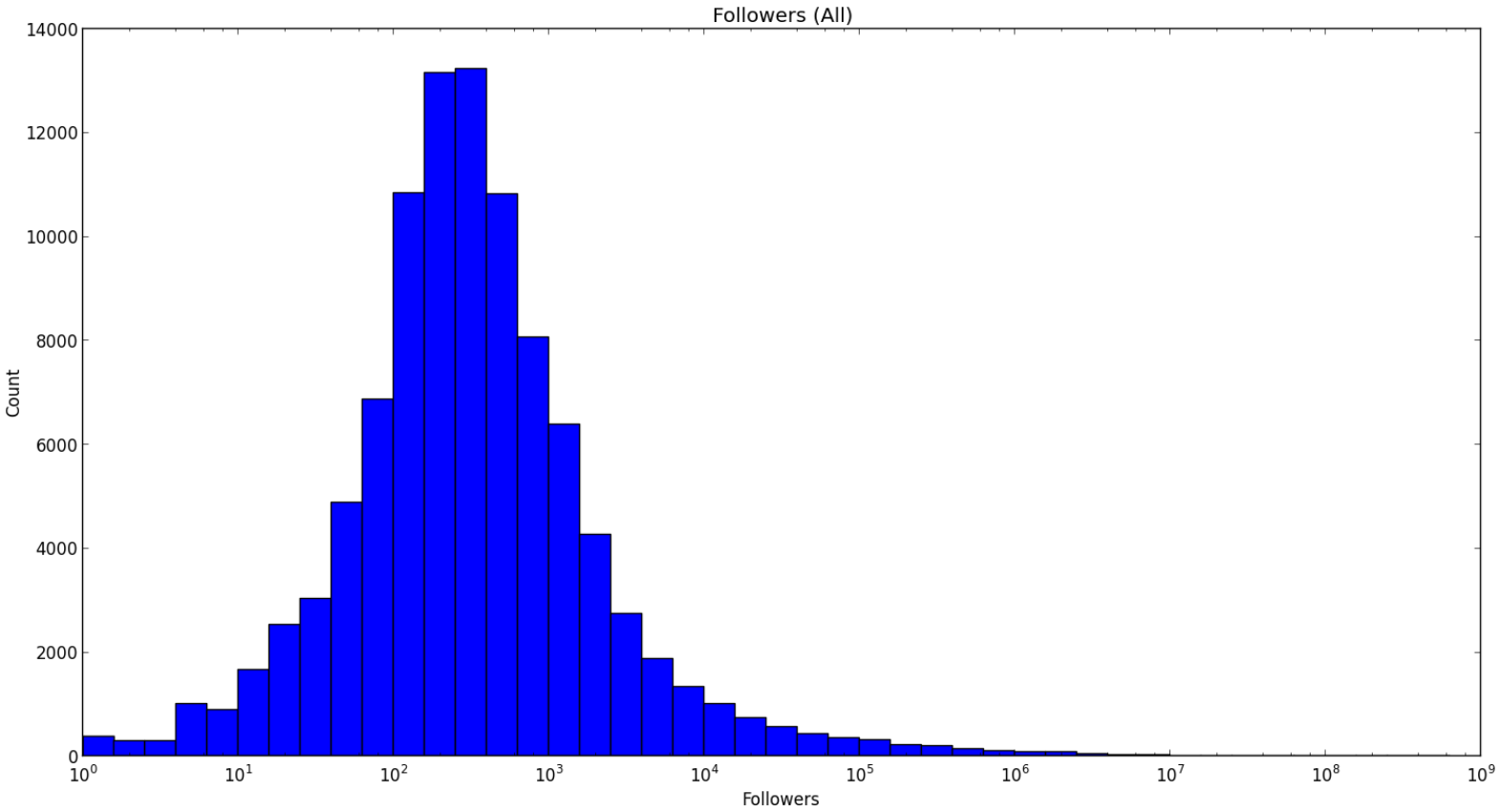

Sebastiano Milardo
@SebMilardo
Happy
Risposta
Elimina
Aggiungi ai preferiti
Altro
1 Giu 13
VS
Justin Bieber
@justinbieber
Happy
Risposta
Retweet
Aggiungi ai preferiti
Altro
173.914
RETWEET
114.198
PREFERITI
6:28 PM - 31 Mag 13

Top Retweeted

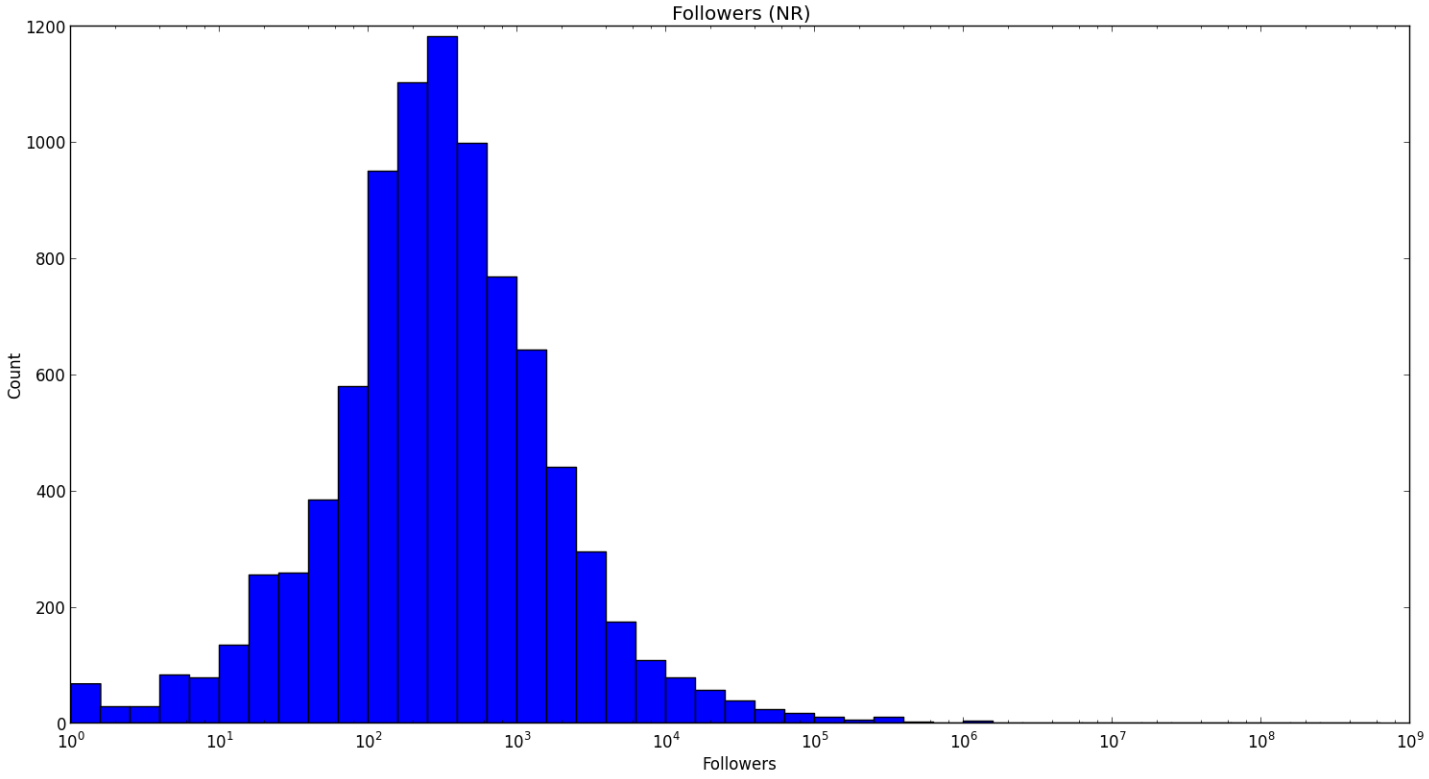
Count	Followers	Screen Name	Text
169266	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Always in my heart @Harry_Styles . Yours sincerely, Louis
157214	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Everybody meet mine and @daniellepeazer new dog Loki :) http://t.co/71lD7azgEW
126309	39610581	justinbieber	RT @justinbieber: My fans... My beliebers... are wild
118880	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: What a great year together @EleanorJCalder :) Love you!!!!
112727	11620806	NiallOfficial	RT @NiallOfficial: Loki is chillin on my bunk! the dog is a legend! Saus monster http://t.co/Yyg9CmY6Nx
108165	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: My heart misses DJ Malik @zaynmalik
71233	11620806	NiallOfficial	RT @NiallOfficial: Love always lads @jlsofficial http://t.co/e1GOSAvMls
66715	11620806	NiallOfficial	RT @NiallOfficial: Barcelona ! Muchas gracias !you were great! Soo loud! Xxx
61954	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Last night was unbelievable. I cannot believe we had the honour to perform at the Olympics! Has to be one of the best …
58188	39610581	justinbieber	RT @justinbieber: @pattiemalle…
57242	13156420	Harry_Styles	RT @Harry_Styles: Cool last na…
54535	11620806	NiallOfficial	RT @NiallOfficial: I can't bel…
52254	11620806	NiallOfficial	RT @NiallOfficial: Morning Bar…
51592	39610581	justinbieber	RT @justinbieber: seeing all t…
49778	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Sad to he…
49162	13156420	Harry_Styles	RT @Harry_Styles: Hapry Son. @…
48931	13156420	Harry_Styles	RT @Harry_Styles: The first ti…
48752	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Had a bri…
45945	11620806	NiallOfficial	RT @NiallOfficial: I hate morn…
44745	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Loving ba…

Count	Followers	Screen Name	Text
169266	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Alway…
157214	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Every…
126309	39610581	justinbieber	RT @justinbieber: My fans.
118880	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: What
112727	11620806	NiallOfficial	RT @NiallOfficial: Loki is
108165	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: My he
71233	11620806	NiallOfficial	RT @NiallOfficial: Love al
66715	11620806	NiallOfficial	RT @NiallOfficial: Barcelo
61954	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Last
58188	39610581	justinbieber	RT @justinbieber: @pattiem
57242	13156420	Harry_Styles	RT @Harry_Styles: Cool las
54535	11620806	NiallOfficial	RT @NiallOfficial: I can't
52254	11620806	NiallOfficial	RT @NiallOfficial: Morning
51592	39610581	justinbieber	RT @justinbieber: seeing a
49778	10698555	Louis_Tomlinson	RT @Louis_Tomlinson: Sad t
49162	13156420	Harry_Styles	RT @Harry_Styles: Hapry So
48931	13156420	Harry_Styles	RT @Harry_Styles: The firs
48752	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Had a
45945	11620806	NiallOfficial	RT @NiallOfficial: I hate
44745	10627917	Real_Liam_Payne	RT @Real_Liam_Payne: Lovin

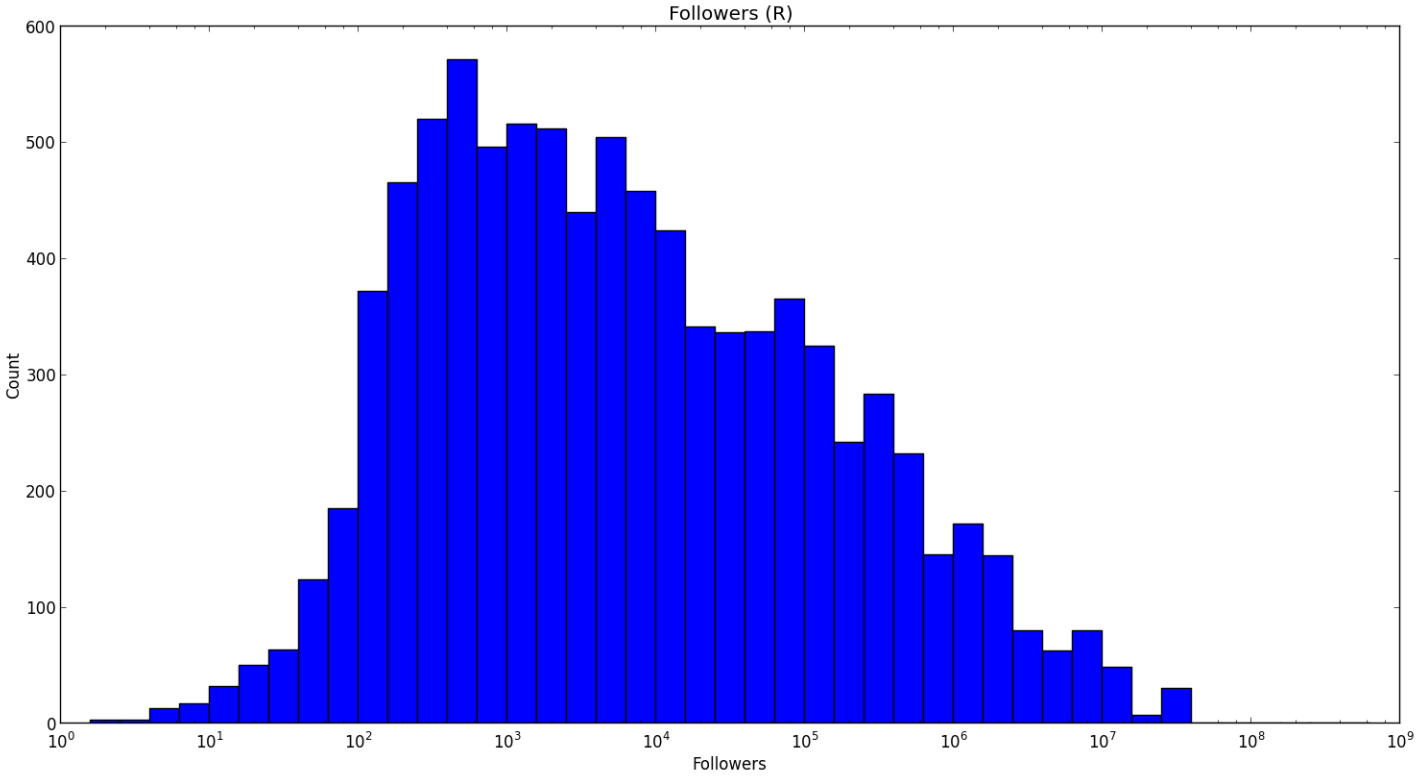
Followers (all tweets)



Followers (non retw.)



Followers (retw.)



Features

- Features related to a particular user.

- Features C: B + User Properties
  - N. Followers
  - N. Friends
  - N. Statuses
  - Account age
  - Verified
  - Default Profile

Results

F-measure: 0.71 - Accuracy: 0.73 - Precision: 0.75 - Recall: 0.68

5.0<=log_flw<7 = True	True : False =	40.1 : 1.0
5.0<=log_frnds<7 = True	True : False =	18.3 : 1.0
verified = True	True : False =	18.2 : 1.0
has(enviroment) = True	False : True =	16.3 : 1.0
log_flw<1.0 = True	False : True =	15.8 : 1.0
has(prometheus) = True	False : True =	11.1 : 1.0
has(f1) = True	False : True =	9.1 : 1.0
has(retweet) = True	True : False =	8.7 : 1.0
has(sale) = True	False : True =	7.5 : 1.0
has(mojito) = True	False : True =	7.3 : 1.0
avg<3 = True	False : True =	7.1 : 1.0
has(syria) = True	False : True =	6.8 : 1.0
has(sopa) = True	False : True =	6.5 : 1.0
has(mi) = True	False : True =	6.4 : 1.0
has(honda) = True	False : True =	5.1 : 1.0
has(toyota) = True	False : True =	4.9 : 1.0
has(nexus) = True	False : True =	3.9 : 1.0
has(manchester) = True	True : False =	3.7 : 1.0
has(architecture) = True	False : True =	3.7 : 1.0
has(alcohol) = True	False : True =	3.7 : 1.0

	F a l s e	T r u e
False	<1421>	408
True	567	<1204>

Features

- There is something missing…

GIORNALE DI SICILIA

Anno LXXXV - N. 114 — Lire 1re

Palermo - Giovedì 3 Maggio 1945

Direzione ed Amministrazione: Piazza Giulio Cesare, 43

La guerra in Italia è finita

Le Forze tedesche e fasciste - un milione di uomini - si arrendono al Comando Supremo Alleato del Mediterraneo

DAL QUARTIERE GENERALE ALLEATO, 3.

(REUTER) - Le forze ne[illegible]iche di terra, del mare e dell'aria al comando del co[illegible]lonnello generale Heinrich Von Vietinghoff - Scheel, Comandante in capo delle Forze tedesche del sud-ovest e del Gruppo d'Armate C si sono arrese incondiziona[illegible]lmente al Maresciallo Sir Harold Alexander, Coman[illegible]dante Supremo alleato nel teatro di operazioni del Me[illegible]diterraneo.

I termini della resa

La firma dei documenti

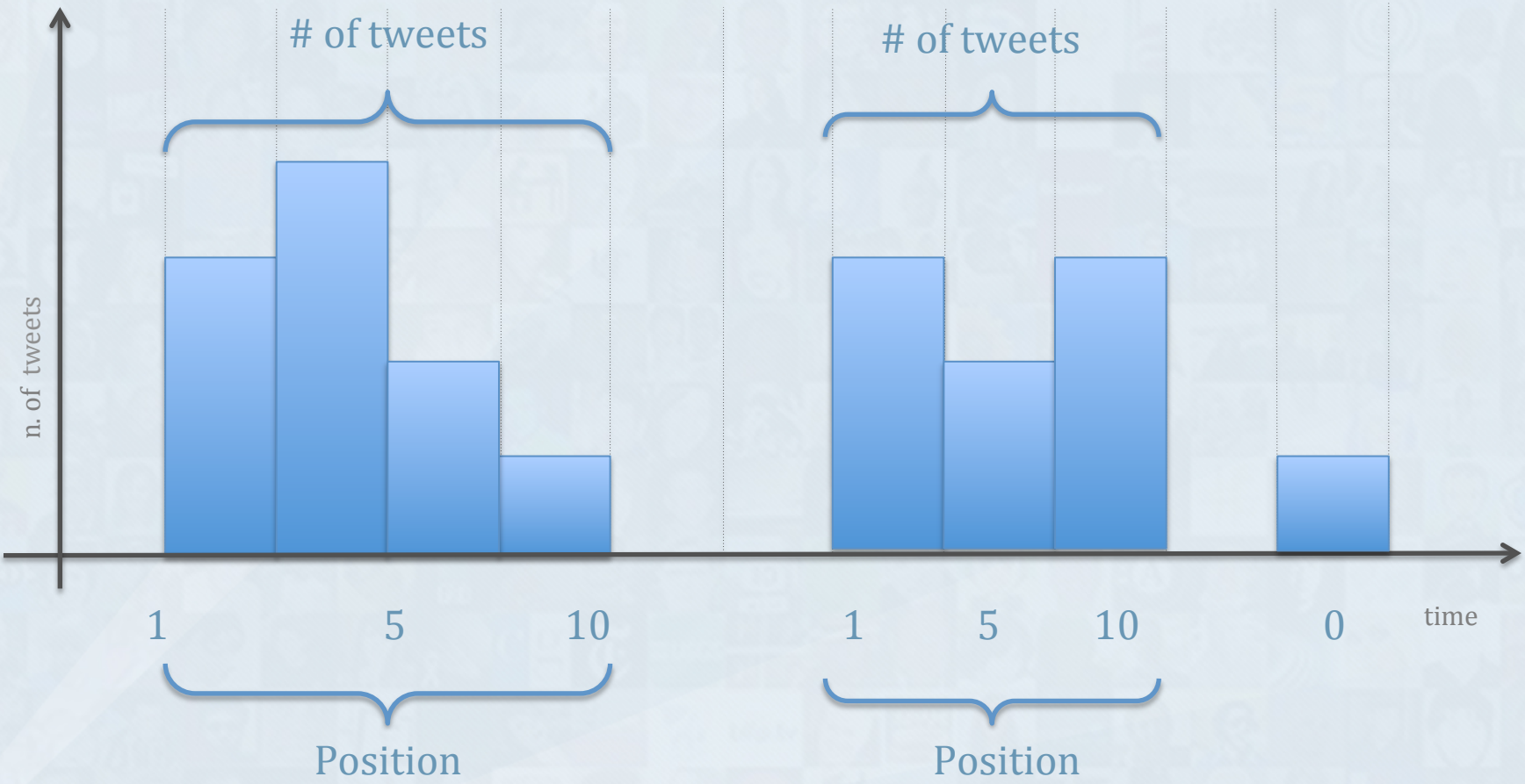
I messaggi di Bonomi del Presidente Truman e dei Comandanti alleati

Il messaggio del Vicecomandante degli Alleati nel Mediterraneo

L'Ambasciatore tedesco si è arreso agli Alleati

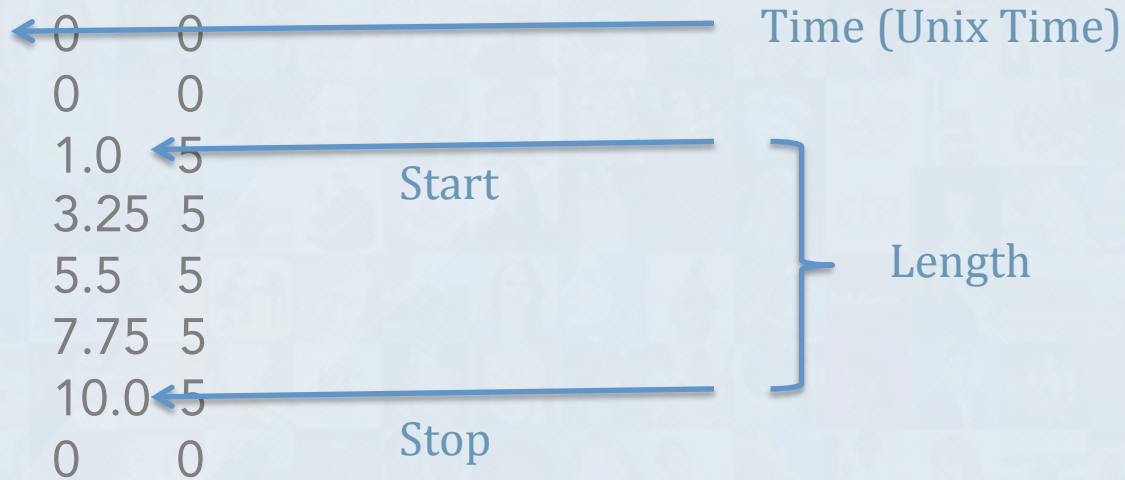
Il nuovo Ministro degli Esteri in Germania

Time

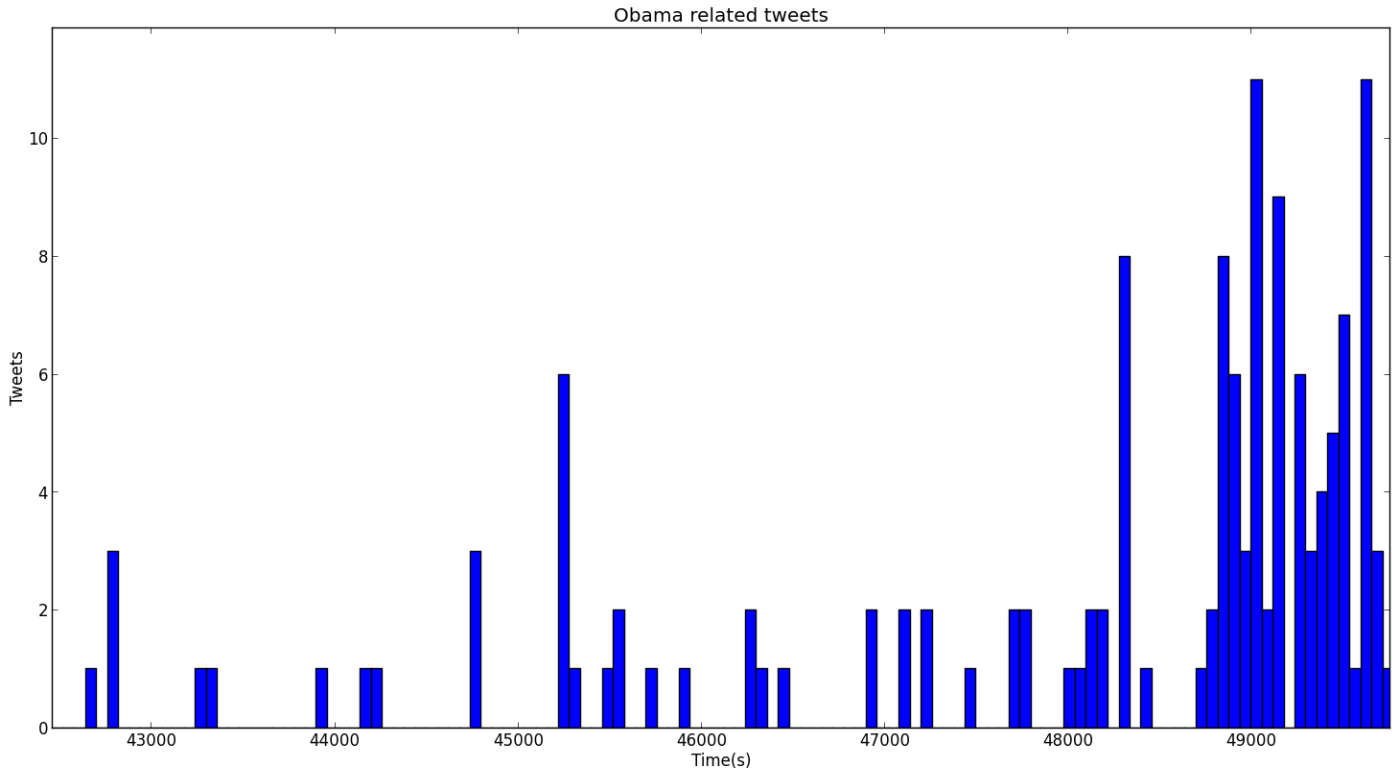

# of tweets
# of tweets
n. of tweets
1
5
10
1
5
10
0
time
Position
Position

Features

1369425527.0	0	0
1369479072.0	0	0
1369485446.0	1.0	5
1369485553.0	3.25	5
1369485829.0	5.5	5
1369486150.0	7.75	5
1369486388.0	10.0	5
1369512154.0	0	0
1369532095.0	0	0
1369539965.0	1.0	3
1369540138.0	5.5	3
1369540254.0	10.0	3
1369540866.0	0	0
1369544351.0	0	0
1369553350.0	0	0



Time


Obama related tweets
Tweets
Time(s)
0
2
4
6
8
10
43000
44000
45000
46000
47000
48000
49000

Features

- Features related to time

- Features D: C + Time Properties
  - Is the first tweet in a series
  - Is the last tweet in a series
  - Is not in a series
  - Position in the series
  - Total number of Tweets in a series

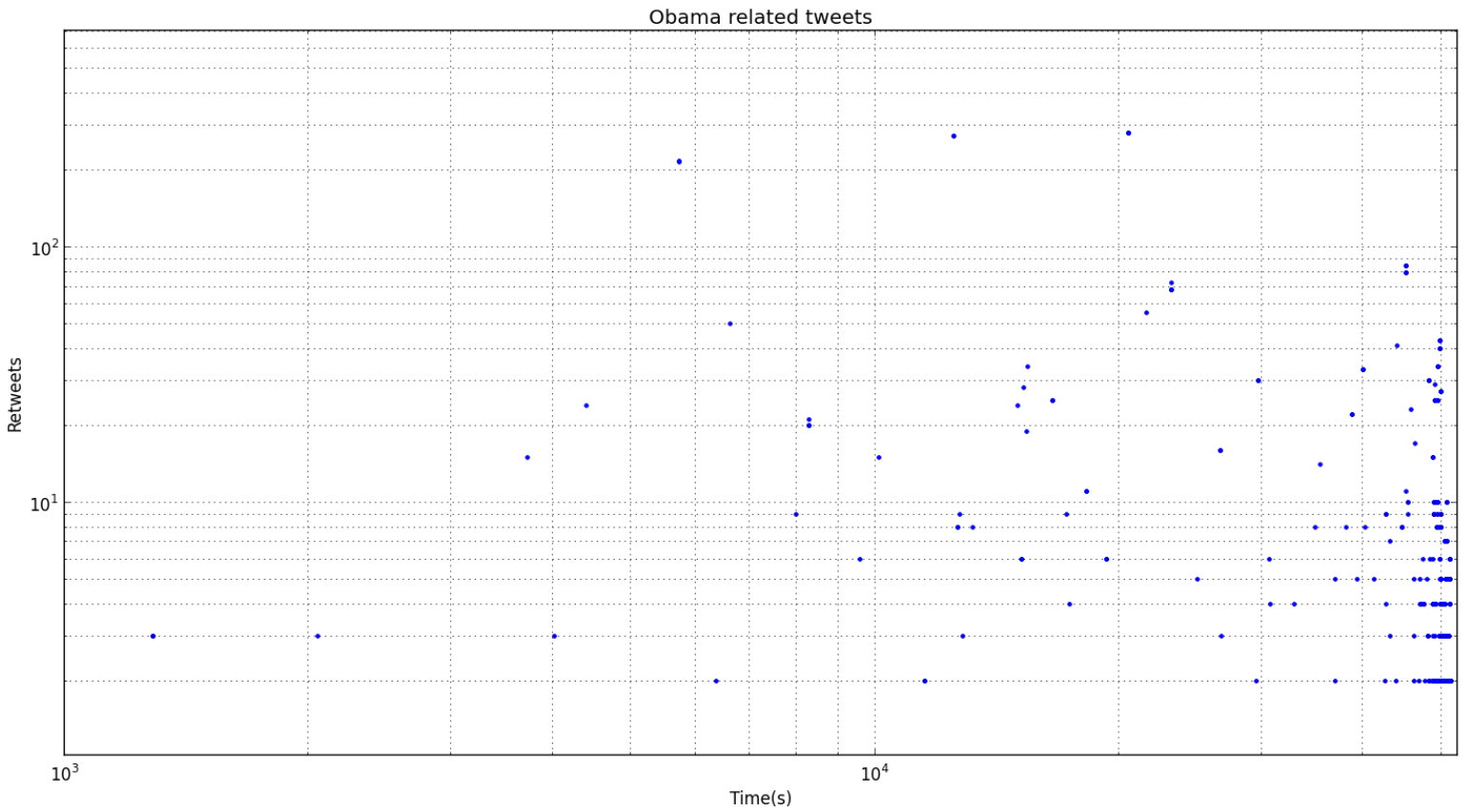
Results

F-measure: 0.86 - Accuracy: 0.87 - Precision: 0.90 - Recall: 0.83

s.pos.==0 = True	True : False =	40.6 : 1.0
5.0<=log_flw<7 = True	True : False =	40.1 : 1.0
5.0<=log_frnds<7 = True	True : False =	18.3 : 1.0
verified = True	True : False =	18.2 : 1.0
has(enviroment) = True	False : True =	16.3 : 1.0
log_flw<1.0 = True	False : True =	15.8 : 1.0
has(prometheus) = True	False : True =	11.1 : 1.0
s.pos.==1 = True	True : False =	10.9 : 1.0
has(f1) = True	False : True =	9.1 : 1.0
has(retweet) = True	True : False =	8.7 : 1.0
has(sale) = True	False : True =	7.5 : 1.0
has(mojito) = True	False : True =	7.3 : 1.0
7.5<s.pos.< 10 = True	False : True =	7.2 : 1.0
avg<3 = True	False : True =	7.1 : 1.0
s.pos.== 10 = True	True : False =	6.8 : 1.0
has(syria) = True	False : True =	6.8 : 1.0
has(sopa) = True	False : True =	6.5 : 1.0
has(mi) = True	False : True =	6.4 : 1.0
s.length<300.0 = True	True : False =	5.3 : 1.0
has(honda) = True	False : True =	5.1 : 1.0

	False	True
False	<1657>	172
True	294	<1477>

...too good to be true



Results

F-measure: 0.85 - Accuracy: 0.86 - Precision: 0.87 - Recall: 0.83

5.0<=log_flw<7 = True	True : False =	40.1 : 1.0
5.0<=log_frnd<7 = True	True : False =	18.3 : 1.0
verified = True	True : False =	18.2 : 1.0
has(enviroment) = True	False : True =	16.3 : 1.0
log_follow<1.0 = True	False : True =	15.8 : 1.0
has(prometheus) = True	False : True =	11.1 : 1.0
s.pos.==1 = True	True : False =	10.9 : 1.0
has(f1) = True	False : True =	9.1 : 1.0
has(retweet) = True	True : False =	8.7 : 1.0
has(sale) = True	False : True =	7.5 : 1.0
has(mojito) = True	False : True =	7.3 : 1.0
7.5<s.pos.< 10 = True	False : True =	7.2 : 1.0
avg<3 = True	False : True =	7.1 : 1.0
s.pos.== 10 = True	True : False =	6.8 : 1.0
has(syria) = True	False : True =	6.8 : 1.0
has(sopa) = True	False : True =	6.5 : 1.0
has(mi) = True	False : True =	6.4 : 1.0
s.length<300.0 = True	True : False =	5.3 : 1.0
has(honda) = True	False : True =	5.1 : 1.0
has(toyota) = True	False : True =	4.9 : 1.0

	False	True
False	<1607>	222
True	296	<1475>

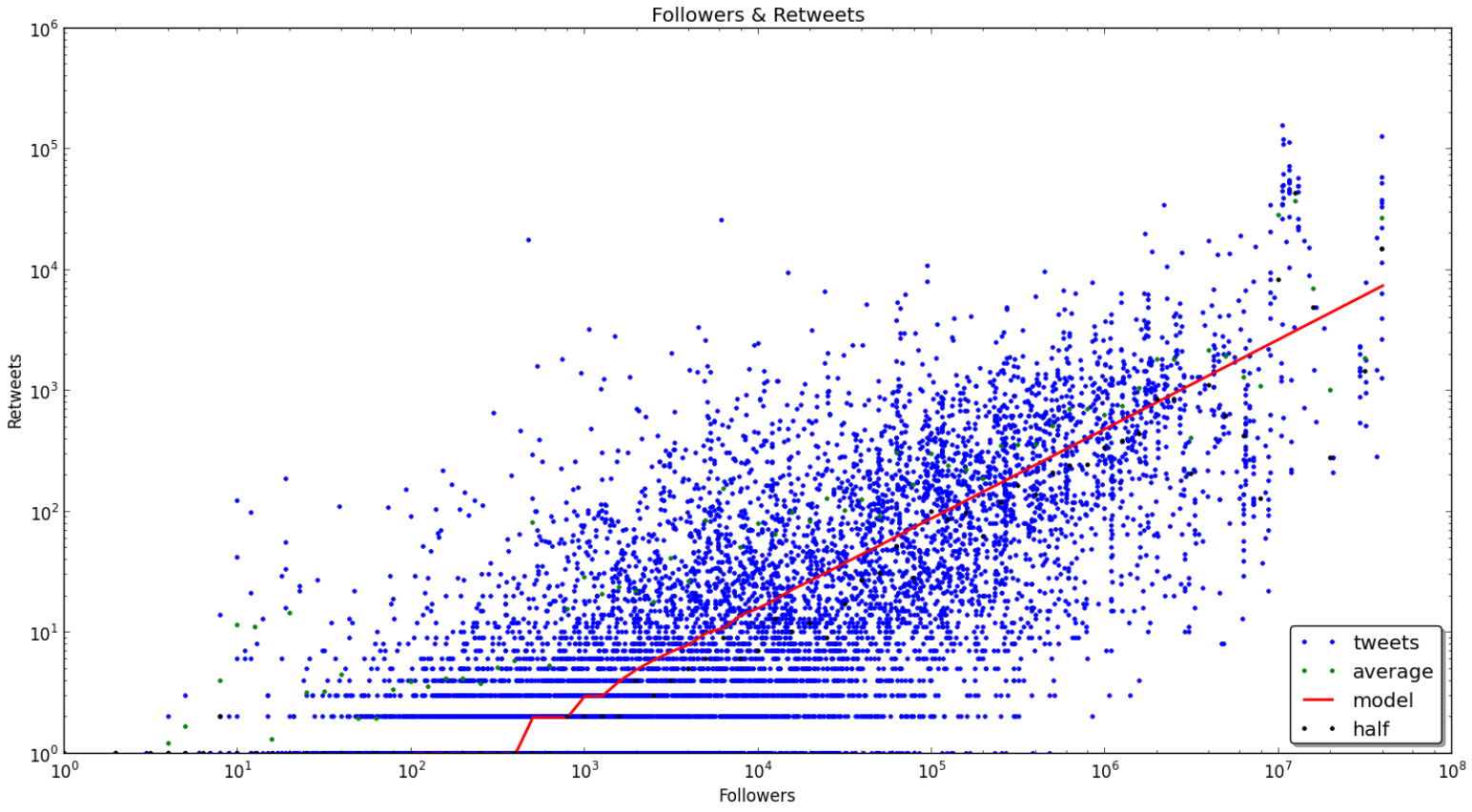
Comparison

Features Set	F-Measure	Accuracy	Precision	Recall
A	0,61	0,58	0,56	0,66
B	0,61	0,62	0,62	0,60
C	0,71	0,73	0,75	0,68
D	0,86	0,87	0,90	0,83
*D	0,85	0,86	0,87	0,83

Problem Description

- *What characterizes a tweet that gets retweeted by many people?*

Results


Followers & Retweets
Retweets
Followers
tweets
average
model
half

Retweet vs Followers

Expected retweet count = a^(log(f)-b)

f = n. of followers

a in [5;6]

b in [2;2.5]

r_score = Retweet count/Expected retweet count

Examples:

5.5^(log(f)-2.37)=1	f=10
5.5^(log(f)-2.37)=100	f=160k
5.5^(log(f)-2.37)=800 000	f=31M

Retweet vs Followers

Was it a golden tweet? (r_score =1 ? Maybe not…)

Retweet vs Followers

- 4319 Good Retweets (r_score>1)
- 4681 Bad Retweets (r_score<=1)

+

Bayesian Magic…

Comparison

Features Set	F-Measure	Accuracy	Precision	Recall
A	0,49	0,57	0,56	0,44
B	0,63	0,57	0,53	0,77
C	0,62	0,67	0,68	0,58
D	0,65	0,69	0,69	0,61
*D	0,64	0,68	0,69	0,59

Results

F-measure: 0.64 - Accuracy: 0.68 - Precision: 0.69 - Recall: 0.59

user<30Days = True	True : False =	12.2 : 1.0
has(goodmorning) = True	True : False =	7.1 : 1.0
has(missing) = True	False : True =	6.0 : 1.0
5.0<=log_statuses<7 = True	False : True =	5.9 : 1.0
1.0<=log_statuses<3.0 = True	True : False =	4.8 : 1.0
has(senate) = True	False : True =	4.7 : 1.0
has(retweet) = True	True : False =	4.2 : 1.0
has(oil) = True	False : True =	4.0 : 1.0
1Month<=user<6Months = True	True : False =	3.7 : 1.0
has(monaco) = True	False : True =	3.5 : 1.0
has(snooki) = True	False : True =	3.5 : 1.0
has(bcci) = True	False : True =	3.4 : 1.0
has(fail) = True	True : False =	3.2 : 1.0
has(uk) = True	False : True =	3.1 : 1.0
has(ufc) = True	False : True =	3.1 : 1.0
has(war) = True	False : True =	3.1 : 1.0
3.0<=log_statuses<5.0 = False	True : False =	3.0 : 1.0
has(lumia) = True	False : True =	2.8 : 1.0
has(union) = True	False : True =	2.8 : 1.0
has(boxing) = True	False : True =	2.8 : 1.0

	F a l s e	T r u e
False	<726>	226
True	348	<500>

Conclusions

- Recipe for a golden tweet:
  - Popularity is the key!
  - Try to be the first one… or the last one
  - At the beginning everything is easier
  - Don’t write too many statuses
  - Ask for a retweet

Questions

Thank You!