CLUSTWEET

Use the power of social media to target your users

Tommaso Agnolazza Alessandro Ciresola Davide Lucchi

Social Media Users

328 Million

2 Billion

700 Million

1 Billion

Learn about users through data

What?

- Habits
- Interests
- Likes
- Dislikes

Why?

- Users' groups
- Sell Targeted Advertisements
- Discover new trends

Learn about users through data

How can we achieve that?

Clustering

Users with the same interests can be gathered into the same cluster

Twitter and K-center Clustering

Dataset:

- 200 000 Tweets
- Tweets between 14-16/04/2016

Main Algorithm:

- K-Center clustering

Framework:

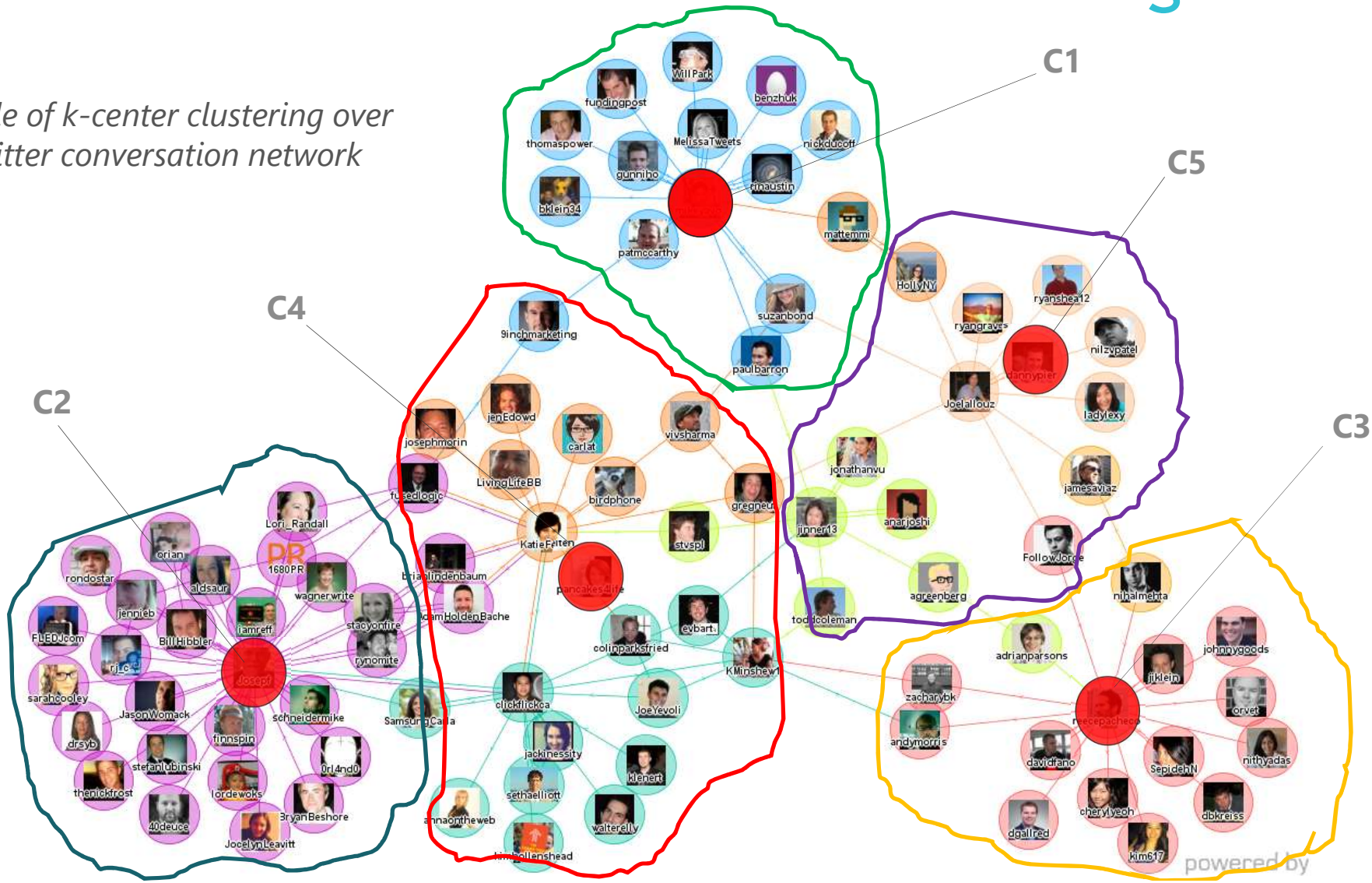
- Apache Spark
- Java 1.8
- Stanford Natural Language Processing

Spark
Lightning-Fast Cluster Computing

Stanford University
S
NLP
Natural Language Processing

Twitter and K-center Clustering

Example of k-center clustering over a Twitter conversation network



Issues with the Twitter fdw

Specific Web Wrappers

Data Source	Type	Licence	Code	Install	Doc	Notes
Database.com	Multicorn	BSD	GitHub			
Dun & Badstreet	Multicorn	PostgreSQL	GitHub			Access to the Data Universal Numbering System (DUNS)
DynamoDB	Multicorn	GPL	GitHub			
Facebook	Multicorn		GitHub			
Fixer.io	based on www_fdw		GitHub			
Google	Multicorn	PostgreSQL	GitHub	PGXN		
Heroku dataclips	Native	PostgreSQL	GitHub			
Mailchimp	Multicorn	PostgreSQL	GitHub			Beta
Parse	Multicorn	MIT	GitHub			
S3	Native	PostgreSQL	GitHub	PGXN		
S3CSV	Multicorn	GPL 3	GitHub			This is meant to replace s3_fdw that does is not supported on PostgreSQL version 9.2+
Twitter	Native	PostgreSQL	GitHub	PGXN		A wrapper fetching text messages from Twitter over the Internet and returning a table
Treasure Data	Native	Apache	GitHub	PGXN		A FDW for Treasure Data internally using a Rust library
Treasure Data	Multicorn	Apache	GitHub			
Google Spreadsheets	Multicorn	MIT	GitHub			

Issues with the Twitter fdw

Issues with the Twitter fdw

- The last stable version dates back to 2012/06/02
- We were able to install Twitter_fdw only for the version 9.1 of Postgre SQL
- Twitter_fdw does not work with the 9.1 version
- Compatibility problems with version 1.1 of the Twitter API

Tweets search

- **Twitter search API :**

1. Requires the registration of a Twitter App
2. Requires the authentication through the Twitter OAuth protocol
3. Provide only results matching with a specific search key

- **Dataset** :

1. Does not need registration and authentication
2. Provide results which belong to different areas of interest

Dataset of 200,000 tweets

Followthehashtag | Home | Features | Plans & Pricing | Help & Tips | Datasets | Contact

200,000 USA geolocated tweets. Free Twitter Dataset

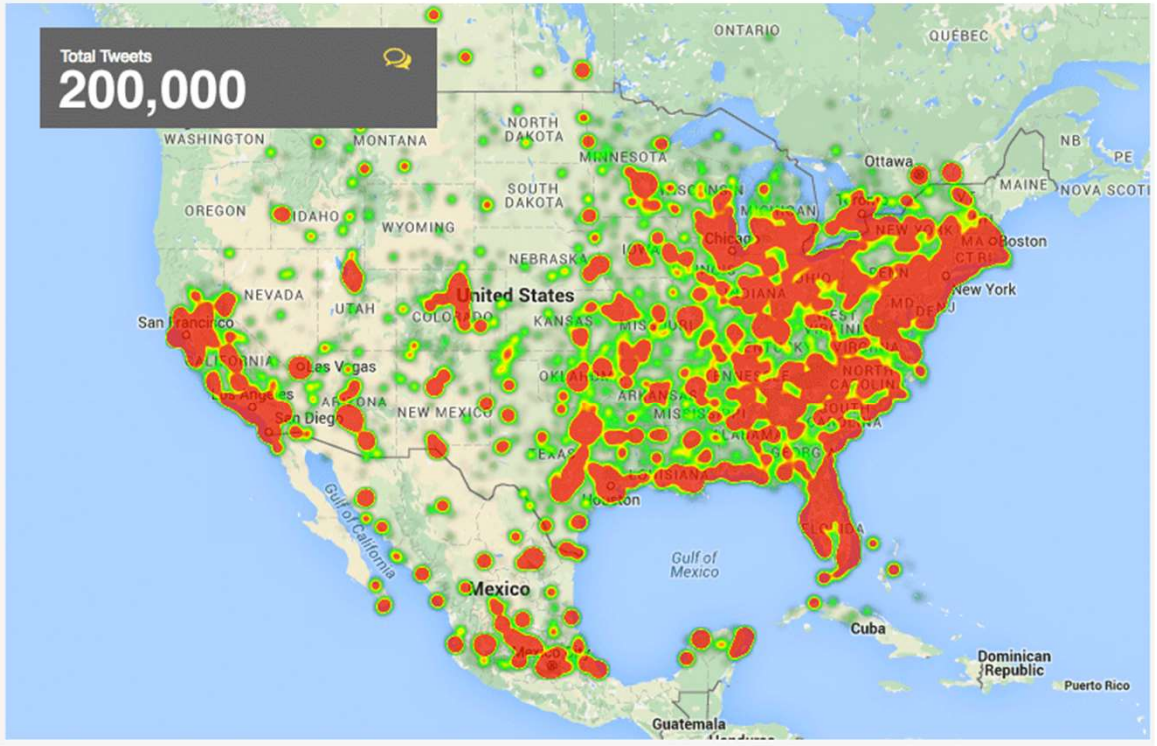
We know data is a valuable research resource, but sometimes can be difficult to get large datasets to be used as corpus. To facilitate researcher´s life we have a new section, "Free Twitter datasets". In this section we will upload different complete large Twitter datasets ready to be used.

Our first Twitter dataset is

USA: Geolocated Twitter Dataset

In this twitter dataset you will get, for free, a database of 200,000 USA geolocated Tweets.

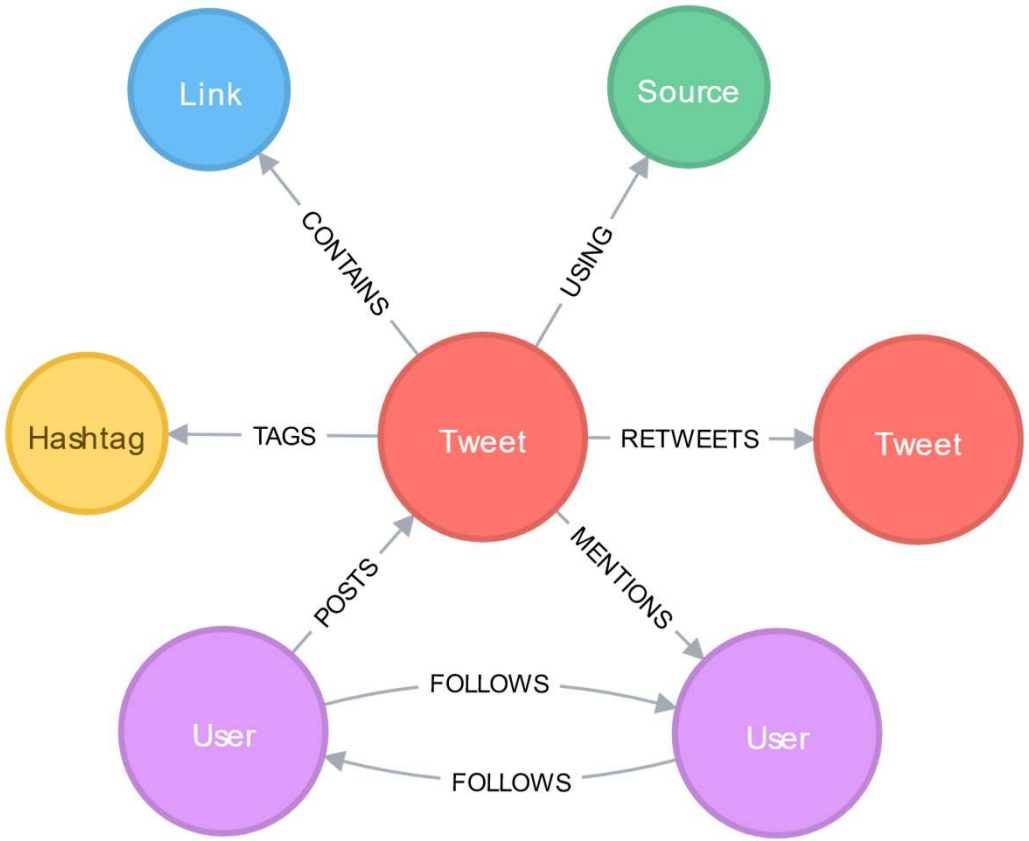
- **Measured Time:** 48h
- **Total Tweets:** 200,000
- **Format:** 4 Excel files
- **Twitter Stream:** Included in "Dashboad" Excel, Sheet: Stream
- **Retweets** are excluded from this search, only original tweets
- **Size:** 47 Mb



Dataset of 200,000 tweets

- **Each row of the dataset is made up of 18 elements** :

1. *Tweet identifier*
2. *Date and hour*
3. *Username*
4. *Nickname*
5. *User biography*
6. *Tweet content (Tweet text)*
7. *Favs(Favorites)*
8. *RTs(Retweets)*
9. *Latitude*
10. *Longitude*
11. *Country*
12. *Place*
13. *Profile picture*
14. *Followers*
15. *Following*
16. *Listed*
17. *Tweet language*
18. *Tweet Url*

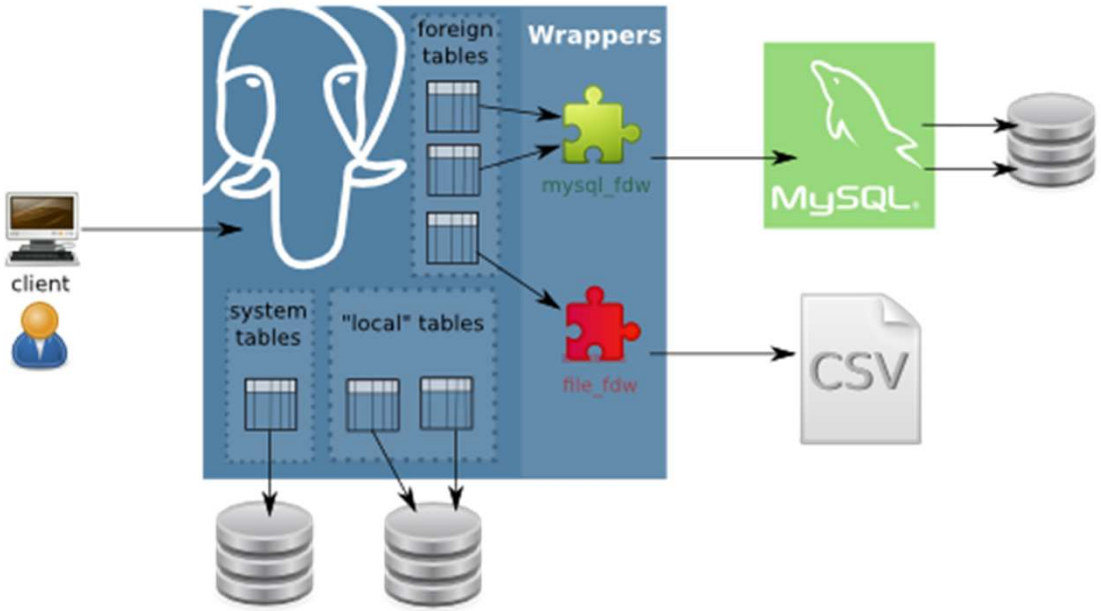


File_fdw

- **Foreign data wrapper able to :**
  1) Read data from text, csv or binary files
  2) Create a foreign table and copy data inside it

- **Options which must or can be set :**
  1) Filename (mandatory)
  2) Format
  3) Header
  4) Delimiter
  5) Quote
  6) Escape
  7) Null
  8) Encoding

- **Import foreign schema function not supported**



Implemented in Postgre SQL starting from 9.2 version

File_fdw

- Installation of the file_fdw with the command : make && make install

- Creation of the extension :

```
CREATE EXTENSION file_fdw ;
```

- Creation of the server :

```
CREATE SERVER server_file_fdw
FOREIGN DATA WRAPPER file_fdw ;
```

- Creation of the user mapping :

```
CREATE USER MAPPING FOR
SERVER server_file_fdw ;
```

File_fdw

- Creation of the foreign table :

```
CREATE FOREIGN TABLE tweets_foreign(
tweet_ID varchar(150) ,
datetweet varchar(50),
hour varchar(50),
username varchar(50) ,
nickname varchar(50),
biography varchar(250),
tweet_content varchar(300) ,
favs varchar(150),
rts varchar(50),
latitude varchar(200) ,
longitude varchar(200) ,
country varchar(200) ,
place varchar(200),
profile_picture varchar(200),
followers integer,
following integer,
listed integer,
language varchar(10),
url varchar(200)
) SERVER server_file_fdw
OPTIONS (format 'csv', header 'true' , filename '/tmp/lista_tweets_USA.csv' , delimiter ',' , null '' );
```

File_fdw

- Creation of the local table :

```
CREATE TABLE tweets_local
AS SELECT * FROM tweets_foreign ;
```

- Removal of the rows in the table with the attributes following, followers or listed equal to NULL :

```
DELETE
FROM tweets_local AS T
WHERE T.followers IS NULL OR T.following IS NULL OR T.listed IS NULL ;
```

Loss of 10k rows from the local table

Why have we converted the csv file into a table before doing the clustering?

Couldn't we take data directly from the csv file?

Three main reasons :

1. Skimming data in order to avoid rows with null attributes

2. Collect information about the dataset through targeted queries

3. Project requirement

Data is the new oil

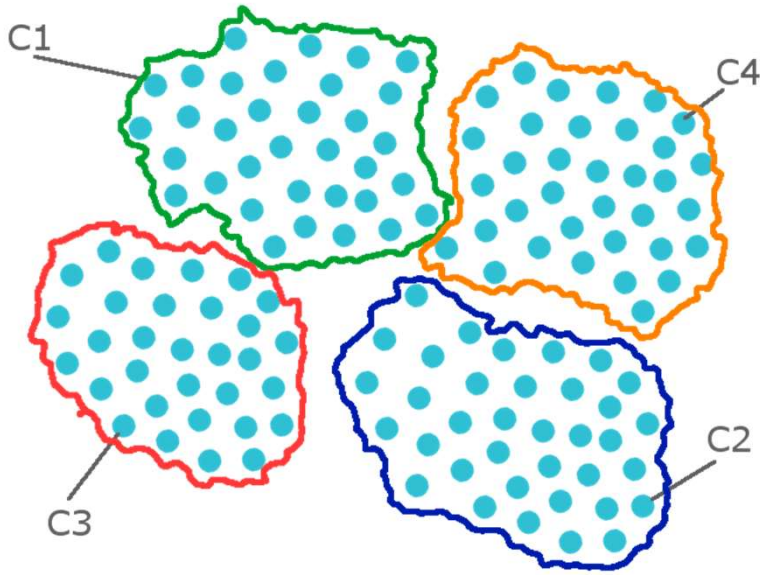
We need to find it,
Extract it, refine it,
Distribute it and
monetize it.

David Buckingham

Delving Into Details

Objective

Clustering tweets contents on common themes



Preprocessing on the tweets

Word2Vec →

1. Lower case
2. Removing non sense words
3. Removing all symbols

Parametres

K′ = 2K

N = 200000

K=150

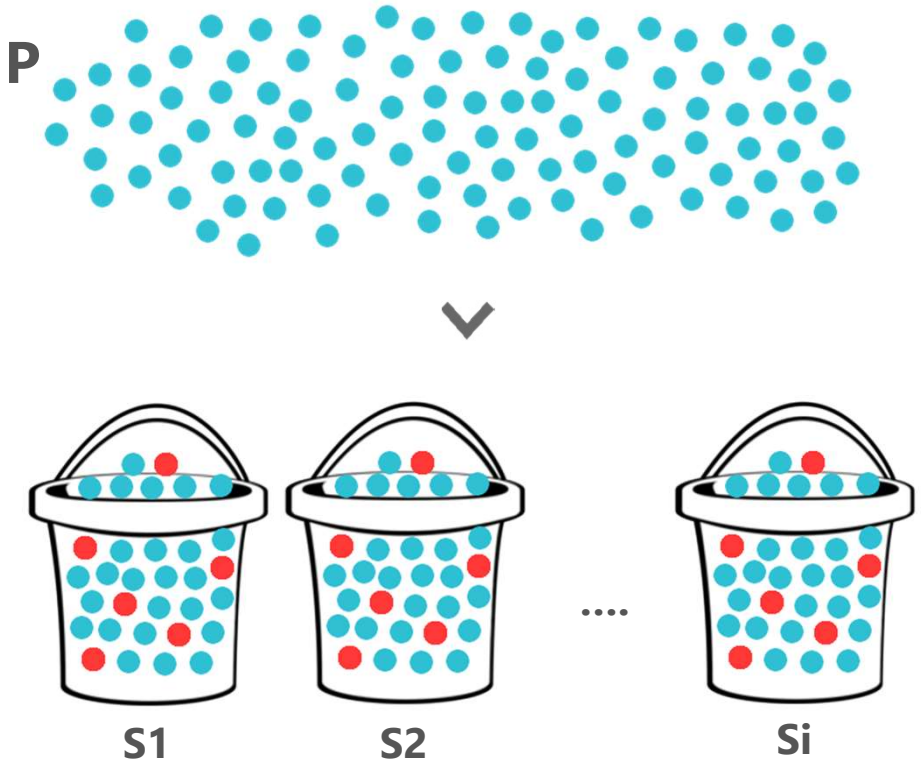
$P = \sqrt{\frac{N}{K}}$

Dim=100

Spark environment

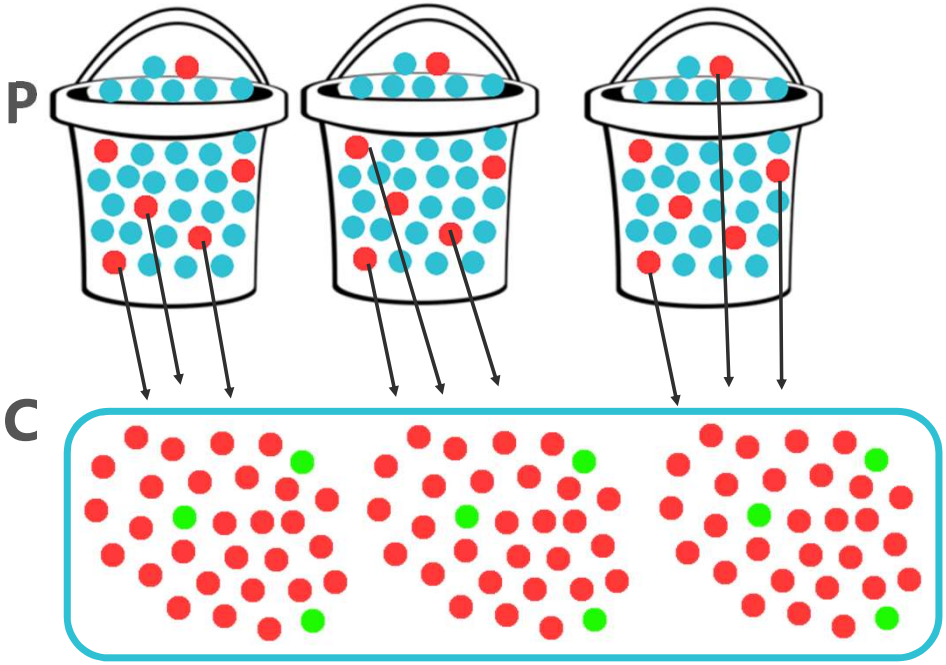
K-Center Clustering with Spark:

- Divide points P randomly in buckets
- Determine 2K centers in each bucket
  - Group all centers in set C
  - Determine K centers in set C
    - Assign points to clusters

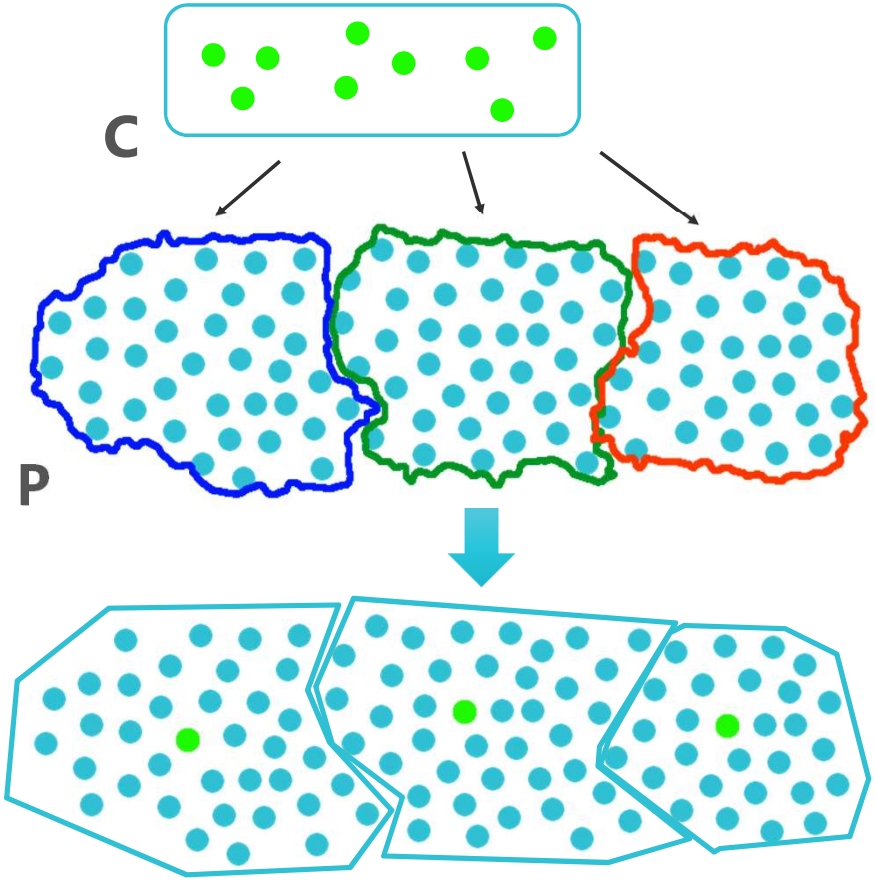


Round 1

Spark environment



Round 2



Round 3

The secret of change is to focus all of your energy, not on fighting the old, but on building the new

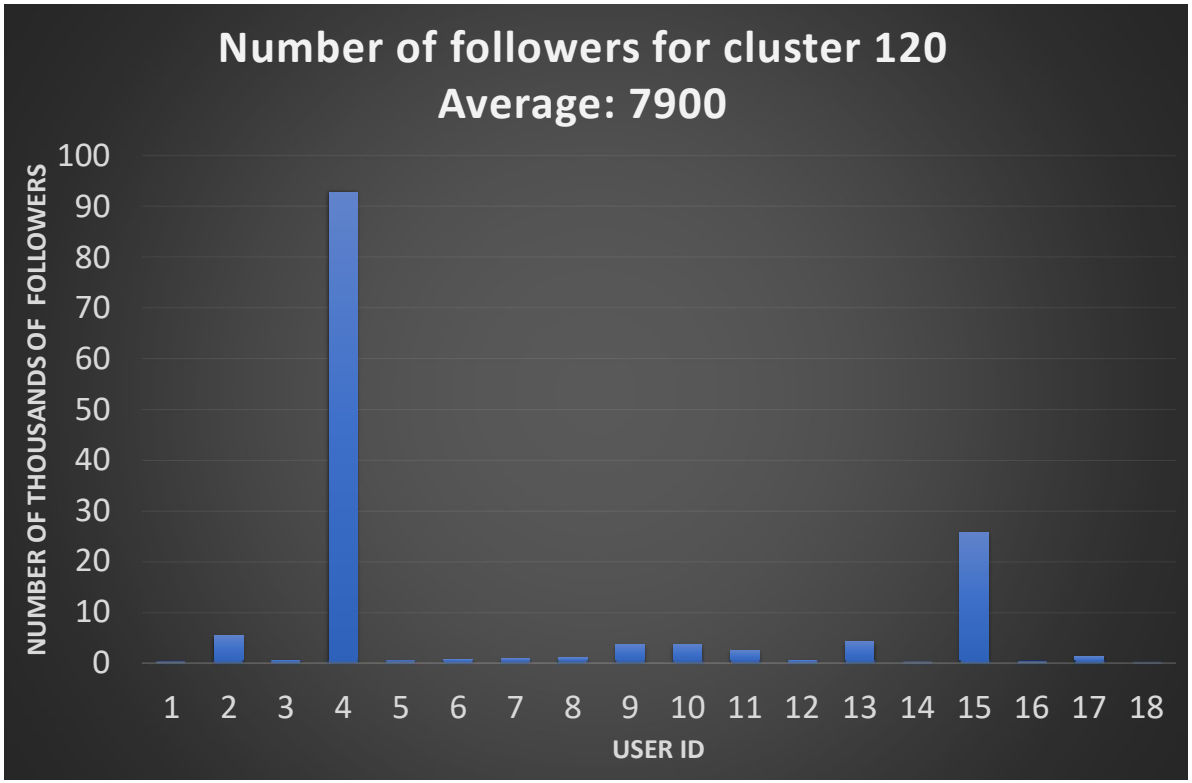
Socrates

RESULTS

What our clustering Doesn't do:

Group tweets based on the number of followers:

- Followers average is not representative



RESULTS

What our clustering DOES:

It groups tweets based on their contents – tweets in the same cluster contain similar topics

How to assess it?

- Assign tags to tweets
- Search for the most frequent tags
- Calculate cluster entropy
- Calculate frequent tags entropy

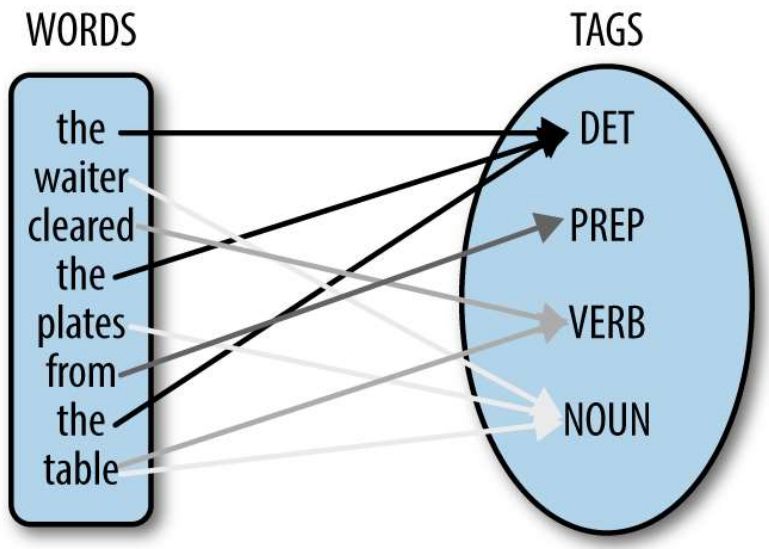
Assign Tags to Tweets
Part-Of-Speech Tagger (POS Tagger)

Stanford Log-linear Part-Of-Speech Tagger

How does it work?

- It reads text in some language
- It assigns parts of speech to each word

For each cluster only the most frequent nouns are kept – threshold 25%



Entropy of a Cluster

$$E(C) = \sum_{for\ each\ noun\ i} \left(\frac{n_{i,C}}{n_C}\right) ln(\frac{n_{i,C}}{n_C})$$

Where:

- $n_{i,C}$ - Number of nouns i in cluster c
- n_C - Number of nouns in cluster c

Why?

To see how representative of a noun a cluster is

Entropy of a Noun

$$E(n) = \sum_{for\ each\ cluster\ c} \left(\frac{n_{i,C}}{n_i}\right) ln(\frac{n_{i,C}}{n_i})$$

Where:

- $n_{i,C}$ - Number of nouns i in cluster c
- n_i - Total number of nouns i

Why?

To see if the noun is distributed in different clusters

RESULTS

Let's take a look - 1

Cluster: 67
Size: 5192 tweets
Most frequent Noun: I'm
Frequency: 97.5%
Cluster entropy: 0.76
Noun entropy: 1.60

SAMPLE TWEETS

I'm at Rivercenter Mall in San Antonio, TX
I'm at Southwest Plaza Mall in Littleton, CO
I'm at @BangSalonDc in Washington, DC
I'm at @UnderArmour Brand House in Baltimore, MD
I'm at park place diner in Lansdale, PA
I'm at @PetSmart in Southfield, MI
I'm at 6PACK in Queens, NY

Applications:

- Targeted advertisements
- Labeling of users with their interests
- Store places that users visit

RESULTS

Let’s take a look - 2

Cluster: 147
Size: 31228 tweets
Most frequent Noun: #Jobs
Frequency: 71.68%
Cluster entropy: 1.01
Noun entropy: 0.27
Other frequent Nouns: #Job, #CareerArc

SAMPLE TWEETS

Want to work at CVS Health? We're #hiring in #Denton, TX! Click for details: https://t.co/65C1Pz8R3C #Retail #Job #Jobs #CareerArc

#IT #Job alert: Senior Infrastructure Provisioning Coordinator | Genuine Parts NAPA | #Atlanta, GA https://t.co/FbwrHkOxwA #Jobs #Hiring

Want to work in #McLean, VA? View our latest opening: https://t.co/v5DDZ5PSqR #Healthcare #Job #Jobs #Hiring #CareerArc

Applications:

- Find people who are looking for a job or offering one
- Make a list of available jobs

RESULTS

Let's take a look - 3

Cluster: 83
Size: 3629 tweets
Most frequent Noun: photo
Frequency: 49%
Cluster entropy: 0.66
Noun entropy: 0.57
Links: most point to Instagram

SAMPLE TWEETS

Just posted a photo @ Randolph-Macon College https://t.co/QL5iXtXhfx

On the lake Tahoe in Sierra Nevada with a paddleboard. Photo by @standup_paddle #lake... https://t.co/e8iC0UhyIF

Just posted a photo @ Grand Canyon Skywalk https://t.co/qgruAKCPVY

Just posted a photo @ San Antonio, Texas https://t.co/9SH5PCoUFH

Applications:

- Labeling of users with their interests
- Store places that users visit
- Get information about their Instagram account

How to improve it?

Improve tweets pre-processing:

- Transform similar nouns into a single noun (#job,#jobs,#Job -> #job)
- Substitute word abbreviations with their extended form (w/ -> with)

Evaluation:

- Merge together similar clusters
- Repeat the entire process on clusters that don’t contain a frequent word

Thanks for your attention

Get in touch with us:

Tommaso Agnolazza
Tommaso.agnolazza@studenti.unipd.it

Alessandro Ciresola
Alessandro.ciresola@studenti.unipd.it

Davide Lucchi
Davide.lucchi@studenti.unipd.it