Twitter Project

Information Retrieval
a.a. 2011/2012

P. Ferragina
-Dipartimento di Informatica, University of Pisa-

Twitter Dataset

- 173049 Users
- Followers of 7 italian politic channels


Beppe_grillo

Pbersani

Idvstaff

Matteorenzi

Govberlusconi

Gianfranco_fini

Rossipresidente

- Semantic Annotation of followers tweets using TAGME

Tagme DEMO

Data and Tools

zola.di.unipi.it

username: ir2011

password: project@2011

User already configured to use TAGME java library for semantic annotation

Tweets

- ~/data/twitter/TweetTwitter-20110912_160840.tweet

Users Info

- ~/data/twitter/TweetTwitter-20110912_160840.usersinfo

Graph

- ~/data/twitter/TweetTwitter-20110912_160840.graph

Tweets File Format

- Every row contains a JSON objects

```
{“tweets”: [...] ,”user”: ”1111111”,timeline=”me/in” }
{“tweets”: [...] ,”user”: ”222222”,timeline=”me/in” }
{“tweets”: [...] ,”user”: ”333333”,timeline=”me/in” }
                    ……
                    ……
{“tweets”: [...] ,”user”: ”777777”,timeline=”me/in” }
```

- “tweets” : array of tweet objects
- “user” : id of the user
- “timeline”
  - me: tweets written by the user
  - in: tweets read by the user

attention: in the file you can find two entries for the same user, one for timeline “me” and the other for timeline “in” (consecutively)

Tweet Object


```
{"contributors": null,
 "truncated": false,
 "text": "RT @Dile: Mah oddio, non che Renzi sia stato proprio supportato eh http://t.co/jl75dat",
 "in_reply_to_status_id": null,
 "id": 111874934254473216,
 "entities": {
        "user_mentions": [{"indices": [3, 8],"id": 514723,"id_str": "514723","name":         "Diletta","screen_name": "Dile"}]
        "hashtags": [],
        "urls": [{
                "indices": [67, 86],
                "url":   "http://t.co/jl75dat","expanded_url":"https://twitter.com/festadem2011/status/111841272519606272",
                "display_url": "twitter.com/festadem2011/s\u2026"}]
         },
"author":   {Author Informations...},
"retweeted": false,
"coordinates": null,
"source": "web",
"in_reply_to_screen_name": null,
"in_reply_to_user_id": null,
"retweet_count": 1,
"id_str": "111874934254473216",
"favorited": false,
"retweeted_status": {Retweet informations...},
"user": { User Informations....}
}
```

Graph File Format

52352494
370324592
145775612
10653372
370315008
370298975
370296062
175429876
370291616
…..
52105134

18762875
186754160
355109448
53454426
9741792
370299829
…….

52352494	pbersani
18762875	matteorenzi
97734603	rossipresidente
314059678	gianfranco_fini
14078646	Idvstaff
19067940	beppe_grillo
17336414	Govberlusconi

To know followers of a specific channel!

User info File Format

- Every row contains a JSON objects

{ "crawler_time": 1315515085, "follow_request_sent": false, "profile_use_background_image": true,
"id": 202192235, "verified": false,
"profile_image_url_https": "https://si0.twimg.com/profile_images/1143950433/Screen_shot_2010-10-13_at_17.43.42_normal.png",
"profile_sidebar_fill_color": "DDEEF6", "geo_enabled": true, "profile_text_color": "333333", "followers_count":120,
"protected": false, "id_str": "202192235", "default_profile_image": false, "location": "Milan, Italy",
"status": {}, "utc_offset": 3600, "statuses_count": 84,
"description": "MR & Associati Comunicazione affianca organizzazioni pubbliche, politiche e private per valorizzarne l'identit\u00e0 e rendere efficace i processi di comunicazione .", "friends_count": 352, "profile_link_color": "0084B4",
"profile_image_url": "http://a1.twimg.com/profile_images/1143950433/Screen_shot_2010-10-13_at_17.43.42_normal.png",
"notifications": false, "show_all_inline_media": true,
"profile_background_image_url_https": "https://si0.twimg.com/images/themes/theme1/bg.png",
"profile_background_image_url": "http://a0.twimg.com/images/themes/theme1/bg.png", "profile_background_color": "C0DEED", "screen_name": "mrassociati",
"lang": "it", "following": false, "profile_background_tile": false, "favourites_count": 0, "name": "MR & Associati",
"url": "http://www.mrassociati.it", "created_at": "Wed Oct 13 14:12:52 +0000 2010", "contributors_enabled": false,
"time_zone": "Rome", "profile_sidebar_border_color": "C0DEED", "default_profile": true,"is_translator": false,
"listed_count": 5, "classified_by_this": 1}

http://twitter.com/#!/mrassociati

How to annotate a text

```
String lang="it";
String  example="Ubuntu attacca, obiettivo smartphone e tablet";

ConfigManager.init("/1/disc3/home/ir2011/ir.tms.xml");
RelatednessCache r=new RelatednessCache(lang);
Annotator annotator = new Annotator(lang);
ArticleSearcher as= new ArticleSearcher(lang);

List<Annotation> annots=annotator.annotates(example,r);//Also without r

System.out.println("\nAnnotations for \""+example+"\"");

for(Annotation a:annots){
        if(a.getSense()!=-2)//Not disambiguated
                System.out.println(as.getTitleByDoc(a.getSense()));
}
//Other Useful Method
a.getRho();
```

OUTPUT

Annotations for "Ubuntu attacca, obiettivo smartphone e tablet"
Ubuntu
Smartphone
Tablet PC

Relatedness between topics

```
String lang="it";
int a=61471 ;  //Smartphone
int b=561859;  //Tablet PC

ConfigManager.init("/l/disc3/home/ir2011/ir.tms.xml");
ArticleSearcher as= new ArticleSearcher(lang);
RelatednessCache r = new RelatednessCache(lang);

//Compute relatedness
float rel=r.rel(a,b);
System.out.println("\nRelatedness value
                    between \""+as.getTitleByDoc(a)+
                    "\" and \""+as.getTitleByDoc(b)+"\": "+rel);
```

OUTPUT

Relatedness value between "Smartphone" and "Tablet PC": 0.7269514

Phrase Similarity

```
String lang="it";
String example1="Ubuntu attacca, obiettivo smartphone e tablet";
String example2="Niente Android Ice Cream Sandwich per Nexus One";

ConfigManager.init("/l/disc3/home/ir2011/ir.tms.xml");
Similarity s= new Similarity(lang);

//Compute Similarity score
float sim_score=s.sim(example1,example2,0.15f);

System.out.println("\nSimilarity between \""+example1+"\"
                   and \""+example2+"\"\nScore:"+sim_score);
```

OUTPUT

Similarity between:
"Ubuntu attacca, obiettivo smartphone e tablet"
and
"Niente Android Ice Cream Sandwich per Nexus One"
Score:0.8257334

Tasks

- Preliminary phase
  - Parse files
  - Annotate texts

- Analysis phase
  - Differences between topics of politcs followers
  - ...

Useful Informations

- Compilation
  - javac filename.java
- Execution
  - java -Xmx2000M classfile
- For long computations
  - nohup yourcommand & (When the connection is closed the program continues to run)
  - tail -f nohop.out (to see the progress)
- For Technical Question and Support
  - d.vitale@di.unipi.it