Twitter Trends Annotator

Hauptseminar "Information Retrieval"

WiSe 15/16
Institut für Computerlinguistik
Universität Heidelberg

Evgenij Reznik

January 2016
Contents

- Introduction
- Twitter
- Google Knowledge Graph
- Implementation
- Demo
- Special Linguistic Cases
- Application
- Sources
• Introduction

• Get current trends and provide a short summary of each

• Considerations
  • Which location?
  • Which language?
• **Twitter**

• Digital real-time microblogging platform, social network or a most publicly accessible online diary

• Communication takes place via "tweets"
  • short text posts (max. 140 characters) created by users around the world

• Used by individuals, organizations, companies and mass media

• Users can follow other people (e.g. celebrities), organizations, companies, etc.
• Twitter: Trends /1

• Topics that are popular now, rather than topics that have been popular for a while
• Help to discover the hottest emerging topics
• Can be tailored for the user based on location and who s/he follows
  • Not considered in my example
• Twitter: Trends /2

• Can start with a hashtag (#)
  • Marks tweets as relating to a topic
  • "serve as a vehicle for otherwise unconnected participants to be able to join in a distributed conversation" (Bruns & Burgess 2011)

• Can consist of compound words
  • #NationalHugDay

• Can consist of regular words (without #)
  • Popular phrases, not annotated by #
  • “Thorsten: Da fallen auch Ballaststoffe ab #IBES” (@RTLde)
• Twitter: API

• Provides access to whole Twitter functionality
  • Access, read, write Twitter data
• OAuth is needed in order to get authorized access to Twitter's API
  • "An open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications."
    (http://oauth.net/)
• Responses are provided in JSON
https://api.twitter.com/1.1/trends/place.json?id=1

[{
  "as_of": "2012-08-24T23:25:43Z",
  "created_at": "2012-08-24T23:24:14Z",
  "locations": [
    {
      "name": "Worldwide",
      "woeid": 1
    }
  ],
  "trends": [
    {
      "tweet_volume": 3200,
      "events": null,
      "name": "#GanaPuntosSi",
      "promoted_content": null,
      "query": "%23GanaPuntosSi",
      "url": "http://twitter.com/search/?q=%23GanaPuntosSi"
    },
    {
      "tweet_volume": 4200,
      "events": null,
      "name": "#WordsThatDescribeMe",
      "promoted_content": null,
      "query": "%23WordsThatDescribeMe",
      "url": "http://twitter.com/search/?q=%23WordsThatDescribeMe"
    },
    {
      "tweet_volume": 1200,
      "events": null,
      "name": "#10PersonasQueExtrañoMucho",
      "promoted_content": null,
      "query": "%2310PersonasQueExtrañoMucho",
      "url": "http://twitter.com/search/?q=%2310PersonasQueExtra%C3%B1oMucho"
    }
  ]
}]
• **Google Knowledge Graph**

  • Semantic technology and knowledge base used by Google
  • Improves the quality of its search engine with semantic-search information collected from various sources
    • Wikipedia
    • Wikidata
    • The World Factbook
  • Provides a Search API

For more information please refer to
• Haenelt_Google-KnowledgeGraph_Teil1.pdf
• Haenelt_Google-KnowledgeGraph_Teil2.pdf
• Google Knowledge Graph: Request

• HTTP request
  
  • [https://kgsearch.googleapis.com/v1/entities:search](https://kgsearch.googleapis.com/v1/entities:search)

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>string</td>
<td>A literal string to search for in the Knowledge Graph</td>
</tr>
<tr>
<td>ids</td>
<td>string</td>
<td>A list of entity IDs to search for in the Knowledge Graph.</td>
</tr>
<tr>
<td>languages</td>
<td>string</td>
<td>The list of language codes (defined in ISO 639) to run the query with, for instance en.</td>
</tr>
<tr>
<td>types</td>
<td>string</td>
<td>Restricts returned entities to those of the specified types.</td>
</tr>
<tr>
<td>intend</td>
<td>boolean</td>
<td>Enables indenting of JSON results.</td>
</tr>
<tr>
<td>prefix</td>
<td>boolean</td>
<td>Enables prefix (initial substring) match. e.g. Jung → Jung, Jungle, Jung-ho Kang</td>
</tr>
<tr>
<td>limit</td>
<td>number</td>
<td>Limits the number of entities to be returned.</td>
</tr>
</tbody>
</table>
Taylor Alison Swift is an American singer-songwriter and actress. Raised in Wyomissing, Pennsylvania, she moved to Nashville, Tennessee, at the age of 14 to pursue a career in country music.
• Implementation

• Used tools
  • **PHP**: webservice
    - Authentication and communication with Twitter API
    - `twitter-api-php`
      - Simple PHP Wrapper for Twitter API v1.1 calls
  • **JavaScript**
    - **JQuery**: Library which simplifies use of JavaScript
    - **AJAX**: Communication with APIs, sending asynchronous requests
Flowchart

1. Choose from list
2. Location
   - Retrieve location
   - Retrieve WOEID
3. Retrieve language
4. Language correct?
   - yes: Retrieve Trends
   - no: Enter desired language
5. Pre-processing
6. Remove stop words?
   - no: Annotate Data (GKG)
   - yes: Remove stop words
• Location

• Current location

  1) Get user's geolocation using the HTML Geolocation API
  2) Retrieve location that Twitter has trends for, closest to user's location
      • Return WOEID (Where On Earth Identifier)
      • Unique identifier for any location on earth, assigned by Yahoo!

• Choose location from list
  • Content comes from Twitter, as well
  • Already with WOEID
• Language Detection

• In order to annotate the trends correctly, the right language needs to be chosen

1) Retrieve language based on chosen location

2) Send request to REST Countries
   • e.g. https://restcountries.eu/rest/v1/name/germany

3) Receive response with country information
   • capital, population, languages, etc.

4) If detected language is wrong, users can specify their own
• Preprocessing

• Removing # if applicable
• Splitting words if CamelCase

#HappyBirthdayCalumHood
=> Happy Birthday Calum Hood
• Stop words removal

• User can decide, whether stop words need to be removed
  • Zusammenbruch in Kreuth => Zusammenbruch Kreuth

• May increase annotation accuracy

• Removal is based on the detected language in the prior step

• Currently, 41 languages are supported
  • Entire list available at http://www.ranks.nl/stopwords
Getting Twitter trends for: **United Kingdom**

- Arsenal
- #Labyrinth
- #growingupwithmyname
- Costa
- Giroud
- Floreni
- Mertesacker
- Walcott
- #GiantDinosaur
- Fabregas
- #NHCOFA
- Ozil
- #TraditionallySubmissive
- Emirates
- Everton
- Cech
- Cesc
- Hazard
- Campbell
- Willian
- Koscielny
- Mikel
- Zidane
- Courtois
- Clattenburg

---

**Zidane**

*Category: Soccer player*

Zinedine Yazid Zidane, nicknamed "Zizou", is a retired French footballer and current manager of Real Madrid. He played as an attacking midfielder for the France national team, Cannes, Bordeaux, Juventus and Real Madrid.

• Special Linguistic Cases: Abbreviations

• Popular abbreviations can be handled
  • After removing # during the preprocessing step

• Examples:
  • #ibes: Ich bin ein Star – Holt mich hier raus!
  • #dbsd: Deutschland sucht den Superstar
  • #MCFC: Manchester City F.C.
• Special Linguistic Cases: Meaningless words

• There are difficulties when there is no "main" word
  • After applying stop word removal, it is unclear what to look for
    • #EinWortWerbung => Wort Werbung
    • #TheyJustDiscovered => Discovered
• Words written without CamelCase notation are treated like a single word
  • It is unclear where to separate them
    • #growingupwithmyname => growingupwithmyname
• **Special Linguistic Cases: Synonyms**

  • Not always clear which word is meant (domain-specific)
    
    • Robben
    
    • Context: Sports or animals?

  

  Person ≠ Animal
• **Outlook**

• Detect events (e.g. sports, festivals)
  - #HSVFCB => Hamburger SV vs. FC Bayern München (Bundesliga)
  - #GERHUN => Germany vs. Hungary (Handball Euro 2016)

• Connect with whatthetrend.com
  • Users can add explanation, *why* trends are trending right now
  • Unfortunately, very sparse
Applications

- Business cases
  - Companies are interested in what is suddenly increasing in popularity
    - Fashion, music, movies, celebrities, etc.
  - very effective real-time decisions can be made
- Natural disaster
  - Detect movements of earthquakes, storms, etc.
    - People replace high tech sensors (6,000 tweets / second)
  - Take appropriate actions in time
  - Twitter is now being used to track earthquakes
- Much more ...

http://www.digitaltrends.com/social-media/twitter-is-now-being-used-to-track-earthquakes/
• Sources

• Twitter

• Google Knowledge Graph

• Tools
  • https://dev.twitter.com/rest/public
  • https://developers.google.com/knowledge-graph/
  • http://oauth.net/
  • http://www.ranks.nl/stopwords

• Applications
  • Chang, L. (2015). Twitter is now being used to track earthquakes. [online] Digital Trends. Available at:  
  http://www.digitaltrends.com/social-media/twitter-is-now-being-used-to-track-earthquakes/  