**Intelligent Web Applications**

**Seminar II – Information Extraction with GATE**

Student: Filipe Fernandes Fiedler

NIA: 203229

1: What is Twitter?

Twitter is a social web app where users create microblogs and interact with other users posts. Each micropost has the maximum length of 280 characters, known as tweets, and may contain hashtags, images, videos, links and mention to other users. Each user can interact with a tweet liking it, sharing it,

2: Find a real application which uses Twitter for Opinion Mining and explain what it does.

3: Check the web site of Twitter4j explain what the library is useful for (comment on some of its main features).

4: What is the objective of the program SearchByKeywords? Where and how is the crawled information stored?

The objective is return all the lastest tweets that can be found through a search in Twitter webpage using a specific keyword in the form of Json files. The crawled information is stored in the folder data/keywords/[name of the keyword used] in a JSON file, that allows the storage of the content and metadata about the tweet.

5: List the text of the first tweet of each of the searches. (Use one of the suggested Websites to display the tweets)

**Teófilo Otoni** - *@Cruzeiro Quero ser sócio do Cruzeiro mas moro no interior em Teófilo Otoni, como faço para me associar?*

**Pudgemont** - Pero quien le esta pagando a pudgemont toda su parafernalia publicitaria ahora mismo??

6: What is the objective of the program SearchByUser?

The objective is to catch all the tweets in the timeline of a specific user. The program takes as input the user ID and creates files with …

7: Show the tweets your program returns and also a screenshot of the user timeline, mark in the timeline the tweets you have retrieved.

8: What is the objective of the program SearchByLocation?

The objective is to create a listener that is able to catch each tweet that was made inside an specific geographical area. The program receives as input the coordinates of a retangle area in the world and prints in almost real time all the tweets made in that area since the start of the execution of the program.

9: Show at least one tweet for each of the searches indicating the landmark used in the search.

10: Consult the gate.ac.uk documentation to find information about the TwitIE plugin. Explain its main functionalities

TwitIE is a GATE pipeline that contains features that helps to process the specifities of texts from tweets. Its main contents are social media data language identification, trained in the most common languages such as English, French, German, Dutch and Spanish; a specific tokenizer to twitter, that it’s able to handle emoticons, hashtags, user names and URLs; a specific part-of-speech tagger for Twitter and a text normalization PR.

11: List the information identified in the corpus:

**Twitter 1**:

* Date and time: Sun Nov 19 18:58:00 +0000 2017
* Organization: NHS
* userID: NHSMillion

**Twitter 2**:

* userID: Shaker\_aphra
* location: Turkey and US
* URL: https://t.co/wB15ZZV15W

**Twitter 6**:

* Organizations: BBC, CNN and SKY

**Twitter 9**: According to Twitter, the language is “es”, which is the abbreviation for spanish and this information can be found at Tweet annotation and in the lang attribute