Activity Report WWSSS'18

Student Details

Name: Fernando Santos Sanchez
Id: 28069862
Programme: iPhD in Web Science
Activity: Attending the WSTNET Web Science Summer School 2018
Day of departure: 27/07/2018
Return date: 06/08/2018

Activity schedule

First day

Welcome and registration

Introduction to Web Science

Data Science in the context of the Web

Data Science process

Machine learning for Web Science

Philosophy & ethics of Web Science

Poster presentation

Second day

Text mining and social media analysis with GATE

Information extraction with GATE

Temporal information retrieval & Web Archive analysis

Truth discovery

Third day

Web analytics for everyday learning

Personal Web analytics

Group work

Social event: BBQ

Fourth day

Social computing & crowdsourcing

Scaling up human-computation through Microtask Crowdsourcing

Knowledge construction in Wikipedia

Social event: City tour and dinner

Fifth day

Web semantics & knowledge graphs

Introduction to scalable Big Data processing

Visualization & Visual Analytics

Group work presentation

Sixth day

Using advertising audience estimates for population studies

Copyright and Data Protection Law in Data Mining

Digital social sciences (Complex networks analysis)

Closing event

Activity summary

I attended the WSTNET Web Science Summer School 2018 as part of my PhD formation, the activities, presentations and tutorials presented during this summer school have greatly enriched my understanding of the vanguard research topics in the Web Science area. During this period, I attended 4 social activities, 18 thematic talks, and 1 practical tutorial. Also, I collaborated in a team effort to design a novel solution for a Web Science problem, and presented my research via poster (attached document). In the following lines I describe in more detail the activities which I found most relevant for my research.

The Data Science Process

This talk approached a practical approach for working with data, it explored the full range of task from data acquisition and pre-processing to the analysis, modelling and interpretation. It explored the issues of having incomplete, incompatible, noisy, and biased data. It presented common approaches to solve this issues when the data is also distributed or too big to be processed whole:

1. Transformation
2. Data Cleaning
3. Data Combination
4. Selection and Sampling
5. Big Data Methods

It showed the importance of carefully exploring the data instead of mechanically following a general procedure:

- Do you need other/more data?

- Do you need data in another format?
- Which are the important variable?
- What the value distributions are saying?
- Detect and analyse the outliers, missing values and anomalies
- Generate hypothesis, test them and discard them if necessary

Finally, it showed the importance of carefully record and report the exact procedure and the actions taken during the process of analysing the data.

Information Extraction

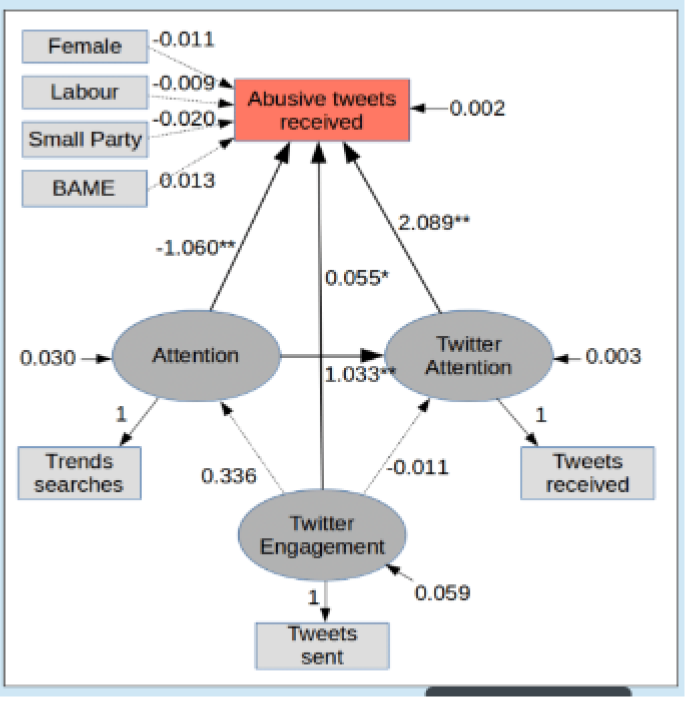
This presentation approached several interesting facts about Twitter data which must be considered when employing it for data analysis:

- 500 million tweets per day
- 24% of all internet male users
- 21% of all internet female users
- 37% of Twitter users are 18-29
- 25% of Twitter users are 30-49
- US has 67 million Twitter users, Mexico 23.5 million, and Germany only 1 million
- 1/3 of all the internet users in Saudi Arabia are on Twitter, making it the country with highest penetration
- Top 10 most followed: media personalities except Barack Obama, YouTube and Twitter.

Which highlights the need to observe research considerations when employing this data source, the results of any analysis of Twitter data must be interpreted in consideration with the how representative the Twitter users are of the general population and the type of common content that is generated. It presented a case study of employing the GATE platform to analyse tweets to find hate speech:

1. Tweet Collection: employ the Twitter streaming API
2. Tokenisation: identify individual components of the tweet
3. Normalization: normalize spelling and abbreviations
4. Pos-Tagging: identify parts of speech
5. Named Entities recognition: identify mentions of entities like people, locations, organizations and products.
6. Political Entity recognition: identify political entities based on MPs mentions and information links from YourNextMP and DBpedia
7. Topic detection: match tweets against a topic ontology
8. Geolocation: match tweets to NUTS regions based on place tags and user home locations
9. User classification: it classified users into groups like journalist, charity, member of the public.
10. Semantic search: it employs a semantic search engine, Mimir, to search and visualise the tweet text and annotations.

11. Employ structural equation modelling



The case study identified 404 abusive terms, but only annotated them when they were used in specific situations. It then employed Structural Equation Modelling using the Lavaan package in R to answer questions like: Did men get more abuse because more men are conservatives? And did more prominent people get more abuse because they got more tweets?

Gate Practical Tutorial

Here we followed a practical tutorial on how to employ the GATE graphic user interface to analyse a given corpus of data. It started by presenting the architecture of the GATE platform form text processing:

- At the base are the documents which are text compositions
- This texts can be explored as strings with the use of character offsets, as sentences or as tokens.
- The GATE platform offers the utility of identifying part of speech
  - Personal pronouns
  - Verbs
  - Prepositions
  - Number
  - Nouns
- It offers morphological analysis like identifying the base form of the employed verbs and tools for knowledge engineering, like finding patterns for dates:
  - Recognizing months' names
  - Recognizing date structure
    - {number}{Month}{number}
    - {number}{slash}{number}{slash}{number}

It introduced the GATE pipeline for common information extraction:

- Pre-processing: tokenisation, sentence splitting, morphological analysis and POS tagging
- Entity finding (Gazetter lookup and NE grammars
- Co-reference: alias finding, orthographic co-reference
- Export the results: to a database, XML or ontology.

The tutorial explored a corpus of 10 documents (Language resources) in which the ANNIE plugin for Named Entity Recognition was run. The entities of each document were explored in employing the annotation sets in display pane of each document. The structure of composite entities was presented via the annotation stack view were the token composition was presented.

Group Work

Context: Complexity is commonly measured by quantitative analysis of the structure of the sentence and the size of the word. This does not translate well into languages like German, Chinese and Japanese in which word length is a poor indicator of complexity.

The problem: Can more meaningful indicators be extracted from written resources (webpages) to measure the complexity of the text.

The idea: Employ context data to score the complexity of the text. Text may vary in complexity depending on the subject it approaches, and the named entities on the text can be used to characterize the topics of the text.

The algorithm:

- Create a database of relevant named entities from texts on several topics
- Evaluate the complexity of the texts from the selected topics by employing Mechanical Turk as a Crowdsourcing platform.
- Employ the page rank algorithm to determine a complexity score for each entity depending on the complexity of the texts it appears on and its relevance to the text:
    - Entities that appear many times or at the beginning are more relevant
    - Entities that appear on complex texts have higher scores
- Employ GATE to identify the entities from an acquired test corpus
- Employ the scores for the weighted entities to compute a complexity score for the test texts.
- Evaluate the complexity of the test corpus with Crowdsourcing and determine the validity of the approach when correlated with other commonly employed sentence metrics
- Collect text sources with low levels of recognition for relevant named entities as possible

Presented results:

The team worked together for 5 hrs in which time a methodology was designed but deemed unviable given the time and resource limitations. A simplified approach was presented by selecting a 10 Webpage dataset with similar format from the given corpus, employing the manual annotation of complexity scores for each page in the set, identifying ~10K entities in the pages and employing a naïve model to score the complexity of each entity based only on the complexity scores of the pages it appeared in.