Network Analysis on Twitter Data

Progress Presentation

Betül Bilgin
Hakime Öztürk
N.Yağmur Aydın

Outline

- Project Definition
- Progress Made So Far
  - Data Collection
  - Data Processing
  - Ontology
- Remaining Work

Project Definiton

In our Social Semantic Web project, we aim to :

- Make word normalization on twitter data using a slang word dictionary
- Create an ontology to represent semantic relations between word versions and users
- Perform network analysis on these relations and induce new possible relations between different words

Project Progress - Data Collection

- 3000 tweets with the keyword ‘xmas’ are collected using Direnaj in json format
- Data is parsed using jq with the following filter:

```
[.["direnaj_pages"][]["results"][] | {poster: .tweet.user.screen_name, texts: .tweet.text, loc: .tweet.user.location, language: .tweet.lang}] | .[] | [.poster, .texts, .loc, .lang]
```

- and resulting data is in the following format:
  - "lee_lizard",
  - "@MrDanRigby nope not a thing I do it all on Xmas eve !!,
  - "Liverpool",
  - "en"

Project Progress - Data Processing

- By using filtered data, JSON objects are manually created to see the functioning of the parsing.

ELEMENTS IN CURRENT OBJECT:

name=name, value=Lauren_Mayhew

name=location, value=Moncton, NB

name=language, value=en

name=tweet, value=RT @KetoKrate: Christmas is here! Gift yourself delicious Keto snacks.

Project Progress - Data Processing

- Parsed data are represented as Java objects and tweets are splitted into tokens. Resulting objects have five fields:
  - username
  - tweet text
  - tweet words
  - location
  - language
- Tweet objects which include slang words are processed using our slang word dictionary which can be found at our github repository.

Project Progress - Extracting Slang Words

First, each TweetObject is filtered according to their languages, where the english ones are kept.

There are total 2540 English TweetObjects we processed.

Then, each word in each TweetObject is cross-checked with our Slang Dictionary.

Project Progress - Slang Dictionary

Idiot	I
ideal	I-D-L
identification	ID
jealous	Jealz
OK	K
over	KKK
knock	KOK
laugh	L
later	L8R
Last-In-First-Out	LIFO
hello	lo
loser	luser
love	luv
mom	M
mate	M8
mates	M8s
yeah	mhhm
mobile	mob
message	MSG

okay	OK
all correct	OK
over	ova
Australia	OZ
Partner	P
Pizza	P-ZA
person	P3r50n
Passengers	PAX
people	peeps
pictures	pix
photos	pix
Please	PLS
Please	PLZ
post-modern	po-mo
computer	puter
password	pw
own	pwn
owned	pwnt

Project Progress - Extracting Slang Words cont.

So far, we kept two lists:

- Slang Word - Frequency
- Cooccurence(Slang Word1, Slang Word2) - Frequency

We intend to keep one more list:

- User - slang Words

Project Progress - Extracted Slang Words

Total 34 different Slang Words are extracted.

Table: Top-ten frequent slangs

Slang	Actual Word	Frequency
ı	idiot / I	288
we	whatever	88
4	for / four	58
u	you	25
f	friend	12
w/	with	11
20	location	11
r	are	10
ur	your	10
pls	please	8

Co-occurence is **not** spotted which indicates that we need much more data.

Project Progress - Ontology

- We have built an ontology named TweetWordOntology to represent semantic relations based on Twitter users and their word usages.

- We created:
  - user-user relations
  - user-word relations
  - word-word relations

- The specifications about the TweetWordOntology are on our github repository.

Remaining Work

- Represent data on the TweetWord ontology
- Import RDF data to Gephi
- Network Analysis