Es hora de decir bye-bye

Using python for sociolinguistic analysis

Who am I

Figure 1: me

Who are you

Figure 2: Photo by davide ragusa on Unsplash

What is this talk about

- scraping webpages
- processing linguistic data
- interpreting linguistic data

Motivation

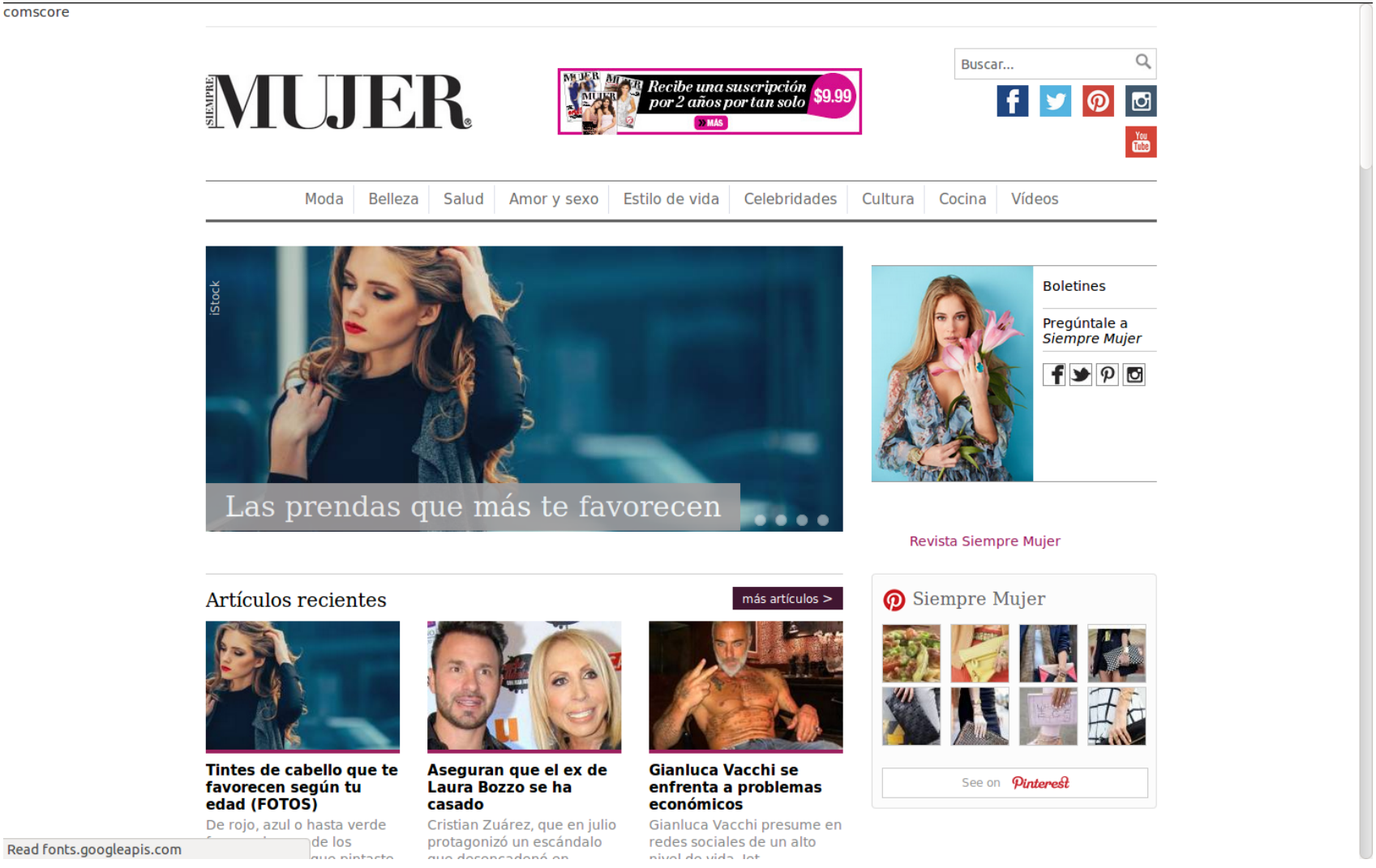

Figure 3: http://siempremujer.com/ (visited: Aug 13th 2017)

Si tu amiga se la pasa poniendo memes que critican tu postura religiosa, se burlan abiértamente de los homosexuales y apoyan frases que parecen de la época de la esclavitud que te incomodan. Es hora de decir bye-bye.

What have I done?

- scrape articles: 1419
- possible problems: messy html structure of the magazine
- alternative approach: systematic crawling with `scrapy`?
- strip articles from html markup

Token segmentation

- “self-esteem”/“self esteem”
- “machine learning”
- “you’re”
- segment article texts in tokens: non-trivial
- `nltk.word_tokenize()`

What have I done? (II)

- download and clean openoffice spelling dictionaries
- intersection and differences (Spanish common and Spanish regional);
- label each token according to appearance in dictionaries (EN, ES, ES_regional –> multiple labels possible; PUNCT; UNK)

Problems:

- maintainer: dictionaries non-exhaustive: (seldom) words are missing
- linguistic: not every word is present in all its forms (e.g. singular, plural, m, f forms; or every tempus, modus, etc. for the verbs)
- linguistic: some words exist in both languages (e.g.: “me”, “no”, “hay”, “sin”, “soy”, “sea” etc.)

- consider all tokens marked as EN only and their surrounding context
- 3271 distinct appearances
- use NLTK `Texts.ConcordanceIndex()` for generating context

```
Article # 53990
[baby, ['EN']]
Displaying 4 of 4 matches:
epollo púrpura , y todos los "bebés" baby carrots ,
, y todos los "bebés" baby carrots , baby remolachas ,
és" baby carrots , baby remolachas , baby greens ,
ts , baby remolachas , baby greens , baby arúgula .

--------------------------------------------------

Article # 51108
[charter, ['EN']]
Displaying 1 of 1 matches:
as , tanto públicas como privadas y charter . El sitio es
```

sorting into categories

- manually!
- alternative: automatic classification into categories
- supervised learning
- clustering

Resulting Categories

named entities

"Converse All-Star", "iPod Shuffle", "Daily Mail", "American Diabetes Association", ...

Resulting Categories

tech/internet

"Haz clic aquí" (Click here)

"se convirtió en un trending topic en las redes sociales"
(it became a trending topic in the social networks)

"por nuestros followers en las redes sociales"
(for our followers in the social networks)

"tiene capacidad de WiFi integrada"
(has an integrated wifi capacity)

"hace compras online" (do shopping online)

Resulting Categories

food/cooking

```
"panceta ( bacon )"
"la anchoa y la caballa ( mackerel )"
"Por ejemplo: Cranberry (arandanos)"
"con un albaricoque (apricot)"
"col rizada (kale, en ingles)"
"curcuma fresca (turmeric)"
```

Resulting Categories

food/cooking: life style

"prefieres alimentos gluten-free,"
(prefer gluten-free foods)

"pavo wild con salvia y ajo"
(wild turkey with sage and garlic)

"el saludable smoothie de fresas"
(the healthy strawberry smoothie)

"mirarán de forma extraña en tu coffeehouse favorito"
(will look funny in your favourite coffeehouse)

Resulting Categories

and more life style

"alimentos trendy" (trendy foods)

"cambia radicalmente de look" (radically change your look)

"los productos vintage originales"
(the original vintage products)

"con pequenos toques de glamour"
(with little glamour touches)

Resulting Categories

English idioms/collocations

"casarse y vivir su happily ever after"
(get married and live their happily ever after)

"son verdaderas decision makers"
(are the real decision makers)

"echale un vistazo al behind the scenes"
(take a look behind the scenes)

"Mix and match , la pareja ideal"
(Mix and match, the perfect partner)

Takeaways

- socio-linguistic questions are fascinating
- coding skills go a long way
- non-trivial problems
- “lifestyle”/“women”-magazines are evil

Sources

- code: `https://github.com/lusy/hora-de-decir-bye-bye`
- paper: `https://github.com/lusy/hausarbeiten/tree/master/spanishNYC/hausarbeit`
- Eckert, Penelope: Variation and the indexical field. In: Journal of Sociolinguistics 12 (2008), Nr. 4, S. 453–476
- Thomason, Sarah G.: Contact as a Source of Language Change. In: Joseph, Brian D. (Hrsg.) ; Janda, Richard D. (Hrsg.): The Handbook of Historical Linguistics. Oxford, UK : Blackwell Publishing Ltd, 2003, S. 687–712
- Zentella, Ana C.: Lexical Leveling in Four New York City Spanish Dialects: Linguistic and Social Factors. In: Hispania 73 (1990), Nr. 4

Resources

- nltk book: `https://www.nltk.org/book/`
- my favourite python tutorial: `https://learnpythonthehardway.org/python3/`

Thank you!

slides: `https://github.com/lusy/hausarbeiten/tree/master/spanishNYC/tacos28`

Questions

How would you solve this?

alternatives

- generate more complete word lists?
- `langid.py`?
- `nltk.corpus.words.words()` for English?

Resulting Categories

sex

"elevar tu nivel de sex-appeal"
(boost your sex-appeal level)

"tiene sex appeal y te llama la atencion"
(has sex appeal and attracts your attention)

"no trata de ser sexy" (don't try to be sexy)

"para tener una noche súper hot"
(in order to have a super hot night)