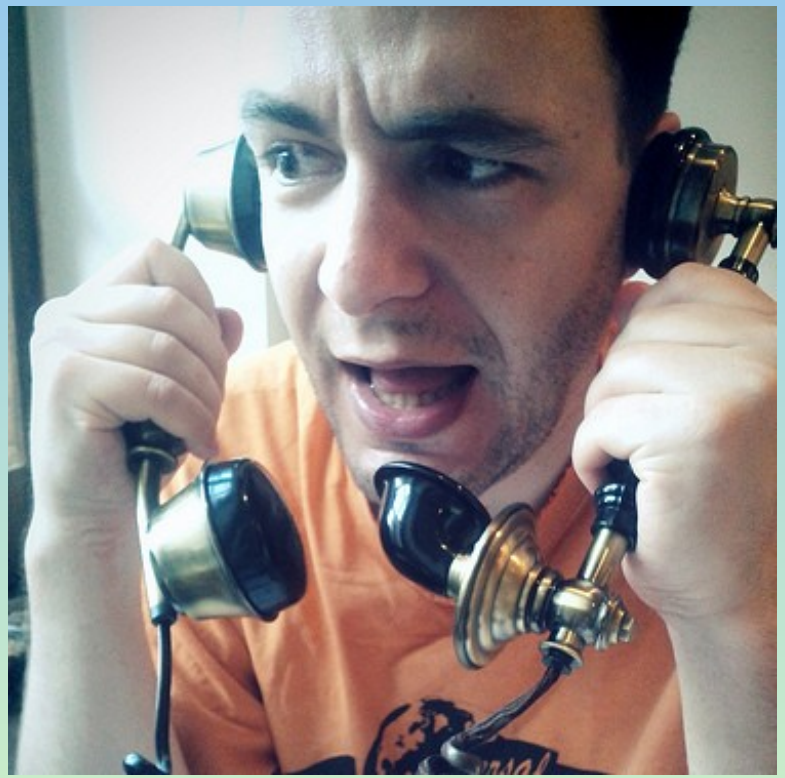
Organizando chistes con NLP

Julio Martínez

Who am I?

- Julio Martinez
- @liopic
- M.A.S. in Natural Language Processing
- Developing Webs since 2001
- Recently interested in ML
- Currently at Ulabox (3 years)



Why jokes, in Spanish?

- Real Problem:
  - Classify 17K products in categories
  - Complex categories' hierarchy
  - New products (batch) inserted by interns

  - PRONE TO ERRORS!

Torrent

Pollo Entero eviscerado

Básicos

1,5 Kg. / Precio 3,10 € / kg.

Why jokes, in Spanish?

- Solution:
  - Use NLP techniques
  - Find wrongly classified products
  - Based on product descriptions in Spanish

- As I cannot open private data…
- ...we will be doing the same with jokes
- Plus: in Spanish!

NLP?

- Natural Language Processing
  - Interaction between computers and humans
  - Understand human language
- Human language is complex
  - Context, backreferences, vagueness
  - Meta-language (irony, jokes…)
- Research in English

NLP step by step

- Classic techniques
  - From chars to a meaning
  - Tools we will use
- Modern approach
  - Word embeddings

NLP step by step

- Language detection
  - Standard 3-grams frequencies
    - the, and, ing, ent...
    - del, que, ent, ion…
    - per, ent, tat, ret…
- Word tokenization
  - Text, voice, other languages

NLP step by step

- Word cleaning
  - Stemmer, lemma
  - Removing stopwords
  - Punctuation, numbers, acronyms, tildes
- Named Entity recognition
  - “the lunch time”, “the president of France”

NLP step by step

- Part of Speech
  - Noun, Verb, etc
- Syntactic trees
  - Alternative trees, anaphora
  - “Le di un regalo a Juan; al abrirlo se manchó”
- Word semantics: Wordnet
- Semantic trees

NLP tools we will use

- The problem:
  - Given a collection of documents, create a numerical representation of each document
- The classic solution:
  - Bag of words
  - td-idf

Word embeddings

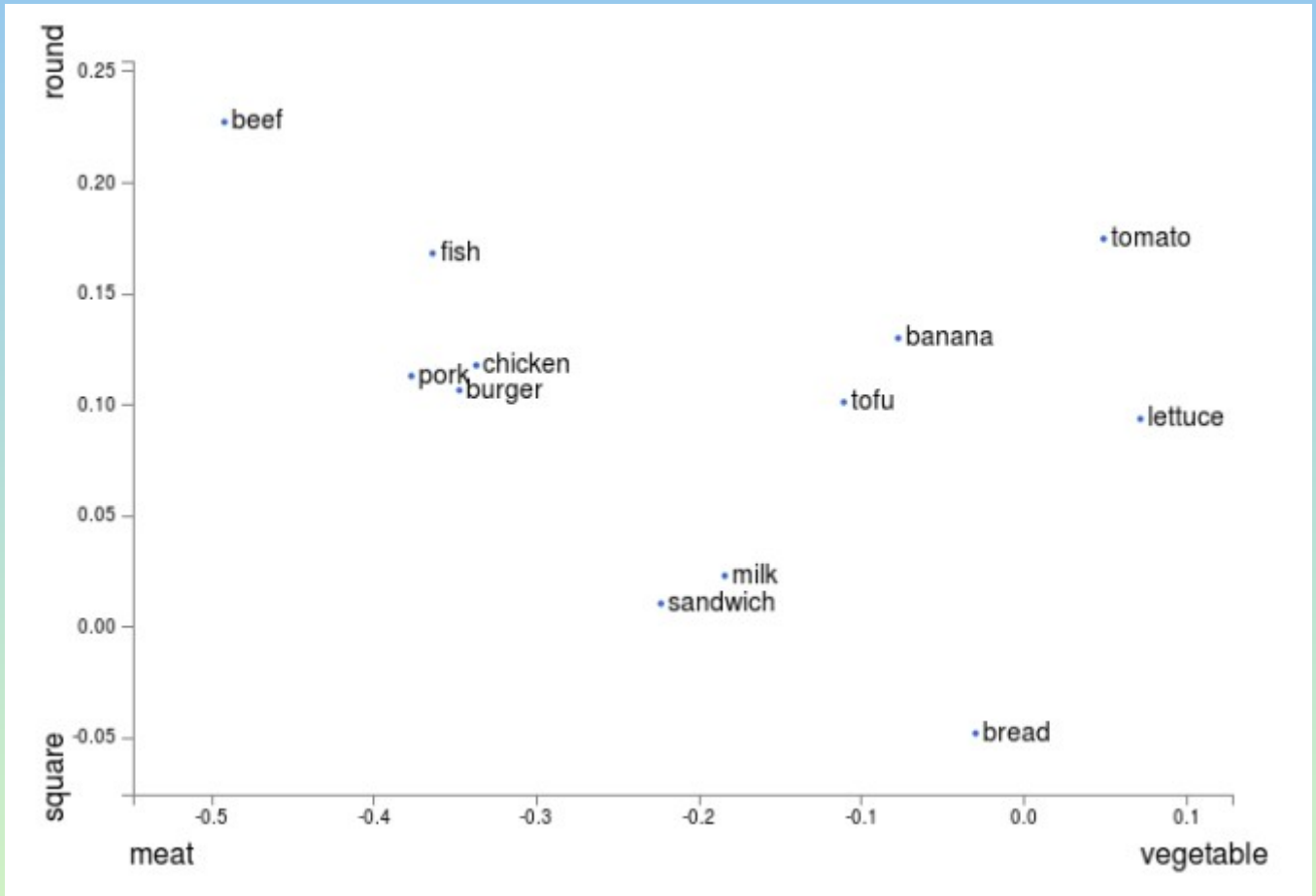
- Words dimensions trained by NN (2008)
- Google’s word2vec toolkit (2013)
- Solves the need of human labeled text
- It’s magic:

  king - man + woman ~= queen

Word embeddings

word	masculinity	royalty	age	… 300 features
king	0.999	0.999	0.60	
queen	0.001	0.998	0.50	
boy	0.01	0.001	0.05	

Word embeddings



Time to execute some code!

github.com/liopic/chistes-nlp