Project 1: Lexical Normalisation of Twitter Data

COMP90049
Knowledge Technologies

Jeremy Nicholson and Tim Baldwin and Sarah Erfani

Semester 2, 2017

THE UNIVERSITY OF MELBOURNE

Lexical Normalisation

Same idea can be expressed in different ways:

- `United States of America`
- `United States`
- `USA`
- `U.S.A`
- `US`
- `U.S.`
- `...`

"The United States of America" ← **canonical** form

Natural Language Processing

Natural Language Processing:

- Computer attempts to derive useful information from textual data
- For example:
  - Understanding language
  - Generating language
  - Translating between languages
  - Having a conversation with a computer (c.f. Turing Test)
  - ...

Natural Language Processing

Natural Language Processing:

- Computer attempts to derive useful information from textual data
- For example:
  - Understanding language
    - Identifying the concepts used in the data
    - Identifying the relationships between entities
    - ...

Natural Language Processing

Natural Language Processing:

- Computer attempts to derive useful information from textual data
- For example:
  - Understanding language
    - Identifying the concepts used in the data:
      - tokenisation
      - **canonicalisation**
      - part–of–speech tagging
      - named entity recognition
      - ...

Natural Language Processing

Natural Language Processing:

Way, way beyond the scope of this subject!

Assumptions

Some typical assumptions in NLP:

- Edited text
- Static data
- Long(ish) documents; plenty of context
- All context is language context
- Well-defined domain/genre
- Sentence boundaries are known
- Grammatical sentences
- Everything is English

What is Social Media?

`https://en.wikipedia.org/wiki/Social_media`, 6 August 2017:

> *Social media are computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks.*

What is Social Media?

For example:

- Social Networking (e.g. Facebook, Google+, ...)
- Content sharing (e.g. Instagram, YouTube, Flickr, ...)
- Blogs (e.g. Gizmodo, Mashable, ...)
- Micro-blogs (e.g. Twitter, Weibo, Tumblr, ...)
- User forums (e.g. StackOverflow, CNet, ...)
- Wikis (e.g. Wikipedia, Wiktionary, ...) ...

What is Social Media?

Common elements:

- Posts
- Social networks
- Comments
- Likes
- Aggregation
- Volume
- ...

Natural Language Processing on Social Media

Some typical assumptions in NLP:

- Edited text
- Static data
- Long(ish) documents; plenty of context
- All context is language context
- Well-defined domain/genre
- Sentence boundaries are known
- Grammatical sentences
- Everything is English

Natural Language Processing on Social Media

Some typical assumptions in NLP:

- ~~Edited~~ Unedited text
- ~~Static~~ Dynamic, streamed data
- ~~Long(ish)~~ Short documents; ~~plenty of~~ little context
- ~~All~~ Little context is language context, potential other kinds of context
- ~~Well-defined~~ Unclear domain/genre
- ~~Sentence boundaries are known~~ Idea of “sentence” is unclear; little punctuation
- ~~Grammatical sentences~~ I can haz grammar?
- Everything is ~~English~~ incredibly diverse

Tweets

omg boring #yawn

Tweets

Typical Twitter message:

(1) *c* *u* *2morw* *!!!*
see you tomorrow PUNCT
"See you tomorrow!"

Tweets

Typical Twitter message:

(2)

W00t	*!*	*i*	*got*	*RTs*	*from*	*7*	*tweeps*	*2day*
???	PUNCT	I	got	???	from	NUM	???	today

"Yay! Seven Twitter users re-posted my messages today."

Tweets

Typical Twitter message:

(3) *zomg* *gal* *bqhatevwr* *smh*
??? ?woman? ??? ?Sydney Morning Herald?
???

Project 1

Your job:

(1)

c	*u*	*2morw*	*!!!*	← **input**
`see`	`you`	`tomorrow`	`!!!`	← **output**

Project 1

Your job:

2morw $\rightarrow$ `tomorrow`

Maybe spelling correction will work?

Project 1

Your job:

More on Friday!

`kthxbye!`