Your text analysis pipeline and you

...

Kasper Welbers
Vrije Universiteit van Amsterdam

Who am I

- Kasper Welbers
- VU University Amsterdam
    - Postdoc on Responsible Terrorism Coverage (ResTeCo) project
    - Teach data science courses, in particular using R
- Substantive focus on Journalism
    - Gatekeeping theory
- Methodological focus on computational methods
    - Mostly automatic text analysis
    - Developer / maintainer of several R packages
        - corpustools
        - Rsyntax
        - RNewsflow
        - Tokenbrowser

Opportunities and pitfalls of specializing in computational research

- Opportunities. Computational skills in communication research...
  - lets you address new problems and old problems in new ways
  - opens up many avenues for collaboration
  - makes you part of an exciting and promising new field

Opportunities and pitfalls of specializing in computational research

- Opportunities. Computational skills in communication research...
  - lets you address new problems and old problems in new ways
  - opens up many avenues for collaboration
  - makes you part of an exciting and promising new field
- Pitfalls. Programming costs valuable time that you need to build your career
  - Programming takes time. Publishing and maintaining software even more so
  - Being ‘the method guy/girl’ probably won’t get you tenured
  - Lack of institutional level incentives in communication science

Opportunities and pitfalls of specializing in computational research

- Opportunities. Computational skills in communication research...
  - lets you address new problems and old problems in new ways
  - opens up many avenues for collaboration
  - makes you part of an exciting and promising new field
- Pitfalls. Programming costs valuable time that you need to build your career
  - Programming takes time. Publishing and maintaining software even more so
  - Being 'the method guy/girl' probably won't get you tenured
  - Lack of institutional level incentives in communication science

- Learn while you can (after PhD time goes downhill)
- Look strategically at what existing software you can and need to use, and what you can and need to develop yourself

Computational text analysis

- What is text analysis?
    - The analysis of natural language data
- Why is text analysis important?
    - Text is still one of the most prominent forms of communication
    - Majority of online data consists of unstructured, natural language data

Computational text analysis

- What is text analysis?
  - The analysis of natural language data
- Why is text analysis important?
  - Text is still one of the most prominent forms of communication
  - Majority of online data consists of unstructured, natural language data

- Opportunity and pitfall
  - There are many amazing state-of-the-art techniques for processing natural language
  - We can and should use these techniques to advance our research

Computational text analysis

- What is text analysis?
  - The analysis of natural language data
- Why is text analysis important?
  - Text is still one of the most prominent forms of communication
  - Majority of online data consists of unstructured, natural language data

- Opportunity and pitfall
  - There are many amazing state-of-the-art techniques for processing natural language
  - We can and should use these techniques to advance our research
  - Don’t reinvent the pipeline
    - Read up on some computational linguistics and computer science
    - Be careful with ambitions to ‘improve’ natural language processing

Four general components of a text analysis project

Obtaining text

- Existing archives
- APIs
- Scrapers
- Cleaning

Text to data

- Preprocessing
- Feature preparation

Analysis

- Dictionary / rule-based
- Supervised approaches
- Unsupervised approaches

Validation

Verify (and prove) that you measure what you claim to measure

Four general components of a text analysis project

Obtaining text	Text to data	Analysis	Validation
Existing archives APIs Scrapers Cleaning	Preprocessing Feature preparation	Dictionary / rule-based Supervised approaches Unsupervised approaches	Verify (and prove) that you measure what you claim to measure

Basic preprocessing

Simple string operations to transform texts to data

- Fast, and easy to install
- Limited features
- Can be inaccurate, especially for non-English languages

Basic preprocessing

Simple string operations to transform texts to data

- Fast, and easy to install
- Limited features
- Can be inaccurate, especially for non-English languages

Advanced preprocessing

Advanced models, based on annotated text corpora

- Slower, and bit more of a hassle. Not available for every language
- More accurate and enables new forms of analysis

Basic preprocessing

- Tokenization
- Lowercasing
- Stopword removal
- Stemming

Basic preprocessing

- Tokenization
  - Splitting texts into individual words
- Lowercasing
  - Make all text lowercase
- Stemming
  - Reduce word to stem
- Remove stopwords
  - Remove list of common words

"All humans are mortal"

"All" "humans" "are" "mortal"

"All" → "all"

"humans" → "human"

"human" "mortal"

Basic preprocessing

- Tokenization
- Lowercasing
- Stopword removal
- Stemming

Advanced preprocessing

- Tokenization
- Part-of-speech tagging
- Lemmatization
- But also...
    - Dependency parsing
    - Named entity recognition
    - Coreference resolution
    - Geotagging
    - Word embeddings

Stemming

- Simple rule-based approach
- Cuts off suffix

token	stem
The	the
walking	walk
dead	dead
were	were
walking	walk

Lemmatization

- Dictionary approach
- Uses part-of-speech (POS) tags

token	POS	lemma
The	DET	the
walking	NOUN	walking
dead	ADJ	dead
were	VERB	be
walking	VERB	walk

Named entity recognition

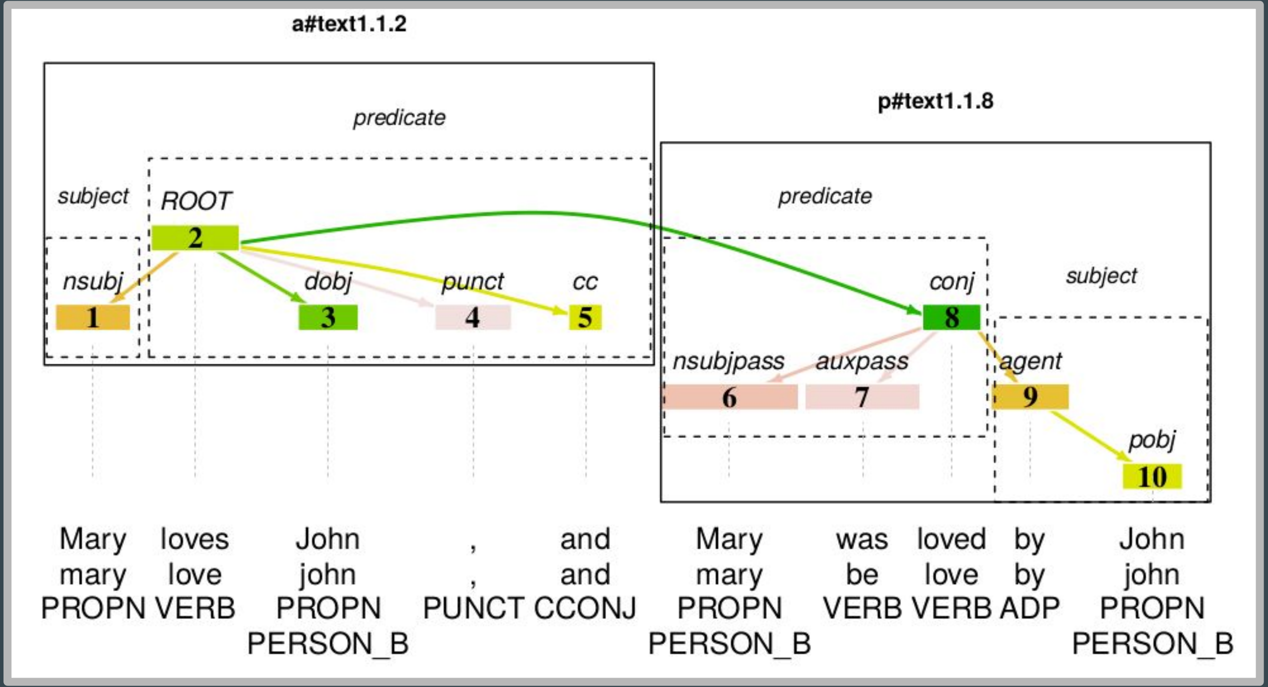
Locate and classify named entities (persons, locations, events)

token	POS	lemma	entity
Bob	PROPN	bob	PERSON_B
Smith	PROPN	smith	PERSON_I
worked	VERB	work	
for	ADP	for	
Apple	PROPN	apple	ORG_B

Dependency parsing

- Represent syntactical structure as a dependency graph

Example:
rsyntax package



Where to start

- What language do you work in?
  - Python
  - R
  - Other?

- Python
  - Spacy
  - StanfordNLP
- R
  - UDpipe (c++ wrapper, easy to install)
  - Spacyr
- Other
  - CoreNLP runs in Java
  - UDpipe runs in c++
  - Many parsers run as a server, and there are often (Docker) containers available.

Language support

- Most parsers focus on English
- Several parsers now support any language that has training data
  - Spacy (pretrained models for 8 languages)
  - UDpipe (pretrained models for 60+ languages)
  - CoreNLP (six languages)
  - StanfordNLP (53 languages)
- Not all techniques are supported for all languages, and accuracy can be much lower for languages with less support
  - POS tagging and lemmatization are often ok
  - Dependency parsing is often problematic (if it exists at al)
  - Hard tasks such as coreference resolution almost only supported for English

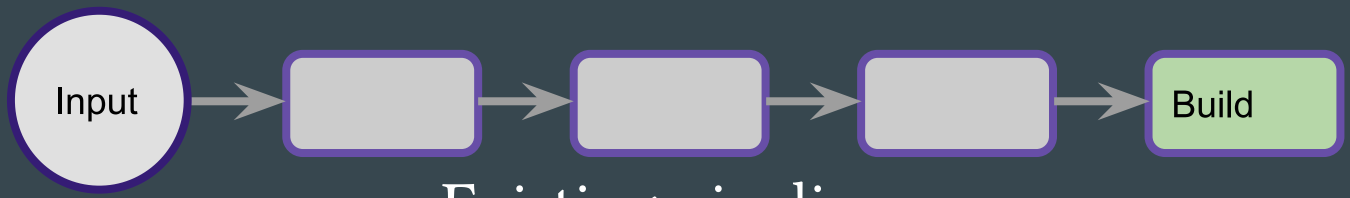
Plugging in to the pipeline

If you write code yourself...

- Make proper software packages (even if you do not aim to share the work)
  - R packages
  - Python modules
  - Proper documentation
  - Unit tests
  - Use Github (or similar)
- If worth sharing, publish on popular repositories
  - CRAN for R
  - PyPI for Python

Do not reinvent the wheel

- Natural Language Processing is a field in itself
  - If your focus is Communication Science, do not compete with NLP research teams
  - Know enough so that you can use the state-of-the-art
- Keep the focus on your own research
  - What existing techniques and software can help you perform your research?
  - What do you need to add or develop to perform your research?
  - Can you publish it in a way that gives you credit for it?



Existing pipeline

your contribution

Your text analysis pipeline and you

...

Kasper Welbers
Vrije Universiteit van Amsterdam

What about word embeddings?

One-hot vector

	Dogs	Cats	Paris	France
Dogs	1	0	0	0
Cats	0	1	0	0
Paris	0	0	1	0
France	0	0	0	1

Word Embeddings
(3 dimensional)

	Dim 1	Dim 2	Dim 3
Dogs	0.3	0.2	0.01
Cats	0.2	0.1	0.03
Paris	0.01	0.5	0.6
France	0.04	0.4	0.8

What about word embeddings?

- Can be considered a form of dimensionality reduction
- Used in many of the advanced (machine learning based) preprocessing techniques

- What can I use if for?
  - Supervised machine learning
    - Especially neural net
  - Vector space models
    - Document similarity
    - Term similarity

Pre-trained word representations

Target Word	BoW5	BoW2	Deps
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina

From: Goldberg, 2017