University of BRISTOL

Week 4: Text Analytics

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Questions and Answers

- Please post questions about:
  - Lecture videos
  - Labs
  - This session…
- Post to:
  - Teams QA channel
  - Blackboard discussion forum (anonymous)

University of BRISTOL

Introduction

Why does text require specialised methods?

- Please post answers to Padlet: https://uob.padlet.org/edwinsimpson/khs5rhkbpco7kdtk

What Makes Text Special?

- Text data is **discrete.** Discrete units are combined in sequences to form meaning.
- Many observations are **rare**, many possible sentences are never observed in any given dataset.
- Text is **compositional:** words combine into phrases, which combine to form sentences, and so on.
- Ambiguity, errors and variations in the way people use language also present major challenges.

Ways to View Meaning in Language

Umashanthi interviewed Ana. She works for the college newspaper.

- Relational
  - Relationships between words represent meaning
  - E.g., synonyms, categories, …
- Compositional
  - The meaning of larger units is formed by combining smaller units
  - E.g., sentences from phrases, words from suffixes, prefixes and stems
- Contextual (or ‘distributional’)
  - We can understand a word from its context
  - The context of a word alters its meaning

Ethical Considerations

- Give some examples of ethical considerations when building a text analytics system.

- Please post answers on Padlet: https://uob.padlet.org/edwinsimpson/khs5rhkbpco7kdtk

Ethical Considerations

- Privacy and freedom of speech: whose data are we processing, and does doing so restrict their privacy or freedom?
- Labour: who created the data that we’re using and do they benefit from our technology?
- Bias: technology can amplify societal biases, so have we done enough to identify and address potential problems? Do we serve different communities (e.g., speakers of different languages) equally?

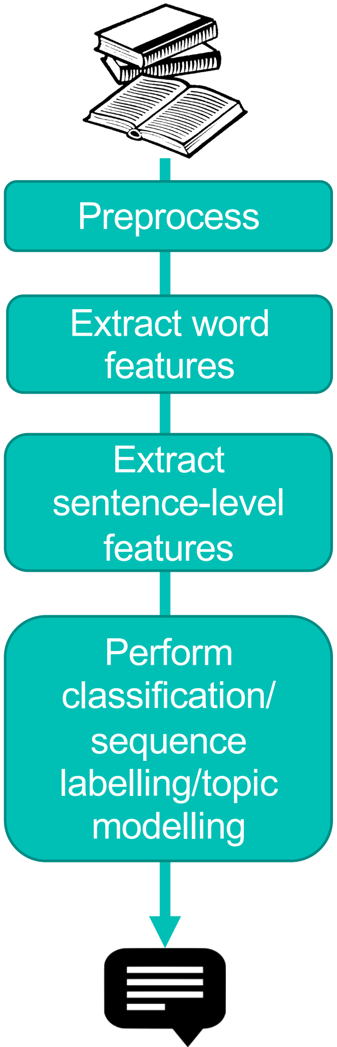
University of
BRISTOL

Regular Expressions

Regular Expressions

- How can we use regular expressions in Text Analytics?
- Where in the pipeline does it fit?

- Write answers on the Padlet: https://uob.padlet.org/edwinsimpson/khs5rhkbpco7kdtk



Preprocess

Extract word features

Extract sentence-level features

Perform classification/ sequence labelling/topic modelling

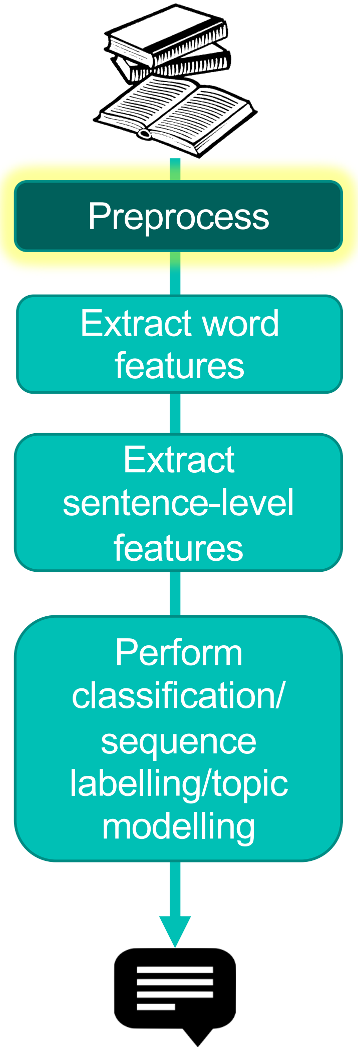
University of BRISTOL

Text Normalisation

bristol.ac.uk

Text Normalisation

- Most text processing methods require us to normalise the text first as a preprocessing step.
- Typical steps are…?



Preprocess

Extract word features

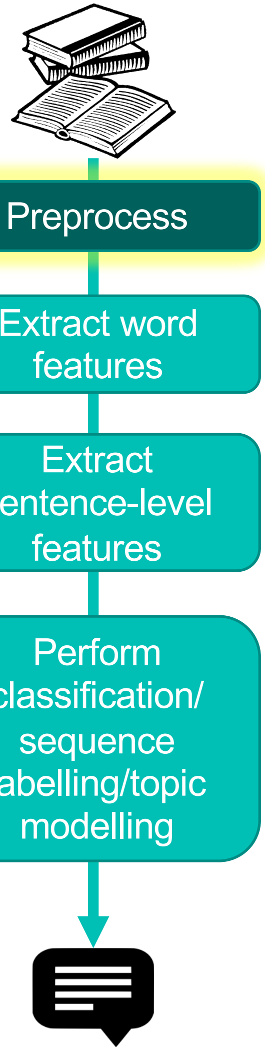
Extract sentence-level features

Perform classification/ sequence labelling/topic modelling

Text Normalisation

- Most text processing methods require us to normalise the text first as a preprocessing step.
- Typical steps are:

1. Tokenisation;
2. Word normalisation;
3. Sentence segmentation.

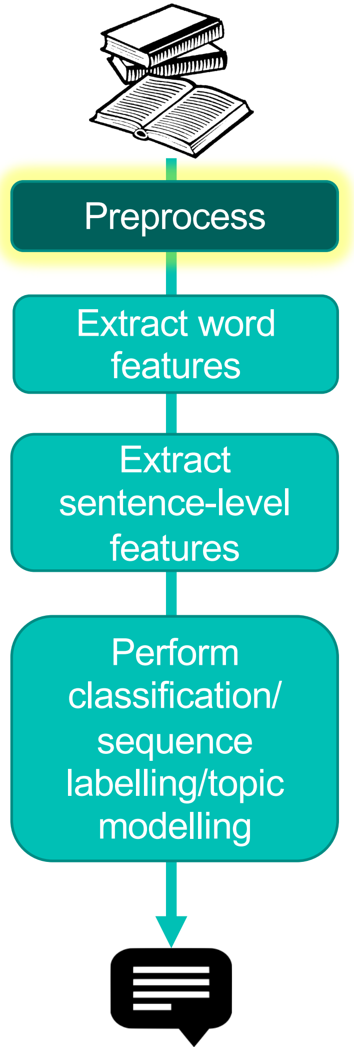

Preprocess
Extract word features
Extract sentence-level features
Perform classification/ sequence labelling/topic modelling

Text Normalisation

- Most text processing methods require us to normalise the text first as a preprocessing step.
- Typical steps are:

1. Tokenisation;
2. Word normalisation;
3. Sentence segmentation.

- Why do we need them?



Preprocess

Extract word features

Extract sentence-level features

Perform classification/ sequence labelling/topic modelling

Tokenisation

- Split on whitespace, punctuation
- Use dictionaries to identify multi-word phrases or to split words
- How does tokenisation differ between languages?

- https://uob.padlet.org/edwinsimpson/khs5rhkbpco7kdtk

Normalising Word Formats

Step	Example input	Output
Replacing emojis with text	🔥 🎅🏾 🇲🇽	*Fire* *Santa Claus: medium-dark skin tone* *flag: Mexico*
Normalising URLs, hashtags	*http://www.bristol.ac.uk* *#NLProc*	*URL* *HASHTAG*
Stopword removal	*the*	
Case folding	*The* *THE*	*the* *the*
Lemmatisation/ Stemming	*is* *reading*	*be* *read*

Lemmatization

- Words have internal structure:
  - They are composed of stems and affixes
  - “Cats” contains the stem “cat” and an affix “s”
- Replace each word with its root form or **lemma**
- Implement by applying a series of regular expression substitutions
- Example implementation: WordNet Lemmatizer
- Porter stemmer provides a quicker but more error-prone alternative.

Blackboard Quiz

- Post your answers here anonymously:
- https://uob.padlet.org/edwinsimpson/p51jjczj4baw1fl

Lab 4: Regexp and Text Normalisation

bristol.ac.uk

Switch to Jupyter notebook…