estimand

Natural language processing

Dr Gianluca Campanella

14th June 2016

Contents

Natural language processing

Natural language processing

NLP is the task of extracting **meaning** and **information** from text documents, for example:

- Text categorisation
- Sentiment analysis
- Machine translation

Natural language processing

NLP is the task of extracting **meaning** and **information** from text documents, for example:

- Text categorisation
- Sentiment analysis
- Machine translation

- Text is often unstructured
- **Pre-processing** is required before we can use the algorithms we learned

Tokenisation

Separating a sentence into its constituent parts (tokens)

Tokenisation

Separating a sentence into its constituent parts (tokens)

Example

- Data science is the future!

$\rightarrow$ { Data, science, is, the, future, ! }

Stemming and lemmatisation

Identifying **roots** of words

- **Stemming** removes common endings such as ‘—ing’
- **Lemmatisation** uses language-specific knowledge

Stemming and lemmatisation

Identifying **roots** of words

- **Stemming** removes common endings such as '—ing'
- **Lemmatisation** uses language-specific knowledge

Examples

- Badly $\rightarrow$ bad
- Best $\rightarrow$ good

Tagging and parsing

Identifying **parts of speech** and **named entities**

Tagging and parsing

Identifying **parts of speech** and **named entities**

Examples

- What are the nouns, adjectives, verbs, ...?
- Which pieces go together?
- Which tokens correspond to proper nouns (people, business names, locations, ...)?

Common problems

All these tasks are difficult because...

- Language is complex and sometimes inconsistent
- Usage changes frequently

Common problems

All these tasks are difficult because...

- Language is complex and sometimes inconsistent
- Usage changes frequently

Approaches

1. Rule-based systems (grammar)
2. Usage of words (inferred from training corpora)

Bag-of-words classification

Text classification

Examples

- Is this article about science or sports?
- Is this comment positive or negative?

Text classification

Examples

- Is this article about science or sports?
- Is this comment positive or negative?

- Each word becomes a predictor
  (e.g. whether it's present in each document or not)

$\rightarrow$ 'Bag of words'

Bag-of-words in `scikit-learn`

`CountVectorizer` converts documents to a matrix where...

- Each row is a sample (document)
- Each column is a count or indicator for words or contiguous sequences of words (*n*-grams)

Bag-of-words in `scikit-learn`

`CountVectorizer` converts documents to a matrix where...

- Each row is a sample (document)
- Each column is a count or indicator for words or contiguous sequences of words (*n*-grams)

- 'Rare' words may cause overfitting

$\rightarrow$ Filter or use regularisation

tf–idf representation

Term frequency–inverse document frequency (**tf–idf**) reflects how important a word is to a document

tf

Number of times a given word occurs in a document

idf

Inverse proportion of documents with the word

tf–idf representation

Term frequency–inverse document frequency (**tf–idf**) reflects how important a word is to a document

tf

Number of times a given word occurs in a document

idf

Inverse proportion of documents with the word

High tf–idf words...

- Appear frequently in a given document
- Appear rarely in other documents

Latent variable models

Latent variable models

Traditional NLP

- Language in theory
- Preprogrammed set of rules (grammar)
- ‘Bad’ and ‘badly’ are related because they share a common root

Latent variable models

- Language in practice
- Unsupervised learning of structure
- ‘Bad’ and ‘badly’ are related because they are used similarly or near similar words

Latent variable models

Assumption

There is some hidden (**latent**) structure to the data that we'd like to learn

$\rightarrow$ Ignore grammar

$\rightarrow$ Learn rules directly from the data

Redundancy of bags of words

Problem

- Bags of words are a **redundant** representation
- Many words are likely to represent the same concept

$\rightarrow$ Many columns are repetitive

Redundancy of bags of words

Problem

- Bags of words are a **redundant** representation
- Many words are likely to represent the same concept

$\rightarrow$ Many columns are repetitive

Solutions

- Regularisation
- **Dimensionality reduction**
- **Mixture models**

Dimensionality reduction

- Identify correlated columns
- Replace them with a new column that ‘encapsulates’ the others

Dimensionality reduction

- Identify correlated columns
- Replace them with a new column that ‘encapsulates’ the others

Example

- { car, cat, truck, van }

$\rightarrow$ { cat, vehicle }

Latent Dirichlet allocation

- Identify correlated columns
- Create clusters of common words
- Generate probability distributions for relatedness

- Each word belongs to a (latent) ‘topic’
- Each document is a mixture of topics

Latent Dirichlet allocation

LDA tries to learn...

- The word distribution of each topic

$\rightarrow$ Pr(word | topic)

- The topic distribution of each document

$\rightarrow$ Pr(topic | document)

Model evaluation is mostly about interpretation:

- Do the topics make sense?
- Do the constituent words of each topic make sense?

word2vec

- Based on neural networks
- Focus is on words, not documents
- **Idea**: define a word by listing all the ways it's used

word2vec

- Based on neural networks
- Focus is on words, not documents
- **Idea**: define a word by listing all the ways it's used

Example

+ ... is the capital of
+ ..., UK
+ The restaurant in ...
− Can I have a ...
− There's too much ... on this