Foundations of Natural Language Processing
Lecture 3c
Text Corpora: Tokenisation

Alex Lascarides

School of informatics

Corpora in NLP: Tokenisation

Last Time

- Using a corpus to inform NLP
  - Defining input and output
  - Training vs. testing (keep data separate!)
  - Define evaluation metric
- How to access corpora in NLTK.

Now: Preparing a corpus for NLP: tokenisation

Tokenisation

Let’s take another look at the movie_reviews corpus:

```
>>> print('\n'.join(' '.join(sent) for sent in
  ↪ movie_reviews.sents()[:5]))
plot : two teen couples go to a church party , drink
  ↪ and then drive .
they get into an accident .
one of the guys dies , but his girlfriend continues to
  ↪ see him in her life , and has nightmares .
what ' s the deal ?
watch the movie and " sorta " find out .
```

What do you notice about spelling conventions? Spacing?

Tokenisation

Normal written conventions sometimes do not reflect the “logical” organisation of textual symbols. For example, some punctuation marks are written adjacent to the previous or following word, even though they are not part of it. (The details vary according to language and style guide!)

Given a string of raw text, a **tokeniser** adds logical boundaries between separate word/punctuation **tokens** (occurrences) not already separated by spaces:

Daniels made several appearances as C-3PO on numerous TV shows and commercials, notably on a Star Wars-themed episode of The Donny and Marie Show in 1977, Disneyland’s 35th Anniversary.

⇒

Daniels made several appearances as C-3PO on numerous TV shows and commercials , notably on a Star Wars - themed episode of The Donny and Marie Show in 1977 , Disneyland ’s 35th Anniversary .

To a large extent, this can be automated by rules. But there are always difficult cases.

Tokenisation in NLTK

```
>>> nltk.word_tokenise("Daniels made several
  ↪ appearances as C-3PO on numerous TV shows and
  ↪ commercials, notably on a Star Wars-themed episode
  ↪ of The Donny and Marie Show in 1977, Disneyland's
  ↪ 35th Anniversary.")
['Daniels', 'made', 'several', 'appearances', 'as',
  ↪ 'C-3PO', 'on', 'numerous', 'TV', 'shows', 'and',
  ↪ 'commercials', ',', 'notably', 'on', 'a', 'Star',
  ↪ 'Wars-themed', 'episode', 'of', 'The', 'Donny',
  ↪ 'and', 'Marie', 'Show', 'in', '1977', ',',
  ↪ 'Disneyland', "'s", '35th', 'Anniversary', '.']
```

Tokenisation in NLTK

```
>>> nltk.word_tokenise("Daniels made several
  ↪ appearances as C-3PO on numerous TV shows and
  ↪ commercials, notably on a Star Wars-themed episode
  ↪ of The Donny and Marie Show in 1977, Disneyland's
  ↪ 35th Anniversary.")
['Daniels', 'made', 'several', 'appearances', 'as',
  ↪ 'C-3PO', 'on', 'numerous', 'TV', 'shows', 'and',
  ↪ 'commercials', ',', 'notably', 'on', 'a', 'Star',
  ↪ 'Wars-themed', 'episode', 'of', 'The', 'Donny',
  ↪ 'and', 'Marie', 'Show', 'in', '1977', ',',
  ↪ 'Disneyland', "'s", '35th', 'Anniversary', '.']
```

English tokenisation conventions vary somewhat—e.g., with respect to:

- **clitics** (contracted forms) *’s*, *n’t*, *’re*, etc.

- hyphens in compounds like *president-elect* (fun fact: this convention changed between versions of the Penn Treebank!)

Preprocessing/normalisation: The tip of the iceberg

(Word-level) tokenisation is just part of the larger process of preprocessing or normalisation, which may include

- encoding conversion
- removal of markup
- *insertion* of markup
- case conversion
- sentence boundary detection

NLTK provides `nltk.sent_tokenize()` for **sentence tokenisation**, but it is far from perfect (and indeed the fact of the matter is not always clear).

Preprocessing/normalisation: an example

Consider the following Wikipedia extract (from https://en.wikipedia.org/wiki/The_U.S._Air_Force_%28song%29)

> In April 1938, Bernarr A. Macfadden, publisher of *Liberty* magazine stepped in, offering a prize of $1,000 to the winning composer, stipulating that the song must be of simple “harmonic structure”, “within the limits of [an] untrained voice”, and its beat in “march tempo of military pattern”.
>
> The contest rules required the winner to submit his entry in written form, and Crawford immediately complied. However his original title, *What Do You think of the Air Corps Now?*, was soon officially changed to *The Army Air Corps*.

The actual marked-up original for the latter part of the second paragraph above is actually the following (without the line breaks):

```
However his original title, <i>What Do You think of the
 Air Corps Now?</i>, was soon officially changed
 to <i>The Army Air Corps</i>.
```

Preprocessing/normalisation: an example, cont'd

It should be evident that a large number of decisions have to be made, many of them dependent on the eventual intended use of the output, before a satisfactory preprocessor for such data can be produced.

Documenting those decisions and their implementation is then a key step in establishing the credibility of any subsequent experiments.

Such documentation is especially important if some preprocessing has been done on a corpus before it is distributed publically. You may have noted, for example, that the movie review corpus we looked at earlier has already had case conversion (in this case, lower-casing) performed, as well as some separation of punctuation...

Choice of training and evaluation data

We know that the way people use language varies considerably depending on **context**. Factors include:

- *Mode of communication:* speech (in person, telephone), writing (print, SMS, web)
- *Topic:* chitchat, politics, sports, physics, . . .
- *Genre:* news story, novel, Wikipedia article, persuasive essay, political address, tweet, . . .
- *Audience:* formality, politeness, complexity (think: child-directed speech), . . .

In NLP, **domain** is a cover term for all these factors.

Choice of training evaluation data

- Statistical approaches typically assume that the training data and the test data are sampled from the same distribution.
  - I.e., if you saw an example data point, it would be hard to guess whether it was from the training or test data.

- Things can go awry if the test data is appreciably different: e.g.,
  - different tokenisation conventions
  - new vocabulary
  - longer sentences
  - more colloquial/less edited style
  - different distribution of labels

- **Domain adaptation** techniques attempt to correct for this assumption when something about the source/characteristics of the test data is known to be different.

Summary: Why do we need text corpora?

Two main reasons:

1. To evaluate our systems
   - Good science requires controlled experimentation.
   - Good engineering requires benchmarks.

2. To help our systems work well (data-driven methods/machine learning)
   - When a system's behavior is based on manual rules or databases, it is said to be **rule-based**, **symbolic**, or **knowledge-driven** (early days of NLP)
   - **Learning**: collecting statistics or patterns automatically from corpora to govern the system's behavior (dominant in contemporary NLP)
     - **supervised learning**: the data provides example input/output pairs (main focus in this course)
     - core behavior: **training**; refining behavior: **tuning**
     - Either way, corpus preparation is (probably) necessary.