Language Analytics

Session 3: NLP for Linguistic Analysis

Ross Deans Kristensen-McLachlan
rdkm@cas.au.dk

Course outline

- 1. Introductions
- 2. String Processing with Python
- **3. NLP for linguistic analysis**
- 4. Text Classification 1
- 5. Text Classification 2
- 6. Word embeddings
- 7. Language modelling 1
- 8. Language modelling 2
- 9. BERT
- 10. More BERT
- 11. Project pitches
- 12. Generative models
- 13. Social impact

Plan for today

- Catch-up

- 1. What is natural language processing?

- 2. Key concepts in NLP
  - Tokenization
  - Part-of-speech tagging
  - Named entity recognition

- 3. Code-along session
  - An intro to spaCy
  - Starting on Assignment 1

What is natural language processing?

What is natural language processing?

- Put simply, the goal of NLP is…

> *[…] to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.*
>
> (Jurafsky & Martin 2021: 1)

- For our purposes in this course, NLP is a useful way of extracting structured linguistic information from raw text

What is natural language processing?

- NLP as a discipline has roots going back to (at least) the 1940's

- It exists at the intersection of a number of related fields:
  - Linguistics
    - Computational (psycho-) linguistics
  - Computer science
  - Artificial intelligence
    - Cognitive and computational psychology
  - Information retrieval

Tokenization

Tokenization

- In order to count how many times a word appears or to calculate association measures between two words, we need to know what words there are!

- Why is this a potential challenge for a computer?
  - Hint: think about data types

Tokenization

- Think about the following sentence:

  I don’t know how to tokenize this string of text!

- How might you choose to tokenize this?

One strategy

- Whitespace tokenization

  [*I, don't, know, how, to, tokenize, this, string, of, text!*]

- Can you see any problems here?

Tokenization – some problems

- One problem with whitespace tokenization
  - Assumes whitespace!

Tokenization – some problems

- One problem with whitespace tokenization
  - Assumes whitespace!

当前，世界面临食品、气候变化以及金融等多重危机，而正是这些危机更加突出了农业在发展中国家至关重要的地位。

(Eng: Currently, the world is facing multiple crises such as food, climate change, and finance, and it is these crises that have highlighted the vital role of agriculture in developing countries.)

Tokenization – some problems

- One problem with whitespace tokenization
  - Assumes whitespace!

当前，世界面临食品、气候变化以及金融等多重危机，而正是这些危机更加突出了农业在发展中国家至关重要的地位。

(Eng: Currently, the world is facing multiple crises such as food, climate change, and finance, and it is these crises that have highlighted the vital role of agriculture in developing countries.)

Tokenization – some problems

- One problem with whitespace tokenization
  - Assumes whitespace!

当前，世界面临食品、气候变化以及金融等多重危机，而正是这些危机更加突出了农业在发展中国家至关重要的地位。

(Eng: Currently, the world is facing multiple crises such as food, climate change, and finance, and it is these crises that have highlighted the vital role of agriculture in developing countries.)

"世", "界", "面", "临” ✗

"世界", "面临" ✓

Tokenization – some problems

- Desired result:

  当前，世界面临食品、气候变化以及金融等多重危机，而正是这些危机更加突出了农业在发展中国家至关重要的地位。

  "当前","，","世界","面临","食品","、","气候变化","以及","金融","等","多重","危机","，","而","正是","这些","危机","更加","突出","了","农业","在","发展中国家","至关重要","的","地位","。"

Tokenization

- So whitespace is a problem, insofar as it assumes whitespace, causes trouble with punctuation, etc

- Let's try again

  I don't know how to tokenize this string of text!

- What else might work?

A second strategy

- We could try to craft more specific rules to only tokenize on alphanumeric characters instead

  [*I, don, ’t, know, how, to, tokenize, this, string, of, text, !*]

- Can you see any other problems here? Is this better or worse than whitespace tokenization?

Some problems

- Another problem with alphanumeric clusters
  - What about multiword units?

Some problems

- Another problem with alphanumeric clusters
  - What about multiword units?

"We went for a walk where we go to walk. As we walked, all of a sudden, we saw New York!”

Some problems

- Another problem with alphanumeric clusters
  - What about multiword units?

"We went for a walk where we go to walk. As we walked, all of a sudden, we saw New York!”

- Is ‘New York’ *really* two separate tokens? What about an adverbial like ‘all of a sudden’ or the infinitive ‘to walk’?

Some problems

- Another problem with alphanumeric clusters
  - What about multiword units?

"We went for a walk where we go to walk. As we walked, all of a sudden, we saw New York!”

- Is ‘New York’ *really* two separate tokens? What about an adverbial like ‘all of a sudden’ or the infinitive ‘to walk’?

[we, went, for, a, walk, where, we, go, to, walk,
as, we, walked, all_of_a_sudden, we, saw, new_york]

Some problems

- Similarly, English has contractions like *don't* which cause problems – should it be *don't, do not, do + n't,…?*

- Other languages have similar features which are tricky for tokenization. Can you think of other examples?

  - *French:*
    - *Je mange <u>du</u> pain* du = de + le

  - *Italian*
    - *Seduto sulla sedia* sulla = su + la

  - Danish
    - *spillet 'the game'*
    - *Spillet 'played (past participle, e.g har spillet)'*

  - *Finnish*
    - *talo* *'house'*
    - *talo-n* *'of the house'*
    - *talo-ssa* *'in the house'*
    - *talo-i-ssa* *'in the houses'*

What is a word?

What is a word?

- There are *at least* three different answers to this:
  - Word form
    - Physical, concrete realisation of a word in text or speech
      - *Walk*, *walked*
      - *Go, went*

What is a word?

- There are *at least* three different answers to this:

  - Lexeme
    - Abstract cognitive representation
      - [WALK] <-> walk, walked, walking
      - [GO] <-> go, went, going

What is a word?

- There are *at least* three different answers to this:

  - Morphosyntactic word
    - Combines lexeme to specific property such as part-of-speech
      - *as we walked* -> [WALK + simple past]
      - *where we go to walk* - > [GO + habitual], [WALK + infinitive]

What is a word?

- There are *at least* three different answers to this:
  - Word form
    - Physical, concrete realisation of a word in text or speech
      - *Walk*, *walked*
      - *Go, went*
  - Lexeme
    - Abstract cognitive representation
      - [WALK] <-> walk, walked, walking
      - [GO] <-> go, went, going
  - Morphosyntactic word
    - Combines lexeme to specific property such as part-of-speech
      - *as we walked* -> [WALK + simple past]
      - *where we go to walk* - > [GO + habitual], [WALK + infinitive]

What is a word?

- A working definition of a word:

What is a word?

- A working definition of a word:

  A word is a unit of language which has a semantic 'nucleus' and a word class to which it belongs.

- Is this satisfactory? Can you think of any 'edge cases'?

Words or tokens?

- Given this definition, how do we treat non-words?

More problems

- Given this definition, how do we treat non-words?
- This is especially relevant for things like emojis and punctuation

See you tomorrow, I guess 😂
See you tomorrow, I guess 🙃
See you tomorrow, I guess 😕
See you tomorrow, I guess 😭

More problems

- Given this definition, how do we treat non-words?
- This is especially relevant for things like emojis and punctuation

See you tomorrow, I guess 😂
See you tomorrow, I guess 🙃
See you tomorrow, I guess 😕
See you tomorrow, I guess 😭

- (For more on emojis as language, see McCulloch (2019), *Because internet)*

More problems

- What about other linguistic features?
  - Uhuh, Hmmm, nå, nåååååå
  - …
  - …!
  - ???

- What about numbers?
  - British English: £500,500.50
  - Danish: £500.500,50

Summary

- All NLP tasks essentially involve counting words and modelling their distribution
- This depends on the way we choose to tokenize our text
- While it may seem intuitive as a human reader, tokenization is *not a trivial task*
- Tokenization is *language dependent*. Different languages have different requirements
- What counts as a token is *task dependent*
  - All words are tokens but not all tokens need be words

Break

Part of speech tagging

Part of speech tagging

- We said that a word has a semantic nucleus and belongs to a word class

- But what exactly is a word class?

Part of speech tagging

- We said that a word has a semantic nucleus and belongs to a word class

- But what exactly is a word class?

  - *Noun, verb, adjective, adverb, pronoun, preposition, conjunction, participle, article*

  - NB: traditional Danish grammar uses different names for word classes

Part of speech tagging

Yesterday, I was in the park and I saw a {___}

Part of speech tagging

Yesterday, I was in the park and I saw a {___}

dog

slide

tree

fight

Part of speech tagging

Yesterday, I was in the park and I saw a {___}

dog

slide

tree

fight

Yesterday, I was in the park and I saw a {NOUN}

Part of speech tagging

This is my favourite place to {____} in Rome

Part of speech tagging

This is my favourite place to {____} in Rome

eat

sleep

party

fight

This is my favourite place to {VERB} in Rome

Part of speech tagging

1. A word's class – its part of speech – tells us a lot about how we should expect it to behave
   - Remember: *You shall know a word by the company it keeps*

Part of speech tagging

1. A word's class – its part of speech – tells us a lot about how we should expect it to behave
   - Remember: *You shall know a word by the company it keeps*

2. The distribution of word classes across different registers, genres, and styles of language varies in pronounced and predictable ways
   - E.g. narrative discourse in English features prominently more past tense verbs, third-person pronouns, and attributive adjectives than non-narrative discourse (Biber 1995: 152)

Other linguistic techniques

- In the code-along session, we're going to see how we can use the NLP framework *spaCy* to extract linguistic information

- Grammatical analysis
  - https://demos.explosion.ai/displacy

- Named entity recognition
  - https://demos.explosion.ai/displacy-ent

Take-home points

- Tokenization is not always a simple task, but it's fundamental to how we work with natural language data

- Clean text is essential for nearly all downstream NLP tasks
  - Think: garbage in -> garbage out

- Despite sometimes being taken for granted and as ‘solved problems’, these are are often not trivial tasks, and one-size-fits-all universal solutions rarely exist

- If you haven’t already, read Tahmasebi & Hengchen (2019) from the syllabus

Additional reading

- **Biber, D. (1995).** *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

- **Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020).** "spaCy: Industrial-strength natural language processing in python". Zenodo. https://doi.org/10.5281/zenodo.1212303

- **Jurafsky, D. & Martin, J.H. (2023)**. *Speech and Language Processing*, 3rd edition online pre-print.

- **McCulloch, G. (2019).** *Because Internet.* New York, NY: Riverhead Books.

Break

And head over to UCloud...